Lower Bounds for Error-Correcting Codes with Local Recovery

GUANGDA HU

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science

Advisor: Zeev Dvir

January 2017

Abstract

Error-correcting codes (ECCs) are ubiquitous in computer science. A common property of ECCs is *local recovery*, which demands that given a corrupted codeword, a single lost code symbol can be recovered by reading only a small part of the codeword. An intriguing problem is to find the most "efficient" ECCs (e.g., codes with short length, codes over a small alphabet) with certain types of local recovery. Both constructions and lower bounds have been proven in the literature. However, the problem is still largely open. In this thesis, we prove three lower bound results on different types of ECCs with local recovery:

Firstly, we propose an approximate version of locally decodable codes (LDCs) and prove lower bounds that are similar to the known ones for traditional LDCs. The concerned approximate LDCs are over real numbers and they support recoveries by querying constant number of codeword symbols. The 2-query case (the bulk of our work) is partially related to the lower bound of constant query LDCs, which is a major open problem.

Secondly, we generalize the Sylvester-Gallai (SG) theorem to a subspace version. Generally speaking, the setting of the SG theorem is equivalent to 2-query locally correctable codes (LCCs), and our generalization corresponds to the block version of 2-query LCCs.

Thirdly, we consider a realistic storage model that is a unification of several families of codes studied in the literature. We prove negative results for codes that attain the maximal recovering capability under this model. Our lower bound rules out the possibility of constructions of efficient codes for most parameter settings. We will also explore some results in the construction direction in the appendix.

Acknowledgements

First and foremost, I would like to thank my advisor Zeev Dvir, for his guidance throughout my PhD studies and numerous great advices on both research and paper presenting. Without him, none of the papers that this thesis depends on would be possible.

I would also like to thank Sergey Yekhanin, for his mentoring during my internship at MSR and many helpful discussions on a work that has been included in this thesis.

Thanks to my thesis committee Zeev Dvir, Shubhangi Saraf, Sanjeev Arora, Mark Braverman and Elad Hazan for their time and valuable feedbacks.

My research would not be possible without my coauthors: Jop Briët, Zeev Dvir, Parikshit Gopalan, Swastik Kopparty, Shunhangi Saraf, Carol Wang and Sergey Yekhanin. Thanks to all of them.

I'm grateful to Xiaoming Sun and Wei Chen for their guidance during my undergraduate studies and introduction to theoretical computer science.

On a more personal level, I would like to thank my family for their unconditional constant supports throughout my PhD studies and my life.

This thesis is based on research done during my PhD, which was supported by NSF grants CCF-0832797, CCF-1217416, DMS-1451191 and by the Sloan fellowship.

Contents

List of Results

List of Figures

Chapter 1

Introduction

Error-correcting codes (ECCs) are very widely used in all kinds of applications and theoretical studies. They encode a *message* into a *codeword* in such a way that certain errors in the codeword can be corrected. Most known families of ECCs are *linear*, which means that the codeword is obtained by a linear transformation on the message. In a linear ECC, all data symbols are elements of some field $\mathbb{F}$, which we call the *alphabet*, and the message and the codeword are therefore vectors consisting of elements of $\mathbb{F}$. Let $\boldsymbol{m}$ and $\boldsymbol{c}$ denote the message vector and the codeword vector respectively (for consistency we will only use column vectors throughout this thesis). Then there is

$$\boldsymbol{c}^{\mathsf{T}} = \boldsymbol{m}^{\mathsf{T}} \cdot G,$$

where G is the matrix that defines the code and is known as the *generator matrix*. In other words, every symbol of the codeword is a linear combination of the symbols of the message. Note that the range of $\boldsymbol{c}$ is some vector space (the span of the rows of G). One also uses the *parity check matrix* H to define the space of all codewords:

$$\Big\{\boldsymbol{c} : H \cdot \boldsymbol{c} = \boldsymbol{0}\Big\},$$

where the rows of H are coefficients of the linear constraints (known as *parity check equations*) on codeword coordinates. Typically, the length of the codeword is greater than that of the message, which means there is redundant information so that errors in the codeword can possibly be corrected. See for example [MS77] for a comprehensive introduction on linear ECCs. One simple example of a linear ECC is as following: Let the alphabet be $\mathbb{F}_2$. For a message $\boldsymbol{m}$, we calculate the sum of all entries (bits) of $\boldsymbol{m}$ (which is known as a *parity*), and let the codeword $\boldsymbol{c}$ be the concatenation of $\boldsymbol{m}$ and this parity. In this code, if one symbol (bit) of the codeword is lost, it can be easily recovered from the remaining part. However, this code cannot handle errors on two or more symbols.

Given a corrupted codeword, in order to recover one (recoverable) symbol in the original message or the original codeword, we might need to query many symbols in the corrupted codeword, which could be inefficient. We follow the convention in the literature to use the word “local” to describe the recovery if the number of queries needed is always small. In this thesis, we are interested in two classes of linear ECCs in which certain errors can be corrected locally, namely *locally decodable/correctable codes* and *maximally recoverable codes*, and we will give impossibility results on the constructions of some types of these codes.

1.1 Locally decodable and correctable codes

Locally decodable codes (LDCs) are ECCs that every individual symbol of the original message can be retrieved with high probability by querying a *small number* of random codeword coordinates, where the codeword may contain up to a constant fraction of erroneous symbols at unknown locations. The notion of LDCs was formally defined in [KT00] (and implicitly in prior works such as [Lip90, BF90, BFLS91, GLR+91, FF93, BK95]). Many well-known ECCs are LDCs (with different parameters), e.g., Hadamard codes, Reed-Solomon codes. Obvious applications of LDCs include efficient reliable data transmissions and storage systems. LDCs are also used in areas of theoretical computer science, e.g., average-case complexity [Lev87]. And one other problem closely related to LDCs is *private information retrieval (PIR)* schemes, which are data retrieval protocols that hide the user's request from each individual database server [CGKS98]. Linear LDCs over infinite fields ($\mathbb{R}$ or $\mathbb{C}$) are studied as well as those over finite fields. They found applications in compressed sensing [Don06, CRT06], and are considered in relevant topics that are for arbitrary fields, e.g., [DS07] considered LDCs over any field in the context of polynomial identity testing for arithmetic circuits.

Let d denote the length of the message (which is also the dimension of the space of all codewords), n denote the length of the codeword, and q denote the maximum number of queries allowed to retrieve each symbol. A major question in the area of LDCs is to find out the minimum value of the code length n needed (as a function of d) for given values of q (other parameters such as the fraction of errors in the codeword are usually constants). We briefly summarize the known results as follows:

1. For $q = d^{\varepsilon}$, where $\varepsilon > 0$, there are LDCs known as *multiplicity codes* with n close to d [KSY14].

2. For $q = (\log d)^t$, where $t > 1$, traditional Reed-Muller codes yield q-query LDCs with lengths $n \approx d^{1+1/(t-1)}$.

3. For the case that $q \geq 3$ is a constant independent of d, the length n had been conjectured to be at least exponential of d until a series of surprising works [Yek08, Rag07, KY09, Efr12, IS10, CFL+13, DGY11, BET10] constructed *matching-vector codes* with sub-exponential (still super-polynomial) lengths $n \approx \exp\exp\big((\log d)^{1/\log q}\big)$. The best known lower bound, however, is only $n = \widetilde{\Omega}\big(d^{1+1/(\lceil q/2\rceil-1)}\big)$ [KdW04, Woo07] and $n = \Omega(d^2)$ for the special case of 3-query linear LDCs [Woo12].

4. For $q = 2$, it is known that the length $n = \exp\big(\Omega(d)\big)$ [GKST06, KdW04, DS07], and Hadamard codes are 2-query LDCs that attain this lower bound.

5. For $q = 1$, it was shown in [KT00] that no 1-query LDCs exist over finite fields, and it is also easy to see that 1-query linear LDCs do not exist over characteristic-zero.

See also the survey [Yek12] for the mentioned LDCs. For the regime of constant number of queries, we see that the minimum value of the code length n is well understood only for the cases $q = 1$ and 2. For $q \geq 3$, there is a huge gap between the lower bounds (at most quadratic) and the best constructions (super-polynomial). It is a difficult and intriguing open problem to make any progress on closing this gap. For the lower bound side, several different techniques yield $n = \widetilde{\Omega}(d^2)$ but there is no way to go beyond this.

Locally correctable codes (LCCs) are similar to LDCs except that we want to retrieve a symbol of the codeword instead of the message. Precisely, LCCs are ECCs that every individual symbol of the original *codeword* can be retrieved with high probability by querying a *small number* of random codeword coordinates, where the codeword may contain up to a constant fraction of erroneous symbols at unknown locations. For any LCC, we can obtain an LDC from it by fixing d coordinates of the codeword to be the same as the message. This can be easily done through a linear transformation that changes a submatrix of the generator G to the $d \times d$ identify, and such an operation does not break the properties of LCCs since it does not change the space of codewords. Thus, LCCs are considered as a stronger version of LDCs, and lower bounds on LCCs might be an easier problem to start with if lower bounds on LDCs are difficult. Many known LDCs are also LCCs, e.g., Hadamard codes over $\mathbb{F}_2$, Reed-Muller codes and multiplicity codes over finite fields. Besides the aforementioned applications of LDCs, LCCs (and LDCs) are also closely related to rigid matrices which in turn imply results on circuit complexity [Dvi11] and some combinatorial geometry problems that will be discussed later.

The known results for the minimum lengths of LCCs are similar to those of LDCs. In particular, for constants $q \geq 3$, there seems to be a lot of room to improve the lower bounds on n. For characteristic-zero fields $\mathbb{R}$ and $\mathbb{C}$, some noticeable differences between LCCs and LDCs are as follows: (1) 2-query LCCs do not exist over $\mathbb{R}$ or $\mathbb{C}$ [BDWY11] whereas Hadamard codes are 2-query LDCs over any field; (2) No construction of LCCs is known over $\mathbb{R}$ or $\mathbb{C}$. (For finite fields, the known constructions are also weaker. LDCs with sub-exponential lengths, i.e., matching vector codes, do not seem to generalize to LCCs); (3) For 3-query linear LCCs over $\mathbb{R}$, a lower bound of the form $n = \Omega(d^{2+\varepsilon})$ for some $\varepsilon > 0$ was proved in [DSW14a], whereas the quadratic barrier of LDC lower bounds is still not broken (over any field).

For linear LDCs and LCCs with constant number of queries, we now state an equivalent and convenient version of their definitions that will be used in all our discussions. Let $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ be the columns of the generator matrix G and $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ be the standard basis of the message space. From the results of [KT00], the code defined by G is a q-query linear LDC if and only if for every $i \in [d]$, there are $\Omega(n)$

disjoint q-element subsets $\{j_1, j_2, \ldots, j_q\} \subseteq [n]$ such that

$$\boldsymbol{e}_i \in \operatorname{span}\bigl\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}\bigr\}.$$

Correspondingly, the code is a q-query linear LCC if and only if for every $\boldsymbol{v}_j$, where $j \in [n]$, there are $\Omega(n)$ *disjoint* q-element subsets $\{j_1, j_2, \ldots, j_q\} \subseteq [n]$ such that

$$\boldsymbol{v}_j \in \operatorname{span}\bigl\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}\bigr\}.$$

We see that LDCs and LCCs can be equivalently considered as arrangements of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ with many linearly dependent tuples.

Given a q-query LDC (or LCC) as defined above, the decoding procedure is as following: To retrieve the ith symbol of the message (or the jth symbol of the codeword), one simply pick a random q-element subset $\{j_1, j_2, \ldots, j_q\} \subseteq [n]$ such that $\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}$ span $\boldsymbol{e}_i$ (or $\boldsymbol{v}_j$), and then a linear combination of the codeword coordinates at $j_1, j_2, \ldots, j_q$ will yield the required symbol, provided that there are no errors on these coordinates, which happens with high probability if the fraction of errors in the codeword is below some small constant.

1.1.1 Approximate locally decodable codes

We consider linear LDCs over real numbers. Given the above convenient form, one natural generalization is an "approximate" version of LDCs, where the vectors $\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}$ span $\boldsymbol{e}_i$ only "approximately". Precisely, we define *approximate LDCs* as ECCs that for every $i \in [d]$, there are $\Omega(n)$ disjoint q-element subsets $\{j_1, j_2, \ldots, j_q\} \subseteq [n]$ each with a nonzero vector

$$\boldsymbol{u} \in \operatorname{span}\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}\}$$

such that that angle between $\boldsymbol{u}$ and $\boldsymbol{e}_i$ is smaller than some fixed $\theta_{\max} > 0$. We see that approximate LDCs are a relaxed version of LDCs.

Contribution 1: In Chapter 2 (based on [BDHS14]), we prove lower bounds on the lengths of approximate LDCs. Specifically, for any 2-query approximate LDCs, there is $n = \exp\bigl(\Omega(\sqrt{d})\bigr)$, and this can be improved to $n = \exp\bigl(\Omega(d)\bigr)$ for the "almost exact" case (i.e., the angle bound $\theta_{\max}$ is smaller than some absolute constant); for any q-query approximate LDC, where $q \geq 3$ is a constant, there is $n = \Omega\bigl(d^{1+1/(q-1)}\bigr)$.

Although our lower bounds are a bit weaker than the known ones for LDCs, we are not aware of any constructions of approximate LDCs that are shorter than LDCs. It remains an open problem to either improve these lower bounds to match those for LDCs, or give constructions with shorter lengths that take advantage of being approximate.

Our motivation for studying this problem comes from several directions:

Firstly, one could hope to use approximate LDCs in practice. As long as the message vector is bounded and $\theta_{\max}$ is sufficiently small, we can retrieve an approximation of every individual symbol of the message by making q queries to the codeword. Approximate LDCs might be more efficient than LDCs in some applications if there existed a construction with sufficient good parameters.

Secondly, approximate LDCs are related to the problem of LDC/LCC lower bounds. As we have mentioned, the first *super-quadratic* lower bound for LDCs/LCCs with three or more queries was given in [DSW14a], on 3-query LCCs over real numbers. Some cases of 3-query LCCs can actually be reduced to 2-query approximate LDCs (although a different simpler proof was used in [DSW14a]). This raises the possibility that, in the future, perhaps approximate LDCs will find more applications.

Thirdly, there are connections to topics in combinatorial geometry. Recall that LDCs and LCCs are arrangements of vectors with lots of linear dependencies. Approximate LDCs can be naturally considered as a relaxed version of this geometry problem. A related work is [ADSW14], where the Sylvester-Gallai theorem (which is also on arrangements with many dependent tuples and will be discussed later) and LCCs are generalized to a (different) approximate version.

1.1.2 Sylvester-Gallai for subspaces

The Sylvester-Gallai theorem and a series of its generalizations (in for example [Kel86, BDWY13, DSW14b]) consider arrangements of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ (over $\mathbb{R}$ or $\mathbb{C}$) which are in different directions, and satisfy that for every $\boldsymbol{v}_{j_1}$ there are $\Omega(n)$ choices of another vector $\boldsymbol{v}_{j_2}$ such that there is a third vector $\boldsymbol{v}_{j_3} \in \operatorname{span}\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}\}$. This is similar to the setting of 2-query linear LCCs, where for every vector $\boldsymbol{v}_j$ there are $\Omega(n)$ disjoint pairs $\{j_1, j_2\} \subseteq [n]$ with $\boldsymbol{v}_j \in \operatorname{span}\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}\}$ (an important difference is that vectors in the same direction are allowed in LCCs). The Sylvester-Gallai theorem gives an upper bound on the dimension of the arrangement (which is a constant when n grows). Correspondingly, a lower bound on the length n of an LCC (expressed using d) can also be considered as an upper bound on the dimension d (expressed using n). Thus, the Sylvester-Gallai theorem and lower bounds for 2-query linear LCCs are closely related notions. Their equivalence (with loss in parameters and special handling of some cases) has been observed in [BDWY11, DSW14b].

Contribution 2: In Chapter 3 (based on [DH16]), we prove a generalization of the above Sylvester-Gallai theorem where the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ are replaced with k-dimensional vector subspaces $V_1, V_2, \ldots, V_n$ (over $\mathbb{C}$) for any positive integer k. Under the assumption $V_j \cap V_{j'} = \{\mathbf{0}\}$ for all $j \neq j'$ (otherwise a counter-example exists), we show that $V_1, V_2, \ldots, V_n$ are contained in a space of dimension at most $\operatorname{poly}(k)$, which is independent of n.

The setting of our generalized Sylvester-Gallai corresponds to 2-query *block LCCs*, where the coordinates of the codeword are partitioned into blocks of the same size, and every block of codeword symbols can be retrieved by querying two random blocks of codeword symbols. Let k be the block size, n be the number of blocks in the codeword (then the length of the codeword is kn), and $V_1, V_2, \ldots, V_n$ be the subspaces spanned by every k consecutive columns of the generator G. The code defined by G is a 2-query block linear LCC if and only if for every $j \in [n]$, there are $\Omega(n)$ disjoint pairs

$\{j_1, j_2\} \subseteq [n]$ such that

$$V_j \subseteq V_{j_1} + V_{j_2},$$

where $V_{j_1} + V_{j_2}$ denotes the set $\{\boldsymbol{v} + \boldsymbol{v}' : \boldsymbol{v} \in V_{j_1}, \boldsymbol{v}' \in V_{j_2}\}$, i.e., the space spanned by all vectors in $V_{j_1} \cup V_{j_2}$. If $V_j \cap V_{j'} = \{\boldsymbol{0}\}$ is satisfied for all $j \neq j'$ and every V_j has dimension exactly k, one can see that 2-query block linear LCCs are just a special case of our Sylvester-Gallai theorem for subspaces.

Our motivation for studying block LCCs comes from *block LDCs*, where the coordinates of the codeword are partitioned into blocks of the same size like in the above block LCCs, and every individual symbol of the message can be retrieved by querying a small number of random blocks of codeword symbols. Block LDCs are used in aforementioned private information retrieval (PIR) schemes. Suppose we have a q-query block LDC that all blocks are queried with equal probability, in the procedure of retrieving any message symbol. Then the following database retrieval protocol between one user and q servers is a PIR scheme: Let the message of the block LDC be the entire database, and let every server store a copy of the corresponding codeword. To retrieve a symbol of the database, the user simply asks the q servers respectively for the q random blocks needed to query. In this protocol, every individual server sees a uniform distribution of a random block query, and learns no information about the database symbol that the user is interested in. On the other hand, if we have a one-round q-server database retrieval protocol that in order to retrieve a database symbol, the user sends to every server a random string with a distribution independent of the symbol being retrieved, and each server responds with a string of k symbols, then we can obtain a block LDC by defining the codeword as the concatenation of all possible responses of the q servers. Therefore block LDCs and PIR schemes are equivalent. See also for example [KT00, GKST06] for their connections. Block LCCs are the corresponding LCC version of block LDCs.

1.2 Maximally recoverable codes

In a real distributed storage system, the setting is often different from that of the LDCs/LCCs discussed in the previous section:

1. In LDCs/LCCs corruptions are at unknown locations, whereas in a storage system we usually know which hard disks or sectors are bad. Thus it suffices to handle *erasures* at known locations. Moreover, the corrupted locations may be correlated (e.g., in one failed hard disk all sectors are bad) rather than being completely arbitrary.

2. We would like a deterministic recovering algorithm that always gives the correct result instead of a randomized one. And knowing the erroneous locations makes this possible.

3. The *topology* of the code is often predetermined. A *topology* includes the number of code symbols, the number of parities, and the sets of symbols that each parity depends on. One can think of a topology as a specification of the nonzero

locations of the parity check matrix. A predetermined topology simplifies the design and can handle errors at correlated locations better.

For example, in a storage system one may use several hard disks to store the original data, and then add a redundant hard disk in which every byte is a linear combination of the corresponding bytes of the previous hard disks. In this case, the topology is fixed and we can only change the coefficients of the linear combination.

4. The resources are limited. In the LDCs/LCCs problem a lower bound or construction over any field will be interesting, whereas in a storage system we would only prefer small finite fields for computational efficiency. See also [PGM13] for the importance of using a small finite field as the alphabet.

   If the topology is fixed, the length of the code will be fixed. Thus, our goal is to minimize the alphabet size instead of the code length (under certain reliability requirements).

For a fixed topology, there can be many codes *instantiating* this topology (i.e., specifications of the nonzero entries of the parity check matrix) with different levels of reliabilities. We say that a set of code coordinates E is a *recoverable pattern* for a given topology T, if there exists a code instantiating T such that the symbols in E can be recovered if they are all erased. And a *maximally recoverable (MR)* code instantiating T is a code that can recover all recoverable patterns for T. MR code exists for any topology if the underlying field is sufficiently large [GHJY14]. One can consider MR codes as ECCs with the maximal recovering capability for a given topology. The notion of MR codes were first considered in [HCL07, CHL07] for the restricted setting that there are only “parities of data symbols” and no “parities of parities”. MR codes for more general cases and various topologies have since been studied in [BHH13, Bla13, GHJY14, CSYS15, BPSY16].

In this thesis, we are interested in the following topology denoted by $T_{m\times n}(a,b,h)$: The code symbols are arranged as an $m \times n$ matrix, where there are a parity check equations per column, b parity check equations per row, and h additional parity check equations that can depend on all code symbols. Generally speaking, one can equivalently consider a code instantiating $T_{m\times n}(a,b,h)$ as having $(m-a)(n-b)-h$ data symbols with h parities, which form an $(m-a)\times(n-b)$ matrix, plus a additional *local* parities in every column and b additional *local* parities in every row. (One can see that there are $a \times b$ local parities that are for both columns and rows.)

Like LDCs/LCCs, this model of codes also supports a type of “local” recovery. If there are at most a erasures in a column or at most b erasures in a row (which is often the case in practice), the lost data can be recovered using the local parities. (The other h parities provide additional “non-local” reliabilities for more complicated erasure patterns.) In fact, the case $a = 1$, $b = 0$ is a common model of codes studied in the context of *locally recoverable codes (LRCs)*, which are defined as ECCs that every symbol can be recovered locally if it is erased [GHSY12].

The topology $T_{m\times n}(a,b,h)$ can be considered as a unification of several common topologies in the literature:

1. The topology $T_{m\times n}(1,0,h)$ has received a considerable amount of attention especially in the recent work on LRCs [BHH13, GHSY12, GHJY14, BK15, LL15, BPSY16]. One can see that in this topology, the original data and h parities are partitioned into n groups (columns), and a local parity is added to each group.

   Another topology studied in the context of LRCs (e.g., [GHSY12]) is to partition just the original data into n groups where each is added a local parity, and after that add another h parities that depend on both the data and the local parities. MR codes instantiating this topology are equivalent to those instantiating $T_{m\times n}(1,0,h)$ [GHJY14].

   It is also worth mentioning that instead of maximal reliability, another related (but weaker) reliability requirement for LRCs in the literature is to maximize the hamming distance between codewords under certain locality conditions [GHSY12, PKLK12, PD14, TB14, CM15].

2. *Maximum distance separable (MDS)* codes, which are defined as ECCs maximizing the hamming distances between codewords for given values of message length and codeword length, can be viewed as MR codes instantiating the topology $T_{1\times n}(0,0,h)$, where n is the codeword length and $n-h$ is the message length. MDS codes are very widely used in all kinds of applications. One common family of MDS codes are Reed-Solomon codes.

3. The topology $T_{m\times n}(a,b,0)$ can be considered as a tensor product of two codes. Tensor product codes are common in storage systems (see for example [RR72]). A code instantiating $T_{3\times 14}(1,4,0)$ is used by Facebook's f4 storage system [MLR$^+$14]. The code is the tensor product of a Reed-Solomon code within data centers with a parity check code across data centers.

4. An MR code instantiating a topology closely related to $T_{2\times 7}(0,1,2)$ is used by Microsoft's Azure storage [HSX$^+$12].

Contribution 3: In Chapter 4 (based on parts of [GHK$^+$17]), we prove a super-polynomial (consider a,b,h as constants and m,n as growing variables) lower bound on the field size of MR codes instantiating $T_{m\times n}(a,b,h)$ for all $a,b,h\geq 1$.

Prior to our result, the only known lower bound is linear $\Omega(mn)$ for $h\geq 2$ [Bal12, GHJY14]. Our result shows that MR codes with polynomial field size might exist only when one of a,b,h is zero. After accounting for symmetries, there are two cases:

1. Tensor product of codes $T_{m\times n}(a,b,0)$. MR codes instantiating this topology are poorly understood. We know neither explicit constructions nor non-trivial lower bounds. In fact, even the recoverable patterns are not well understood. In [GHK$^+$17], a characterization of the recoverable patterns for $T_{m\times n}(a,b,0)$ was conjectured, and was proved for the case $a=1$. MR codes instantiating $T_{m\times n}(a,b,0)$ might exist over small fields.

2. Generalized LRCs $T_{m\times n}(a,0,h)$. In LRCs, one typically considers the case $a=1$. The generalization to arbitrary a was also studied in the literature

[PKLK12, BHH13, BPSY16]. The only known field size lower bound for MR codes instantiating the topology $T_{m\times n}(a,0,h)$ is the aforementioned $\Omega(mn)$ for $h \geq 2$, and the best construction for the case $a = 1$ and general h has field size roughly $O\big(\min\big\{2^m n^{h-1}, (mn)^{(n+h)/2}\big\}\big)$ (see [GHJY14] and Theorem A.5). It remains an intriguing open problem to close the gap.

Finally, we consider the construction direction for the topology $T_{m\times n}(1,0,h)$. As we have mentioned, this topology is a common type of codes considered in the literature, especially topics on LRCs.

Contribution 4: In Appendix A (based on parts of [GHK$^+$17] and [HY16]), we give two new explicit families of MR codes (which are alternative to some previously known constructions) for the cases $h = 2$ and general h respectively.

Chapter 2

Lower Bounds for Approximate LDCs

In this chapter, we study lower bounds on the encoding length of an approximate version of LDCs. Most of this chapter will be devoted to the 2-query case. In Section 2.1, we define the model of codes formally and state our results. Then in Section 2.2, we reduce 2-query approximate LDCs to a convenient notion about graphs. Based on this, in Section 2.3 we prove a general lower bound for 2-query approximate LDCs and a stronger (tight) one for 2-query "almost exact" LDCs. Finally, in Section 2.4 we consider q-query approximate LDCs for $q > 2$ and prove a simple lower bound similar to the known ones for "exact" LDCs. The results in this chapter are also included in [BDHS14].

2.1 Generalization of LDCs and results

We begin by introducing some notations. Throughout this chapter, we consider linear codes that encodes a message $\boldsymbol{m} \in \mathbb{R}^d$ into a codeword $\boldsymbol{c} \in \mathbb{R}^n$, where $d, n \in \mathbb{Z}^+$ are the *dimension* (or *message length*) and the *code length* respectively. Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ denote the standard basis of $\mathbb{R}^d$. For a nonzero vector $\boldsymbol{u} \in \mathbb{R}^d$ and an index $i \in [d]$, we define

$$\text{weight}_i(\boldsymbol{u}) = \frac{|\langle \boldsymbol{u}, \boldsymbol{e}_i \rangle|}{\|\boldsymbol{u}\|_2}.$$

Clearly, $\text{weight}_i(\boldsymbol{u}) \leq 1$ and the equality holds if and only if $\boldsymbol{u} = \boldsymbol{e}_i$. For an integer $q \geq 2$, we define a *q-matching* as a family of *disjoint* q-element subsets of $[n]$. We will use *q-tuples*, or *pairs* for $q = 2$, to refer to the subsets contained in a q-matching.

Recall that every linear code that encodes a message $\boldsymbol{m} \in \mathbb{R}^d$ into a codeword $\boldsymbol{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ can be defined as

$$c_j = \langle \boldsymbol{m}, \boldsymbol{v}_j \rangle \quad \forall j \in [n],$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^d$ are the columns of the generator matrix. For any constant $q \geq 2$, it is well known that the code is a q-query LDC if and only if for every $i \in [d]$, there exists a q-matching M_i of size $\Omega(n)$ such that

$$\boldsymbol{e}_i \in \operatorname{span}\bigl\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}\bigr\} \tag{2.1}$$

for every q-tuple $\{j_1, j_2, \ldots, j_q\} \in M_i$ [KT00]. Thus, we may equate the notion of q-query LDCs and pairs $(\mathcal{V}, \mathcal{M})$, where $\mathcal{V}$ is a list of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^d$ and $\mathcal{M}$ is a list of q-matchings $M_1, M_2, \ldots, M_d$, such that the above requirements are satisfied. Based on this observation, we define *approximate LDCs* as a generalization of LDCs, where the "span" in Equation (2.1) is replaced with an "approximate span":

Definition 2.1 (approximate LDC)**.** For $q \geq 2$ and $\alpha, \delta \in (0, 1]$, we define a q-query (α, δ)-*approximate LDC* as a pair $(\mathcal{V}, \mathcal{M})$, where $\mathcal{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$ is a list of vectors in $\mathbb{R}^d$, and $\mathcal{M} = (M_1, M_2, \ldots, M_d)$ is a list of q-matchings, such that the follows are satisfied:

1. For every $i \in [d]$ and every q-tuple $\{j_1, j_2, \ldots, j_q\} \in M_i$, there exists a vector $\boldsymbol{u} \in \operatorname{span}\{\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \ldots, \boldsymbol{v}_{j_q}\}$ with $\text{weight}_i(\boldsymbol{u}) \geq \alpha$;
2. $|M_1| + |M_2| + \cdots + |M_d| \geq \delta d n$.

We note that the first item in the above definition is a generalization of Equation (2.1), and the original "exact" LDCs correspond to the case $\alpha = 1$. The second item $|M_1| + |M_2| + \cdots + |M_d| \geq \delta d n$ is also more general than the requirement for "exact" LDCs that demands $|M_i| = \Omega(n)$ for all $i \in [d]$, and this makes our (negative) results stronger.

For "exact" LDCs, there have been exponential lower bounds $n = \exp\bigl(\Omega(d)\bigr)$ for the 2-query case [GKST06, KdW04, DS07]. In this chapter, we will study lower bounds for 2-query approximate LDCs. An observation in Section 2.2 will show that it suffices to consider *simple* 2-query approximate LDCs defined as following:

Definition 2.2 (simplicity)**.** Let $(\mathcal{V}, \mathcal{M})$ be a 2-query (α, δ)-approximate LDC. We say that $(\mathcal{V}, \mathcal{M})$ is *simple* if for every $i \in [d]$ and every pair $\{j_1, j_2\} \in M_i$, there is

$$\text{weight}_i(\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}) \geq \alpha.$$

Equivalently, we can define a simple 2-query approximate LDC as an arrangement of points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ such that for every $i \in [d]$, there are δn (on average) disjoint pairs $\{j_1, j_2\}$ that the line passing through $\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}$ is in a direction that has projection as least α onto $\boldsymbol{e}_i$. An example of such an arrangement is the vertices of the Boolean hypercube $\{0, 1\}^d$ (i.e., the arrangement corresponding to the Hadamard code), which satisfies Definition 2.2 for $\alpha = 1, \delta = 1/2$ if we choose M_i to be the family of all the $n/2$ vertex pairs that differ only at the ith coordinate.

Intuitively, we may consider the arrangement as a graph, where the vertices are the points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, and the edges are the pairs in $M_1, M_2, \ldots, M_d$. Also in Section 2.2, by generalizing a simple lemma used in [GKST06], we will show that if there exists a small cut in every induced subgraph, then the number of edges (at least

δdn) can be bounded from above using the number of vertices (which is n). From this we can derive a lower bound on n expressed using d. Thus the problem is transformed to showing the existence of small cuts in the arrangement of points.

Based on the above reduction and using a probabilistic argument, we are able prove our first result as stated below:

Theorem 2.3. *For any 2-query (α, δ)-approximate LDC of dimension d and length n, we have*

$$n = \exp\bigl(\Omega(\alpha\delta\sqrt{d})\bigr),$$

where $\Omega(\cdot)$ hides an absolute constant independent of α, δ or d.

We note that this bound gets worse as α approaches $1/\sqrt{d}$, at which point we cannot expect any non-trivial lower bounds, since every single vector $\boldsymbol{v}_j$ can have $\text{weight}_i(\boldsymbol{v}_j) \geq 1/\sqrt{d}$ for all $i \in [d]$ and so, an arbitrary list of 2-matchings satisfies Definition 2.1. However, we conjecture that Theorem 2.3 is not tight, and probably a lower bound of the form $n = \exp\bigl(\Omega(\alpha^2\delta d)\bigr)$ should hold, where on the exponent we have d instead of $\sqrt{d}$ just as in the aforementioned lower bounds or 2-query "exact" LDCs. Currently, we are only able to prove this when α is close to 1. The idea is to round the points in the arrangement to the vertices of a grid, where a cut orthogonal to a standard direction will be small. The rounding is found using a beautiful result on space tiling in [KORW12]. Our second result is as following:

Theorem 2.4. *Let α_0 denote the constant $\sqrt{1 - 1/(4\pi^2)} = 0.98725\cdots$. For any constant $\alpha > \alpha_0$ and any 2-query (α, δ)-approximate LDC of dimension d and length n, we have*

$$n = \exp\bigl(\Omega(\delta d)\bigr),$$

where $\Omega(\cdot)$ hides a constant that is directly proportional to $(\alpha - \alpha_0)^{2.5}$.

The above two theorems will be proved in Section 2.3. Theorem 2.4 is tight because of the Hadamard code, where $\alpha = 1$, $\delta = 1/2$ and $n = 2^d$. It is worth mentioning that we are not aware of any approximate LDCs with shorter lengths than "exact" LDCs, and it is an open problem to give such constructions.

For q-query "exact" LDCs, where $q > 2$, the best known lower bounds on the code length are only super-linear or quadratic of the code dimension [KT00, KdW04, Woo07, Woo12], and it is a very difficult open problem to prove better results. For q-query approximate LDCs, where q is a general integer, we will show in Section 2.4 a super-linear lower bound as stated below. The proof is a simple modification of the existing techniques in [KT00] that have given a similar super-linear lower bound for "exact" LDCs.

Theorem 2.5. *Let $q \geq 2$ be a constant. For any q-query (α, δ)-approximate LDC of dimension d and length n, we have*

$$n = \Omega\bigl(\delta^{\frac{q+1}{q-1}}(\alpha^2 d)^{\frac{q}{q-1}}\bigr),$$

where $\Omega(\cdot)$ hides a constant that only depends on q.

2.2 Reduction to arrangements of points

We first prove the following lemma, which shows that any 2-query approximate LDC can be transformed into a simple one (Definition 2.2) with similar parameters:

Lemma 2.6. *If there exists a* 2*-query* (α, δ)*-approximate LDC of dimension* d *and length* n*, then for any* $\alpha' \in (0, \alpha)$ *such that*

$$\delta' = \delta - \frac{1}{(\alpha^2 - \alpha'^2)d} > 0,$$

there exists a simple 2*-query* (α', δ')*-approximate LDC of dimension* d *and length* $2n$*.*

Proof. Let $(\mathcal{V}, \mathcal{M})$ be a 2-query (α, δ)-approximate LDC, where $\mathcal{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$ and $\mathcal{M} = (M_1, M_2, \ldots, M_d)$. We will construct a code with the desired properties based on $(\mathcal{V}, \mathcal{M})$.

For every $j \in [n]$ such that $\boldsymbol{v}_j \neq \boldsymbol{0}$, we replace $\boldsymbol{v}_j$ with the unit vector $\boldsymbol{v}'_j = \boldsymbol{v}_j / \|\boldsymbol{v}_j\|_2$. Let $\mathcal{V}' = (\boldsymbol{v}'_1, \boldsymbol{v}'_2, \ldots, \boldsymbol{v}'_n)$ denote the resulting list of vectors.

Then for all $j \in [n]$, $i \in [d]$ such that $\boldsymbol{v}'_j \neq \boldsymbol{0}$ and $\text{weight}_i(\boldsymbol{v}'_j) \geq \sqrt{\alpha^2 - \alpha'^2}$, we remove any pair in M_i that contains j (there is at most one such pair since the pairs in M_i are disjoint). Let $\mathcal{M}' = (M'_1, M'_2, \ldots, M'_d)$, $M'_i \subseteq M_i$, denote the resulting list of 2-matchings. We see that $(\mathcal{V}', \mathcal{M}')$ is a 2-query (α, δ')-approximate LDC from the following claim:

Claim 2.7. $|M'_1| + |M'_2| + \cdots + |M'_d| \geq \delta dn - n/(\alpha^2 - \alpha'^2) = \delta' dn.$

Proof. For every $j \in [n]$ such that $\boldsymbol{v}'_j \neq \boldsymbol{0}$, there are at most $1/(\alpha^2 - \alpha'^2)$ values of $i \in [d]$ with $\text{weight}_i(\boldsymbol{v}'_j) \geq \sqrt{\alpha^2 - \alpha'^2}$, and so there were at most $1/(\alpha^2 - \alpha'^2)$ pairs containing j removed from $M_1, M_2, \ldots, M_d$. The claim follows immediately. ■

Next, we fix an $i \in [d]$ and a pair $\{j_1, j_2\} \in M'_i$, and study the vectors $\boldsymbol{v}'_{j_1}$, $\boldsymbol{v}'_{j_2}$.

Claim 2.8. *The vectors* $\boldsymbol{v}'_{j_1}$ *and* $\boldsymbol{v}'_{j_2}$ *are linearly independent, i.e.,* $\boldsymbol{v}'_{j_1} \neq \boldsymbol{v}'_{j_2}$*,* $\boldsymbol{v}'_{j_1} \neq -\boldsymbol{v}'_{j_2}$ *and neither of them is* $\boldsymbol{0}$*.*

Proof. We assume the opposite and derive a contradiction. By Definition 2.1, there exists $\boldsymbol{u} \in \text{span}\{\boldsymbol{v}'_{j_1}, \boldsymbol{v}'_{j_2}\}$ with $\text{weight}_i(\boldsymbol{u}) \geq \alpha > 0$. Hence at least one of $\boldsymbol{v}'_{j_1}$ and $\boldsymbol{v}'_{j_2}$ is nonzero. Suppose $\boldsymbol{v}'_{j_1} \neq \boldsymbol{0}$. Then $\boldsymbol{u}$ is a multiple of $\boldsymbol{v}'_{j_1}$, and it follows that $\text{weight}_i(\boldsymbol{v}'_{j_1}) \geq \alpha > \sqrt{\alpha^2 - \alpha'^2}$. This contradicts the definition of M'_i as $\{j_1, j_2\}$ should have been removed. ■

By Claim 2.8, there is a unique plane passing through the origin (i.e., a two-dimensional vector space) that contains $\boldsymbol{v}'_{j_1}$ and $\boldsymbol{v}'_{j_2}$. We set up Cartesian axes on this plane. Let the y-axis be in the direction of the projection of $\boldsymbol{e}_i$ onto this plane, and let the x-axis be in either of the two possible directions. We use $\tau \in [0, \pi/2]$ to denote that angle between $\boldsymbol{e}_i$ and the plane spanned by $\boldsymbol{v}'_{j_1}$, $\boldsymbol{v}'_{j_2}$ (see Figure 2.1a). Since $\{j_1, j_2\} \in M'_i \subseteq M_i$, there is

$$\cos \tau \geq \alpha. \tag{2.2}$$

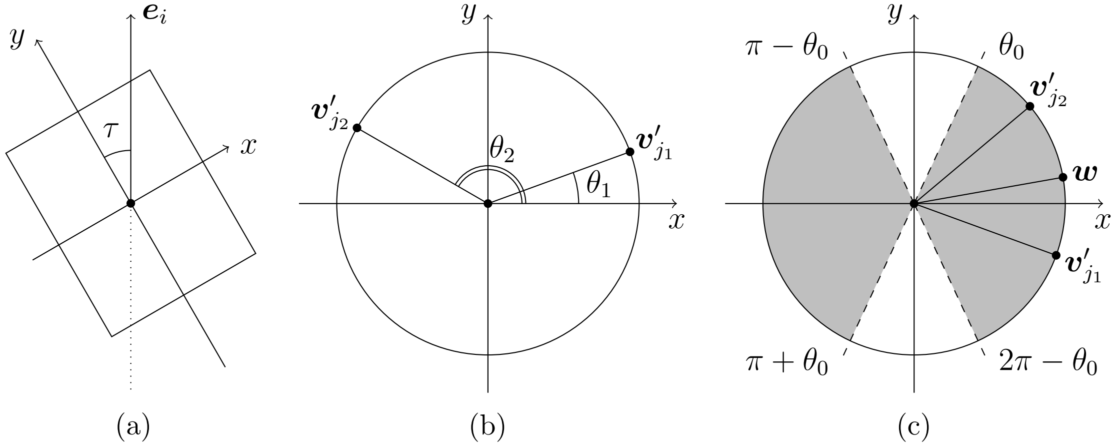

Figure 2.1: (a) The angle between e_i and the plane spanned by $\boldsymbol{v}'_{j_1}$, $\boldsymbol{v}'_{j_2}$; (b) The angles from the x-axis to $\boldsymbol{v}'_{j_1}$ and $\boldsymbol{v}'_{j_2}$; (c) The range of θ_1, θ_2 and the middle point $\boldsymbol{w}$.

Let $\theta_1, \theta_2 \in [0, 2\pi)$ be the angles from the x-axis to $\boldsymbol{v}_{j_1}$ and $\boldsymbol{v}_{j_2}$ respectively (see Figure 2.1b). Then $\boldsymbol{v}_{j_1}$ and $\boldsymbol{v}_{j_2}$ are the points $(\cos\theta_1, \sin\theta_1)$ and $(\cos\theta_2, \sin\theta_2)$ on the plane. By the construction of M'_i, we have $\text{weight}_i(\boldsymbol{v}'_{j_1}) < \sqrt{\alpha^2 - \alpha'^2}$ and $\text{weight}_i(\boldsymbol{v}'_{j_2}) < \sqrt{\alpha^2 - \alpha'^2}$, or equivalently,

$$|\sin\theta_1| \cdot \cos\tau < \sqrt{\alpha^2 - \alpha'^2} \quad \text{and} \quad |\sin\theta_2| \cdot \cos\tau < \sqrt{\alpha^2 - \alpha'^2}.$$

Define $\theta_0 = \arcsin\bigl(\sqrt{\alpha^2 - \alpha'^2}/\cos\tau\bigr)$. By Inequality (2.2), there is $\cos\tau > \sqrt{\alpha^2 - \alpha'^2}$. Hence θ_0 is well defined. We see that θ_1 and θ_2 fall into the following two regions (gray areas in Figure 2.1c):

$$\mathcal{A} = (2\pi - \theta_0, 2\pi] \cup [0, \theta_0) \quad \text{and} \quad \mathcal{B} = (\pi - \theta_0, \pi + \theta_0).$$

Claim 2.9. *If θ_1 and θ_2 are in the same region ($\mathcal{A}$ or $\mathcal{B}$), there is $\text{weight}_i(\boldsymbol{v}'_{j_2} - \boldsymbol{v}'_{j_1}) \geq \alpha'$. If θ_1 and θ_2 are in different regions, there is $\text{weight}_i(\boldsymbol{v}'_{j_2} + \boldsymbol{v}'_{j_1}) \geq \alpha'$.*

Proof. We first consider the case that θ_1 and θ_2 are in the same region. Let $\boldsymbol{w} = (\cos\theta_3, \sin\theta_3)$ denote the middle point of the arc between $\boldsymbol{v}'_{j_1}$ and $\boldsymbol{v}'_{j_2}$ on the unit circle (see Figure 2.1c). Then θ_3 is also in one of the two regions $\mathcal{A}$ and $\mathcal{B}$ (the one that θ_1 and θ_2 fall in), which implies

$$|\cos\theta_3| \geq \cos\theta_0 = \sqrt{1 - \frac{\alpha^2 - \alpha'^2}{(\cos\tau)^2}}.$$

Noting that $\boldsymbol{v}'_{j_2} - \boldsymbol{v}'_{j_1}$ is parallel to the tangent line to the unit circle at $\boldsymbol{w}$,

$$\text{weight}_i(\boldsymbol{v}'_{j_2} - \boldsymbol{v}'_{j_1}) = |\cos\theta_3| \cdot \cos\tau \geq \sqrt{1 - \frac{\alpha^2 - \alpha'^2}{(\cos\tau)^2}} \cdot \cos\tau \geq \alpha'.$$

In the last step we used $\cos\tau \geq \alpha$ (Inequality (2.2)).

For the case that θ_1 and θ_2 are in different regions, one can change $\boldsymbol{v}'_{j_1}$ to $-\boldsymbol{v}'_{j_1}$ and θ_1 to $2\pi-\theta_1$ in the proof of the previous case and show $\mathrm{weight}_i\big(\boldsymbol{v}'_{j_2}-(-\boldsymbol{v}'_{j_1})\big) \geq \alpha'$. ■

Now we construct a code that satisfies the requirements of Lemma 2.6. We define $\mathcal{V}'' = (\boldsymbol{v}''_1, \boldsymbol{v}''_2, \ldots, \boldsymbol{v}''_{2n})$, where $\boldsymbol{v}''_j = \boldsymbol{v}'_j$ and $\boldsymbol{v}''_{j+n} = -\boldsymbol{v}'_j$ for all $j \in [n]$. Then by Claim 2.9, for every $i \in [d]$ and every pair $\{j_1, j_2\} \in M'_i$, at least one of the following two cases holds:

1. $\mathrm{weight}_i(\boldsymbol{v}''_{j_2} - \boldsymbol{v}''_{j_1}) \geq \alpha'$ and $\mathrm{weight}_i(\boldsymbol{v}''_{j_2+n} - \boldsymbol{v}''_{j_1+n}) \geq \alpha'$;
2. $\mathrm{weight}_i(\boldsymbol{v}''_{j_2} - \boldsymbol{v}''_{j_1+n}) \geq \alpha'$ and $\mathrm{weight}_i(\boldsymbol{v}''_{j_2+n} - \boldsymbol{v}''_{j_1}) \geq \alpha'$.

For the first case, we add the pair $\{j_1+n, j_2+n\}$ to M'_i. For the second case, we replace the pair $\{j_1, j_2\}$ with two pairs $\{j_1+n, j_2\}$ and $\{j_1, j_2+n\}$. Let $\mathcal{M}'' = (M''_1, M''_2, \ldots, M''_d)$ denote the resulting list of 2-matchings. Clearly, the code $(\mathcal{V}'', \mathcal{M}'')$ has dimension d and length $2n$. By Claim 2.7, there is

$$|M''_1| + |M''_2| + \cdots + |M''_d| = 2\big(|M'_1| + |M'_2| + \cdots + |M'_d|\big) \geq 2\delta' dn = \delta' d(2n).$$

We see that $(\mathcal{V}'', \mathcal{M}'')$ is a simple 2-query (α', δ')-approximate LDC. □

By setting $\alpha' = \alpha/2$ in Lemma 2.6, we have the following convenient corollary:

Corollary 2.10. *Suppose $d \geq 8/(3\alpha^2\delta)$. If there exists a 2-query (α, δ)-approximate LDC of dimension d and length n, then there exists a simple 2-query $(\alpha/2, \delta/2)$-approximate LDC of dimension d and length $2n$.*

From now on we restrict our attention to simple 2-query (α, δ)-approximate LDCs. For such a code $(\mathcal{V}, \mathcal{M})$, we consider the undirected graph where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ are the vertices and the pairs in $M_1, M_2, \ldots, M_d$ are the edges. Note that we allow parallel edges as one pair $\{j_1, j_2\}$ can appear in multiple 2-matchings. We introduce some notations. For every $i \in [d]$ and every pair $\{j_1, j_2\} \in M_i$, let

$$\big(\{j_1, j_2\}; i\big)$$

denote the edge corresponding to this pair. With abuse of notation, let $\mathcal{M}$ also denote the set of all edges of the graph:

$$\mathcal{M} = \Big\{\big(\{j_1, j_2\}; i\big) : \forall i \in [d], \{j_1, j_2\} \in M_i\Big\}.$$

And we will use the pair $(\mathcal{V}, \mathcal{M})$ to refer to the graph as well as the code. For a general graph $G = (V, E)$ and $S_1, S_2 \subseteq V$, we use $\mathrm{Edge}(S_1, S_2)$ to denote the set of edges between the vertices in S_1 and the vertices in S_2. We define $\mathrm{Edge}(S)$ as a shorthand for $\mathrm{Edge}(S, S)$.

Next, we give the following simple lemma that relates the number of edges in a graph with the sizes of cuts. A variant of this lemma (where the graph is a hypercube) was proved in [Bol86, Section 16] and [GKST06]. Our lemma can be proved using the same proof as in [GKST06]. For completeness, we include the proof here.

Lemma 2.11. *Let $G = (V, E)$ be an undirected graph that parallel edges are permitted and c be a positive real number. Suppose that for every $S \subseteq V$ with $|S| \geq 2$, there exist disjoint nonempty subsets $S_1, S_2 \subseteq S$ with $S_1 \cup S_2 = S$ such that*

$$|\operatorname{Edge}(S_1, S_2)| \leq c \cdot \min\bigl\{|S_1|, |S_2|\bigr\}.$$

Then

$$|E| \leq \frac{c}{2} \cdot |V| \log_2 |V|. \tag{2.3}$$

Proof. We show by an induction on $|S|$ that for every nonempty $S \subseteq V$, there is

$$|\operatorname{Edge}(S)| \leq \frac{c}{2} \cdot |S| \log_2 |S|. \tag{2.4}$$

Then the lemme is proved by setting $S = V$ in this inequality.

For $|S| = 1$, Inequality (2.4) is trivial. Assume that we have proved Inequality (2.4) for all nonempty $S \subseteq V$ of size $|S| < k$, and now we consider the case that $|S| = k$, where $k \geq 2$. Let $S_1, S_2 \subseteq S$ be disjoint nonempty sets with $S_1 \cup S_2 = S$ such that

$$|\operatorname{Edge}(S_1, S_2)| \leq c \cdot \min\bigl\{|S_1|, |S_2|\bigr\}.$$

Then using the induction hypothesis,

$$\begin{aligned}|\operatorname{Edge}(S)| &= |\operatorname{Edge}(S_1, S_2)| + |\operatorname{Edge}(S_1)| + |\operatorname{Edge}(S_2)| \\ &\leq c \cdot \min\bigl\{|S_1|, |S_2|\bigr\} + \frac{c}{2} \cdot |S_1| \log_2 |S_1| + \frac{c}{2} \cdot |S_2| \log_2 |S_2|.\end{aligned}$$

In order to prove Inequality (2.4), it suffices to show

$$c \cdot \min\bigl\{|S_1|, |S_2|\bigr\} + \frac{c}{2} \cdot |S_1| \log_2 |S_1| + \frac{c}{2} \cdot |S_2| \log_2 |S_2| \leq \frac{c}{2} \cdot |S| \log_2 |S|. \tag{2.5}$$

Without loss of generality, we assume $|S_1| \leq |S_2|$. Let $x = |S_1|/|S|$, where $x \in (0, 1/2]$. Then $|S_1| = x|S|$ and $|S_2| = (1-x)|S|$. Inequality (2.5) is equivalent to

$$\begin{aligned}& x + \frac{1}{2} \cdot x \log_2\bigl(x|S|\bigr) + \frac{1}{2} \cdot (1-x) \log_2\bigl((1-x)|S|\bigr) \leq \frac{1}{2} \cdot \log_2 |S| \\ \Longleftrightarrow\quad & 2x + x \log_2 x + (1-x) \log_2 (1-x) \leq 0.\end{aligned}$$

Let $f(x)$ denote the left side of the above inequality. One can verify $\lim_{x \to 0} f(x) = 0$ and $f(1/2) = 0$. By $f''(x) = \bigl(1/x + 1/(1-x)\bigr)/\ln 2 > 0$ for $x \in (0, 1/2]$, we see that $f(x)$ is a convex function. Therefore $f(x) \leq 0$ for all $x \in (0, 1/2]$. □

Recall that in the graph $G = (\mathcal{V}, \mathcal{M})$, the number of edges is at least $\delta d n$ and the number of vertices is n. Plugging these into Inequality (2.3), we have the following corollary:

Corollary 2.12. *Let $(\mathcal{V}, \mathcal{M})$ be a simple 2-query (α, δ)-approximate LDC of dimension d and length n. Consider the graph $G = (\mathcal{V}, \mathcal{M})$. Suppose that for every $S \subseteq \mathcal{V}$*

with $|S| \geq 2$, there exist disjoint nonempty subsets $S_1, S_2 \subseteq S$ with $S_1 \cup S_2 = S$ such that

$$|\operatorname{Edge}(S_1, S_2)| \leq c \cdot \min\big\{|S_1|, |S_2|\big\}.$$

Then

$$n \geq 4^{\delta d/c}.$$

2.3 Lower bounds for 2-query approximate LDCs

With the reduction provided in the previous section, we now proceed to prove lower bounds for 2-query approximate LDCs.

2.3.1 Proof of Theorem 2.3

By Corollary 2.10 and Corollary 2.12, Theorem 2.3 follows immediately from the following lemma:

Lemma 2.13. *Let $(\mathcal{V}, \mathcal{M})$ be a simple 2-query (α, δ)-approximate LDC of dimension d and length n. Consider the graph $G = (\mathcal{V}, \mathcal{M})$. For every $S \subseteq \mathcal{V}$ with $|S| \geq 2$, there exist disjoint nonempty subsets $S_1, S_2 \subseteq S$ with $S_1 \cup S_2 = S$ such that*

$$|\operatorname{Edge}(S_1, S_2)| \leq \frac{2\sqrt{d}}{\alpha} \cdot \min\big\{|S_1|, |S_2|\big\}. \tag{2.6}$$

Proof. If there are no edges in the subgraph induced by S, an arbitrary choice of S_1 and S_2 satisfies Inequality (2.6). We assume that there is at least one edge in $\operatorname{Edge}(S)$.

Let $L \in \mathbb{R}^+$ and $b_1, b_2, \ldots, b_d \in \mathbb{R}$ be such that all points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ are contained inside the hypercube

$$[b_1, b_1 + L] \times [b_2, b_2 + L] \times \cdots \times [b_d, b_d + L].$$

We pick an integer $r \in [d]$ and a real number $t \in [0, L]$ uniformly at random, and define

$$S_1 = \Big\{\boldsymbol{v}_j \in S : \text{the } r\text{th coordinate of } \boldsymbol{v}_j \leq b_r + t\Big\},$$
$$S_2 = \Big\{\boldsymbol{v}_j \in S : \text{the } r\text{th coordinate of } \boldsymbol{v}_j > b_r + t\Big\}.$$

Next, we analyze the random sets S_1 and S_2. We first consider a fixed edge $\big(\{j_1, j_2\}; i_0\big) \in \operatorname{Edge}(S)$. Suppose that the vector $\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}$ is $(u_1, u_2, \ldots, u_d)$, where $u_i \in \mathbb{R}$. Then there is

$$|u_{i_0}| \geq \alpha \|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2 = \alpha \sqrt{\sum_{i=1}^{d} u_i^2} \geq \frac{\alpha}{\sqrt{d}} \sum_{i=1}^{d} |u_i|. \tag{2.7}$$

By the construction of S_1 and S_2, for every $i \in [d]$ we have

$$\Pr\Big[r = i \text{ and } \big(\{j_1, j_2\}; i_0\big) \in \mathrm{Edge}(S_1, S_2)\Big] = \frac{1}{d} \cdot \frac{|u_i|}{L}. \tag{2.8}$$

Using Inequality (2.7) and Equation (2.8),

$$\begin{aligned}\Pr\Big[\big(\{j_1, j_2\}; i_0\big) \in \mathrm{Edge}(S_1, S_2)\Big] &= \sum_{i=1}^{d} \left(\frac{1}{d} \cdot \frac{|u_i|}{L}\right) \le \frac{1}{dL} \cdot \frac{\sqrt{d}}{\alpha} |u_{i_0}| \\ &= \frac{\sqrt{d}}{\alpha} \cdot \Pr\Big[r = i_0 \text{ and } \big(\{j_1, j_2\}; i_0\big) \in \mathrm{Edge}(S_1, S_2)\Big].\end{aligned}$$

Consider this inequality for all edges $\big(\{j_1, j_2\}; i_0\big) \in \mathrm{Edge}(S)$. Define $\mathrm{Edge}_r(S_1, S_2) \subseteq \mathrm{Edge}(S_1, S_2)$ as the subset of the edges of the form $\big(\{j_1, j_2\}; r\big)$, where $\{j_1, j_2\} \in M_r$. Then

$$\frac{1}{2} \cdot \mathbb{E}\Big[|\,\mathrm{Edge}(S_1, S_2)|\Big] < \mathbb{E}\Big[|\,\mathrm{Edge}(S_1, S_2)|\Big] \le \frac{\sqrt{d}}{\alpha} \cdot \mathbb{E}\Big[|\,\mathrm{Edge}_r(S_1, S_2)|\Big].$$

In the first inequality, we used our assumption $\mathrm{Edge}(S) \neq \emptyset$, which implies that the expected value of $|\,\mathrm{Edge}(S_1, S_2)|$ is strictly positive. By the linearity of expectation, we have

$$\mathbb{E}\Big[\frac{\sqrt{d}}{\alpha}|\,\mathrm{Edge}_r(S_1, S_2)| - \frac{1}{2}|\,\mathrm{Edge}(S_1, S_2)|\Big] > 0.$$

Therefore there exist an integer $r \in [d]$ and a real number $t \in [0, L]$ such that

$$\frac{\sqrt{d}}{\alpha}|\,\mathrm{Edge}_r(S_1, S_2)| - \frac{1}{2}|\,\mathrm{Edge}(S_1, S_2)| > 0.$$

From this inequality we see that $|\,\mathrm{Edge}_r(S_1, S_2)|$ is strictly positive. Hence S_1 and S_2 are nonempty. Since the pairs in M_r are disjoint, we have $|\,\mathrm{Edge}_r(S_1, S_2)| \le \min\big\{|S_1|, |S_2|\big\}$. It follows that

$$|\,\mathrm{Edge}(S_1, S_2)| \le \frac{2\sqrt{d}}{\alpha} \min\big\{|S_1|, |S_2|\big\}. \qquad \square$$

2.3.2 Proof of Theorem 2.4

By Corollary 2.10, it suffices to consider simple 2-query approximate LDCs. We observe that, if the vector $\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}$ was in the standard direction $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$ for all $i \in [d]$ and $\{j_1, j_2\} \in M_i$, then for every $S \subseteq \mathcal{V}$ with $|S| \ge 2$, a hyperplane orthogonal to an arbitrary standard direction would "cut" S into two parts S_1 and S_2 such that

$$\mathrm{Edge}(S_1, S_2) \le \min\big\{|S_1|, |S_2|\big\},$$

with which one could derive an exponential lower bound by Corollary 2.12. (In this case, the code would be an "exact" LDC for which an exponential lower bound is known. However, we will need the cut in our actual proof.) Based on this, a simple idea is to round $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ to the vertices of a grid (with an appropriate grid distance) so that the vectors $\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}$ will become "upright". See Figure 2.2 for an illustration.

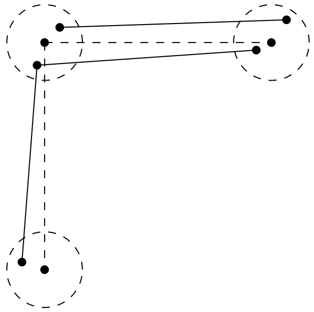

Figure 2.2: The center of every dashed circle is a grid point. We wish to round all points to their corresponding circle centers.

One easy way to do the rounding is to tile the space $\mathbb{R}^d$ with hypercubes, and move every $\boldsymbol{v}_i$ to the center of the hypercube that contains it. However, this does not work because the distances between the vertices of a hypercube can differ greatly (as large as a multiplicative factor $\sqrt{d}$), and it is impossible to find an appropriate grid distance that all (or many) points are rounded to their ideal positions.

It would be helpful if there was a way to tile the space $\mathbb{R}^d$ with "balls", so that during the rounding the points were moved by similar distances regardless of their directions. In [KORW12], the authors gave a randomized algorithm outputting a "spherical cube" such that by moving it along every integral vector $\boldsymbol{s} \in \mathbb{Z}^d$, one can exactly tile (completely cover without overlap) the space $\mathbb{R}^d$. With a simple scaling, this result also works for all positive grid distances other than 1. Let

$$\mathcal{G}_g = \left\{ g\boldsymbol{z} : \boldsymbol{z} \in \mathbb{Z}^d \right\}$$

denote the vertices of the grid with grid distance g, where $g > 0$. We restate the result of [KORW12] in terms of rounding:

Theorem 2.14 ([KORW12])**.** *For every $g > 0$, there exists a randomized algorithm that produces a mapping $R\colon \mathbb{R}^d \to \mathcal{G}_g$ with following properties:*

1. $R(\boldsymbol{x} + \boldsymbol{s}) = R(\boldsymbol{x}) + \boldsymbol{s}$ *for all* $\boldsymbol{x} \in \mathbb{R}^d$ *and* $\boldsymbol{s} \in \mathcal{G}_g$*;*

2. $\Pr\big[R(\boldsymbol{x}) \neq R(\boldsymbol{y})\big] \leq (2\pi/g) \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$*.*

In the following proof of Theorem 2.4, we will use the grid points $R(\boldsymbol{v}_j)$ without actually moving the points $\boldsymbol{v}_j$ to $R(\boldsymbol{v}_j)$. In fact, we will need multiple grids simultaneously with various grid distances, and different grids are used for different subsets $S \subseteq \mathcal{V}$.

Proof of Theorem 2.4. Suppose that $(\mathcal{V}, \mathcal{M})$ is a simple 2-query (α, δ)-approximate LDC of dimension d and length n. By Corollary 2.10, it suffices to prove

$$n = \exp\bigl(\Omega(\alpha\delta\sqrt{d})\bigr).$$

For an edge $\bigl(\{j_1, j_2\}; i\bigr) \in \mathcal{M}$ and a subset of real numbers $I \subseteq \mathbb{R}$, we say that the edge is *contained* in I if

$$\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2 \in I.$$

We define constants $\varepsilon = \sqrt{\alpha - \alpha_0} \in (0, 1)$ and $t = \lceil (\alpha - \alpha_0)^{-1.5} \rceil \in \mathbb{Z}^+$. Partition positive real numbers into disjoint intervals:

$$\mathbb{R}^+ = \bigcup_{\ell \in \mathbb{Z}} \Bigl[(1+\varepsilon)^{\ell}, (1+\varepsilon)^{\ell+1}\Bigr).$$

Let I_j $(j = 0, 1, \ldots, t-1)$ denote the union of the intervals with $\ell \equiv j \pmod{t}$, i.e.,

$$I_j = \bigcup_{k \in \mathbb{Z}} \Bigl[(1+\varepsilon)^{kt+j}, (1+\varepsilon)^{kt+j+1}\Bigr).$$

Then we have $\mathbb{R}^+ = I_0 \cup I_1 \cup \cdots \cup I_{t-1}$. Without loss of generality, we assume that I_0 is the one among $I_0, I_1, \ldots, I_{t-1}$ that contains the most number of edges. (Otherwise, scale the arrangement by $(1+\varepsilon)^j$ for some $j \in [t-1]$.) Let $\mathcal{M}' = (M'_1, M'_2, \ldots, M'_d)$, $M'_i \subseteq M_i$, denote the list of 2-matchings obtained by removing all edges not contained in I_0. We see that $(\mathcal{V}, \mathcal{M}')$ is a simple 2-query $(\alpha, \delta/t)$-approximate LDC. Next, we will restrict our attention to the code $(\mathcal{V}, \mathcal{M}')$.

For an edge $\bigl(\{j_1, j_2\}; i\bigr) \in \mathcal{M}'$, since the edge is contained in I_0, there must exist $k \in \mathbb{Z}$ such that

$$\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2 \in \Bigl[(1+\varepsilon)^{kt}, (1+\varepsilon)^{kt+1}\Bigr).$$

We call the integer k the *level* of $\bigl(\{j_1, j_2\}; i\bigr)$, and denote it by $\operatorname{level}\bigl(\{j_1, j_2\}; i\bigr)$.

Let $k_{\min}$ and $k_{\max}$ denote the minimum and maximum levels of the edges in $\mathcal{M}'$. For every integer $k \in [k_{\min}, k_{\max}]$, we define

$$g_k = \frac{(1+\varepsilon)^{kt} + (1+\varepsilon)^{kt+1}}{2\alpha} = \frac{(2+\varepsilon)(1+\varepsilon)^{kt}}{2\alpha},$$

and let $R_k \colon \mathbb{R}^d \to \mathcal{G}_{g_k}$ be the random mapping satisfying the properties of Theorem 2.14 for the grid $\mathcal{G}_{g_k}$. Note that for different values of k, the corresponding mappings R_k are generated independently.

We say that an edge $\bigl(\{j_1, j_2\}; i\bigr) \in \mathcal{M}'$ is *good* if the following properties are satisfied:

1. For every $k > \operatorname{level}\bigl(\{j_1, j_2\}; i\bigr)$, there is $R_k(\boldsymbol{v}_{j_1}) = R_k(\boldsymbol{v}_{j_2})$, i.e., $\boldsymbol{v}_{j_1}$ and $\boldsymbol{v}_{j_2}$ are mapped to the same grid point by R_k.

2. Let $k_0 = \mathrm{level}\big(\{j_1, j_2\}; i\big)$. Then either $R_{k_0}(\boldsymbol{v}_{j_2}) = R_{k_0}(\boldsymbol{v}_{j_1}) + g_{k_0}\boldsymbol{e}_i$ or $R_{k_0}(\boldsymbol{v}_{j_2}) = R_{k_0}(\boldsymbol{v}_{j_1}) - g_{k_0}\boldsymbol{e}_i$ holds, i.e., $\boldsymbol{v}_{j_1}$ and $\boldsymbol{v}_{j_2}$ are mapped to adjacent grid points along the direction $\boldsymbol{e}_i$ by R_{k_0}.

Claim 2.15. *Every edge $\big(\{j_1, j_2\}; i\big) \in \mathcal{M}'$ is good with probability at least* $21(\alpha - \alpha_0)$.

Proof. We use k_0 to denote $\mathrm{level}\big(\{j_1, j_2\}; i\big)$. There is

$$\frac{\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2}{g_{k_0}} \in \Big[\frac{2\alpha}{2+\varepsilon}, \frac{2\alpha(1+\varepsilon)}{2+\varepsilon}\Big) = \Big[\alpha - \frac{\alpha\varepsilon}{2+\varepsilon}, \alpha + \frac{\alpha\varepsilon}{2+\varepsilon}\Big) \subseteq \Big[\alpha - \frac{\varepsilon}{2}, \alpha + \frac{\varepsilon}{2}\Big).$$

We consider the two requirements of an good edge one by one, and calculate the probabilities that each of them is violated:

1. For every $k > k_0$, we have

$$\begin{aligned}\Pr\Big[R_k(\boldsymbol{v}_{j_1}) \neq R_k(\boldsymbol{v}_{j_2})\Big] &\leq \frac{2\pi}{g_k} \cdot \|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2 \\ &\leq 2\pi \cdot \Big(\alpha + \frac{\varepsilon}{2}\Big) \cdot \frac{g_{k_0}}{g_k} \\ &\leq 4\pi \cdot \frac{1}{(1+\varepsilon)^{(k-k_0)t}}.\end{aligned}$$

The probability that the first requirement is violated is at most

$$4\pi \cdot \sum_{k \geq k_0+1} \frac{1}{(1+\varepsilon)^{(k-k_0)t}} = 4\pi \cdot \frac{1}{(1+\varepsilon)^t - 1} \leq \frac{4\pi}{\varepsilon t}.$$

2. Without loss of generality, we assume $\langle \boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}, \boldsymbol{e}_i \rangle > 0$. (Otherwise interchange j_1 and j_2.) Then $\langle \boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}, \boldsymbol{e}_i \rangle \geq \alpha \|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2$. The probability that the second requirement is violated is at most

$$\begin{aligned}\Pr\Big[R_k(\boldsymbol{v}_{j_2}) \neq R_k(\boldsymbol{v}_{j_1}) + g_k\boldsymbol{e}_i\Big] &= \Pr\Big[R_k(\boldsymbol{v}_{j_2}) \neq R_k(\boldsymbol{v}_{j_1} + g_k\boldsymbol{e}_i)\Big] \\ &\leq \frac{2\pi}{g_k} \cdot \|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1} - g_k\boldsymbol{e}_i\|_2 \\ &= \frac{2\pi}{g_k} \cdot \sqrt{\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2^2 + g_k^2 - 2g_k \cdot \langle \boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}, \boldsymbol{e}_i \rangle} \\ &\leq \frac{2\pi}{g_k} \cdot \sqrt{\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2^2 + g_k^2 - 2g_k \cdot \alpha \|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2} \\ &= 2\pi \sqrt{\Big(\frac{\|\boldsymbol{v}_{j_2} - \boldsymbol{v}_{j_1}\|_2}{g_k} - \alpha\Big)^2 + 1 - \alpha^2} \\ &\leq 2\pi \sqrt{1 - \alpha^2 + \frac{\varepsilon^2}{4}}.\end{aligned}$$

Recall that $\alpha_0 = \sqrt{1 - 1/(4\pi^2)}$, $\varepsilon = \sqrt{\alpha - \alpha_0}$ and $t = \lceil (\alpha - \alpha_0)^{-1.5} \rceil \geq \varepsilon^{-3}$. Using the union bound, the probability that $\bigl(\{j_1, j_2\}; i\bigr)$ is a good edge is at least

$$\begin{aligned}
1 - \frac{4\pi}{\varepsilon t} - 2\pi\sqrt{1 - \alpha^2 + \frac{\varepsilon^2}{4}} &\geq 1 - 4\pi\varepsilon^2 - 2\pi\sqrt{1 - \Bigl(\alpha_0 + \varepsilon^2\Bigr)^2 + \frac{\varepsilon^2}{4}} \\
&\geq 1 - 4\pi\varepsilon^2 - 2\pi\sqrt{1 - \alpha_0^2 - 2\alpha_0\varepsilon^2 + \frac{\varepsilon^2}{4}} \\
&= 1 - \sqrt{1 - 4\pi\sqrt{4\pi^2 - 1} \cdot \varepsilon^2 + \pi^2\varepsilon^2} - 4\pi\varepsilon^2 \\
&\geq 2\pi\sqrt{4\pi^2 - 1} \cdot \varepsilon^2 - \frac{\pi^2\varepsilon^2}{2} - 4\pi\varepsilon^2 \\
&\geq 21\varepsilon^2 = 21(\alpha - \alpha_0). \qquad \blacksquare
\end{aligned}$$

By a standard expected value argument, there exist mappings $R_k \colon \mathbb{R}^d \to \mathcal{G}_{g_k}$ for every integer $k \in [k_{\min}, k_{\max}]$ such that at least $21(\alpha - \alpha_0)$ fraction of the edges are good. We fix such mappings. Let $\mathcal{M}'' = (M_1'', M_2'', \ldots, M_d'')$, $M_i'' \subseteq M_i'$, be the list of 2-matchings obtained by removing all edges that are not good. Then $(\mathcal{V}, \mathcal{M}'')$ is a simple 2-query (α, δ')-approximate LDC, where

$$\delta' \geq (\delta/t) \cdot 21(\alpha - \alpha_0) = \frac{21(\alpha - \alpha_0)}{\lceil (\alpha - \alpha_0)^{-1.5} \rceil} \cdot \delta \geq 20(\alpha - \alpha_0)^{2.5} \cdot \delta.$$

In the last step, we used $(\alpha - \alpha_0)^{-1.5} \geq (1 - \alpha_0)^{-1.5} > 600$. Now we consider the code $(\mathcal{V}, \mathcal{M}'')$.

Claim 2.16. *Consider the graph $G = (\mathcal{V}, \mathcal{M}'')$. For every $S \subseteq \mathcal{V}$ with $|S| \geq 2$, there exist disjoint nonempty $S_1, S_2 \subseteq S$ with $S_1 \cup S_2 = S$ such that*

$$|\operatorname{Edge}(S_1, S_2)| \leq \min\bigl\{|S_1|, |S_2|\bigr\}. \tag{2.9}$$

Proof. If there are no edges in the subgraph induced by S, an arbitrary choice of S_1 and S_2 satisfies Inequality (2.9). We only consider the case that Edge(S) is nonempty. Suppose that $\bigl(\{j_1^*, j_2^*\}; i^*\bigr) \in \operatorname{Edge}(S)$ is the edge with the maximum level and let $k_0 = \operatorname{level}\bigl(\{j_1^*, j_2^*\}; i^*\bigr)$. We will find S_1 and S_2 using the mapping R_{k_0}. Without loss of generality, assume $R_{k_0}(\boldsymbol{v}_{j_2^*}) = R_{k_0}(\boldsymbol{v}_{j_1^*}) + g_{k_0}\boldsymbol{e}_{i^*}$. (If this does not hold, there must be $R_{k_0}(\boldsymbol{v}_{j_2^*}) = R_{k_0}(\boldsymbol{v}_{j_1^*}) - g_{k_0}\boldsymbol{e}_{i^*}$ and we interchange j_1^*, j_2^*.) Partition S into S_1 and S_2 according to the i^*th coordinates of $R_{k_0}(\boldsymbol{v}_j)$:

$$\begin{aligned}
S_1 &= \Bigl\{\boldsymbol{v}_j \in S : \langle R_{k_0}(\boldsymbol{v}_j), \boldsymbol{e}_{i^*}\rangle \leq \langle R_{k_0}(\boldsymbol{v}_{j_1^*}), \boldsymbol{e}_{i^*}\rangle\Bigr\}, \\
S_2 &= \Bigl\{\boldsymbol{v}_j \in S : \langle R_{k_0}(\boldsymbol{v}_j), \boldsymbol{e}_{i^*}\rangle \geq \langle R_{k_0}(\boldsymbol{v}_{j_2^*}), \boldsymbol{e}_{i^*}\rangle\Bigr\}.
\end{aligned}$$

Clearly, S_1, S_2 are nonempty and $S_1 \cup S_2 = S$. It remains to prove Inequality (2.9).

We consider an edge $\bigl(\{j_1, j_2\}; i\bigr) \in \operatorname{Edge}(S)$. By the definition of good edges and $k_0 \geq \operatorname{level}\bigl(\{j_1, j_2\}; i\bigr)$, we see that there are two cases:

1. If $k_0 > \text{level}\big(\{j_1, j_2\}; i\big)$, there is $R_{k_0}(\boldsymbol{v}_{j_1}) = R_{k_0}(\boldsymbol{v}_{j_2})$ and $\boldsymbol{v}_{j_1}$, $\boldsymbol{v}_{j_2}$ must be in the same one of S_1, S_2. Hence $\big(\{j_1, j_2\}; i\big) \notin \text{Edge}(S_1, S_2)$.

2. If $k_0 = \text{level}\big(\{j_1, j_2\}; i\big)$, $R_{k_0}(\boldsymbol{v}_{j_1})$ and $R_{k_0}(\boldsymbol{v}_{j_2})$ differ only at the ith coordinate. In this case, $\big(\{j_1, j_2\}; i\big) \in \text{Edge}(S_1, S_2)$ only if $i = i^*$.

We have that for all edges $\big(\{j_1, j_2\}; i\big) \in \text{Edge}(S_1, S_2)$ there is $i = i^*$. Since the pairs $\{j_1, j_2\}$ in M_{i^*} are disjoint, Inequality (2.9) follows immediately. ■

Using Claim 2.16 and Corollary 2.12, we have

$$n \geq 4^{\delta' d} \geq 4^{20(\alpha-\alpha_0)^{2.5}\delta d}.$$

Therefore Theorem 2.4 is proved. □

2.4 A lower bound for general approximate LDCs

In this section we prove Theorem 2.5. Our idea is based on a result of [KT00] that gives a super-linear lower bound for q-query "exact" LDCs. We briefly sketch the proof in [KT00] (with our notations and assuming δ is a constant) as the following three steps:

1. Sample a random list of indices $S = (j_1, j_2, \dots, j_t) \in [n]^t$, where $t = \Theta\big(n^{\frac{q-1}{q}}\big)$;

2. Show that with high probability such an S contains q-tuples from $\Omega(d)$ different q-matchings M_i;

3. Since each of the above $\Omega(d)$ q-tuples is a set of codeword coordinates that determines a different entry of the original message, there must be $t = \Omega(d)$ by an information theoretic argument, which yields a lower bound $n = \Omega\big(d^{\frac{q}{q-1}}\big)$.

For approximate LDCs, noting that a q-tuple of codeword coordinates does not determine an entry of the original message, we will replace the third step with a spectral argument that shows the rank of the vectors $\boldsymbol{v}_{j_1}, \boldsymbol{v}_{j_2}, \dots, \boldsymbol{v}_{j_t}$ is $\Omega(d)$ (which also implies $t = \Omega(d)$).

We need the following variant of a well-known lemma (see for example [Alo09, BDWY13, DSW14b]) that gives a lower bound on the rank of diagonal dominating matrices:

Lemma 2.17. *Let D be any square matrix with positive real numbers on the diagonal. We have*

$$\text{rank}(D) \geq \frac{\text{tr}(D)^2}{\|D\|_F^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. Let r denote $\operatorname{rank}(D)$. We take a singular value decomposition of D:

$$D = P\Sigma Q^*,$$

where Σ is the diagonal matrix consists of singular values of D, and P, Q^* are unitary matrices. Say the positive singular values of D are $\sigma_1, \sigma_2, \ldots, \sigma_r$. The lemma follows immediately from

$$\operatorname{tr}(D)^2 = \operatorname{tr}\big(P\Sigma Q^*\big)^2 = \operatorname{tr}\big(\Sigma(Q^*P)\big)^2 \leq \left(\sum_{i=1}^{r} \sigma_i\right)^2 \leq r\sum_{i=1}^{r} \sigma_i^2 = r\|D\|_F^2. \qquad \square$$

We will also use [KT00, Lemma 5], which is restated below:

Lemma 2.18 ([KT00])**.** *Let $q \geq 2$ be a constant and M be a q-matching consists of γn disjoint q-tuples of $[n]$, where $\gamma < 1/q$. There exists an integer $t = \Theta\big(\gamma^{-\frac{1}{q}} n^{\frac{q-1}{q}}\big)$ such that if we sample a list $S = (j_1, j_2, \ldots, j_t) \in [n]^t$ uniformly at random,*

$$\Pr\Big[S \text{ contains a } q\text{-tuple in } M\Big] > \frac{3}{4}.$$

We now give the complete proof of Theorem 2.5:

Proof of Theorem 2.5. We first show that there are at least $q\delta d/2$ values of $i \in [d]$ such that $|M_i| \geq \delta n/2$. Assume the opposite. Then, since $|M_i| \leq n/q$ for every $i \in [d]$, there is

$$\delta dn \leq |M_1| + |M_2| + \cdots + |M_d| \leq \frac{q\delta d}{2} \cdot \frac{n}{q} + \left(d - \frac{q\delta d}{2}\right) \cdot \frac{\delta n}{2} = \delta dn - \frac{q\delta^2 dn}{4}.$$

We arrived at a contradiction. Therefore we can find at least $q\delta d/2$ q-matchings among $M_1, M_2, \ldots, M_d$ such that each of them has size at least $\delta n/2$.

Set $\gamma = \delta/2$ in Lemma 2.18. Using a standard expected value argument, we can see that there exists a list $S = (j_1, j_2, \ldots, j_t) \in [n]^t$ containing q-tuples from at least $(3/4) \cdot (q\delta d/2) = \Omega(\delta d)$ different q-matchings, where $t = \Theta\big(\delta^{-\frac{1}{q}} n^{\frac{q-1}{q}}\big)$. Let $M_{i_1}, M_{i_2}, \ldots, M_{i_{d'}}$ be these q-matchings, where $d' = \Omega(\delta d)$ and $i_1, i_2, \ldots, i_{d'} \in [d]$.

Next, we show $t \geq \alpha^2 d'$, which will immediately imply Theorem 2.5:

$$t = \Theta\big(\delta^{-\frac{1}{q}} n^{\frac{q-1}{q}}\big) \geq \alpha^2 d' = \Omega\big(\alpha^2 \delta d\big) \quad \Longrightarrow \quad n = \Omega\big(\delta^{\frac{q+1}{q-1}} (\alpha^2 d)^{\frac{q}{q-1}}\big).$$

For every $i \in \{i_1, i_2, \ldots, i_{d'}\}$, we can find a q-tuple $\big\{j_1^{(i)}, j_2^{(i)}, \ldots, j_q^{(i)}\big\} \subseteq S$ that is in M_i. By Definition 2.1, there is a vector

$$\boldsymbol{u}_i \in \operatorname{span}\Big\{\boldsymbol{v}_{j_1^{(i)}}, \boldsymbol{v}_{j_2^{(i)}}, \ldots, \boldsymbol{v}_{j_q^{(i)}}\Big\} \subseteq \operatorname*{span}_{j \in S}\big\{\boldsymbol{v}_j\big\}$$

with $\operatorname{weight}_i(\boldsymbol{u}_i) \geq \alpha$. Without loss of generality, we assume that $\boldsymbol{u}_i$ is a unit vector and the ith coordinate of $\boldsymbol{u}_i$ is at least α. (Otherwise the ith coordinate must be at most $-\alpha$, and we replace $\boldsymbol{u}_i$ with $-\boldsymbol{u}_i$.) We define $\boldsymbol{u}'_i \in \mathbb{R}^{d'}$ as the subvector of $\boldsymbol{u}_i$ consists of the coordinates with indices $i_1, i_2, \ldots, i_{d'}$. Let $D \in \mathbb{R}^{d' \times d'}$ be the matrix consists of columns $\boldsymbol{u}'_{i_1}, \boldsymbol{u}'_{i_2}, \ldots, \boldsymbol{u}'_{i_{d'}}$. Clearly, every entry on the diagonal of D is at least α, and every column of D has norm $\|\boldsymbol{u}'_i\|_2 \leq 1$. Using Lemma 2.17,

$$\begin{aligned} t &\geq \operatorname*{rank}_{j \in S}\big\{\boldsymbol{v}_j\big\} \geq \operatorname{rank}\big\{\boldsymbol{u}_{i_1}, \boldsymbol{u}_{i_2}, \ldots, \boldsymbol{u}_{i_{d'}}\big\} \geq \operatorname{rank}\big\{\boldsymbol{u}'_{i_1}, \boldsymbol{u}'_{i_2}, \ldots, \boldsymbol{u}'_{i_{d'}}\big\} \\ &= \operatorname{rank}(D) \geq \frac{(\alpha d')^2}{d'} = \alpha^2 d'. \end{aligned}$$

Thus the proof is finished. □

Chapter 3

Sylvester-Gallai for Arrangements of Subspaces

In this chapter, we prove the subspace version of the Sylvester-Gallai theorem. We first state our result in Section 3.1. Then we reduce our version of the Sylvester-Gallai theorem to a more convenient notion called (α, δ)-system in Section 3.2. Next, in Section 3.3 we prove the subspace version of a theorem of Barthe [Bar98], which is a key technique used in our proof. Finally, we prove our main theorem for (α, δ)-systems in Section 3.4, which implies the subspace version of the Sylvester-Gallai theorem. The results in this chapter are also included in [DH16].

3.1 Generalization of the Sylvester-Gallai theorem

The Sylvester-Gallai theorem states that for n points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^\ell$, if for every pair of points $\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}$ there is a third point $\boldsymbol{v}_{i_3}$ on the line passing through $\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}$, then all points must lie on a single line. This was first posed by Sylvester [Syl93], and was solved by Melchior [Mel40]. It was also conjectured independently by Erdös [Erd43] and proved shortly after by Gallai. We refer the reader to the survey [BM90] for more information about the history and various generalizations of this theorem. The complex version of this theorem was proved by Kelly [Kel86] (see also [EPS06, DSW14b] for alternative proofs) and states that if $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{C}^\ell$ and for every pair $\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}$ there is a third $\boldsymbol{v}_{i_3}$ on the same complex line, then all points are contained in some complex plane (there are planar examples and so this theorem is tight).

In [DSW14b] (based on earlier work in [BDWY13]), the following quantitative variant of the Sylvester-Gallai theorem was proved:

Theorem 3.1 ([DSW14b]). *Given n points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{C}^\ell$. Suppose that for every point $\boldsymbol{v}_{i_1}$ ($i_1 \in [n]$) there are at least $\delta(n-1)$ other points $\boldsymbol{v}_{i_2}$ ($i_2 \in [n] \setminus \{i_1\}$) such that there is a third point $\boldsymbol{v}_{i_3}$ ($i_3 \in [n] \setminus \{i_1, i_2\}$) on the the line passing through $\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}$. Then $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ are contained in an affine subspace of dimension at most $12/\delta$.*

The dependence on δ in the above theorem is asymptotically tight as one can place the n points on $1/\delta$ lines so that the dimension of the arrangement is $\Omega(1/\delta)$.

From here on, we will work with homogeneous subspaces (passing through zero) instead of affine subspaces (lines, planes, etc.). The difference is not crucial to our results and the affine version can always be derived by intersecting the homogeneous version with a generic hyperplane. In this setting, the above theorem will be stated for n one-dimensional subspaces (spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ respectively, where no two $\boldsymbol{v}_i$'s are multiples of each other) and collinearity of $\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}, \boldsymbol{v}_{i_3}$ is replaced with the three vectors being linearly dependent (i.e., contained in a two-dimensional subspace).

One natural high-dimensional variant of the Sylvester-Gallai theorem, studied in [Han65, BDWY13], replaces three-wise dependencies with t-wise dependencies for general values of t. We now raise another natural high-dimensional variant in which the *points* themselves are replaced with k-dimensional subspaces. We consider an arrangement of subspaces with many three-wise dependencies (defined appropriately) and prove that the entire arrangement lies in some low-dimensional space. Let $V_1, V_2, \ldots, V_n \subseteq \mathbb{C}^\ell$ be k-dimensional subspaces such that every pair $\{V_{i_1}, V_{i_2}\}$ satisfies $V_{i_1} \cap V_{i_2} = \{\mathbf{0}\}$. A dependency can then be defined as a triple $\{V_{i_1}, V_{i_2}, V_{i_3}\}$ contained in a single $2k$-dimensional subspace. The pairwise zero intersections guarantee that every pair of subspaces defines a unique $2k$-dimensional space (i.e., their span) and so, this definition of dependency behaves in a similar way to collinearity. For example, we have that if $V_{i_1}, V_{i_2}, V_{i_3}$ are dependent and $V_{i_2}, V_{i_3}, V_{i_4}$ are dependent then also $V_{i_1}, V_{i_2}, V_{i_4}$ are dependent. This would not hold if there were pairs with nonzero intersections. In fact, if nonzero intersections were allowed, we can construct an arrangement of two-dimensional subspaces with many dependent triples and with dimension as large as $\sqrt{n}$ (see below). We now state our main theorem, generalizing Theorem 3.1 (with slightly worse parameters) to the case $k > 1$:

Theorem 3.2. *Let $V_1, V_2, \ldots, V_n \subseteq \mathbb{C}^\ell$ be k-dimensional subspaces such that $V_i \cap V_{i'} = \{\mathbf{0}\}$ for all $i \neq i' \in [n]$. Suppose that for every $i_1 \in [n]$ there exists at least δn values of $i_2 \in [n] \setminus \{i_1\}$ such that $V_{i_1} + V_{i_2}$ contains some V_{i_3} with $i_3 \in [n] \setminus \{i_1, i_2\}$. Then*

$$\dim(V_1 + V_2 + \cdots + V_n) = O(k^4/\delta^2)^1,$$

where $O(\cdot)$ hides an absolute constant independent of δ, k or n.

In the statement of this theorem, we use the standard $V + U$ notation to denote the subspace spanned by all vectors in $V \cup U$, and for a set $S \subseteq \mathbb{C}^\ell$ we denote by $\dim(S)$ the smallest d such that S is contained in a d-dimensional subspace of $\mathbb{C}^\ell$.

Theorem 3.2 can be considered as a result for 2-query linear block LCCs over $\mathbb{C}$. See Section 1.1.2 for the connections. The condition $V_i \cap V_{i'} = \{\mathbf{0}\}$ is needed due to the following example: Set $k = 2$ and $n = \ell(\ell-1)/2$, and define the n subspaces to be $V_{ij} = \operatorname{span}\{\boldsymbol{e}_i, \boldsymbol{e}_j\}$, where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_\ell\}$ is the standard basis of $\mathbb{R}^\ell$. Then for

[1] In a recent work [DGOS16], the upper bound was improved to $O(k/\delta)$. This is optimal since one can always construct an arrangement with dimension $2k/\delta$ by partitioning the subspaces into $1/\delta$ groups, where each group is contained in a $2k$-dimensional space.

each $(i,j) \neq (i',j')$, the sum $V_{ij} + V_{i'j'}$ contains a third subspace (since the size of $\{i,j,i',j'\}$ is at least three). However, this arrangement has dimension $\ell > \sqrt{n}$.

3.1.1 Overview of the proof

A preliminary observation is that it suffices to prove the theorem over $\mathbb{R}$. This is because an arrangement of k-dimensional complex subspaces can be translated into an arrangement of $2k$-dimensional real subspaces (this will be proved at the end of Section 3.2). Hence, we will now focus on real arrangements.

The proof of the theorem is considerably simpler when the arrangement of subspaces $V_1, V_2, \ldots, V_n$ satisfies an extra "robustness" condition, namely that every two subspaces in a dependent triple have an angle bounded away from zero. More formally, for every two unit vectors $\boldsymbol{v}_1 \in V_{i_1}$ and $\boldsymbol{v}_2 \in V_{i_2}$ we have $|\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle| \leq 1 - \tau$ for some absolute constant $\tau > 0$. This condition implies that, when we have a dependency of the form $V_{i_3} \subseteq V_{i_1} + V_{i_2}$, every unit vector in V_{i_3} can be obtained as a linear combination with *bounded coefficients* (in absolute value) of unit vectors from V_{i_1}, V_{i_2}. Fixing an orthonormal basis for every subspace, we are able to construct many local linear dependencies among the basis vectors, as there are many three-wise dependencies among the subspaces. We then show (using the bound on the coefficients in the linear combinations) that the space of linear dependencies between all basis vectors, considered as a subspace of $\mathbb{R}^{kn}$, contains the rows of a $kn \times kn$ matrix that has large entries on the diagonal and small entries off the diagonal. Since matrices of this form have high rank (by a simple spectral argument), we conclude that the original set of basis vectors must have small dimension.

To handle the general case, by generalizing a theorem of Barthe [Bar98], we show that unless there exists some low-dimensional subspace W intersecting many of the subspaces V_i in the arrangement, we can find a change of basis that makes the angles between the subspaces large on average (in which case the previous argument works). This gives us the overall strategy of the proof: If such a W exists, we project W to zero and continue by induction. The loss in the overall dimension is bounded by the dimension of W, which can be chosen to be small enough. Otherwise (if such W does not exist) we apply the change of basis and use it to bound the dimension.

3.2 Reduction to (α, δ)-systems

We introduce the notion of (α, δ)-system, which is used to "organize" the dependent triples in an arrangement of subspaces in a more convenient form. In an (α, δ)-system, every subspace appears in many triples and every pair of subspaces appears together only in a few triples.

Definition 3.3 ((α, δ)-system)**.** Given a list of subspaces $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, we call a list of index sets $\mathcal{S} = (S_1, S_2, \ldots, S_w)$, $S_j \subseteq [n]$, an *(α, δ)-system* ($\alpha \in \mathbb{Z}^+, \delta > 0$) of $\mathcal{V}$ if

1. $|S_j| = 2$ or 3 for every $j \in [w]$;

2. If $|S_j| = 2$, say $S_j = \{i_1, i_2\}$, we have $V_{i_1} = V_{i_2}$;

   If $|S_j| = 3$, say $S_j = \{i_1, i_2, i_3\}$, we have $V_{i_1} \subseteq V_{i_2} + V_{i_3}$, $V_{i_2} \subseteq V_{i_1} + V_{i_3}$ and $V_{i_3} \subseteq V_{i_1} + V_{i_2}$;

3. Every $i \in [n]$ is contained in at least δn sets of $\mathcal{S}$;

4. Every pair $\{i_1, i_2\} \subseteq [n]$ $(i_1 \neq i_2)$ appears together in at most α sets of $\mathcal{S}$.

We note the following differences in the setting of (α, δ)-systems and that of the Sylvester-Gallai theorem for subspaces: (1) We allow dependent pairs as well as triples in (α, δ)-systems, as pairs might arise when we apply a linear map on the arrangement; (2) We allow $\delta > 1$ in (α, δ)-systems, whereas in the statement of the Sylvester-Gallai theorem we have $\delta \in [0, 1]$; (3) We only consider subspaces over real numbers in (α, δ)-systems, and we will show that the Sylvester-Gallai theorem for complex subspaces can be reduced to an (α, δ)-system for real subspaces.

We now state a few simple observations.

Lemma 3.4. *Let $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, be a list of subspaces that has an (α, δ)-system $\mathcal{S} = (S_1, S_2, \ldots, S_w)$, $S_j \subseteq [n]$. Then $\delta n^2/3 \leq w \leq \alpha n^2/2$ and $\delta/\alpha \leq 3/2$.*

Proof. We consider the sum $\sum_{j \in [w]} |S_j|$. By the definition of (α, δ)-systems,

$$n \cdot \delta n \leq \sum_{j \in [w]} |S_j| \leq 3w \quad \Longrightarrow \quad \delta n^2/3 \leq w.$$

Then we consider the number of pairs $\sum_{j \in [w]} \binom{|S_j|}{2}$,

$$w \leq \sum_{j \in [w]} \binom{|S_j|}{2} \leq \alpha \binom{n}{2} \leq \alpha n^2/2.$$

It follows that $\delta/\alpha \leq 3/2$. □

Lemma 3.5. *Given a list of subspaces $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, and a list of index sets $\mathcal{S} = (S_1, S_2, \ldots, S_w)$, $S_j \subseteq [n]$. If $w \geq \delta n^2$ and $\mathcal{S}$ satisfies the first, second and fourth requirements in Definition 3.3, there must exist a sublist $\mathcal{V}'$ of $\mathcal{V}$ and a sublist $\mathcal{S}'$ of $\mathcal{S}$ such that $|\mathcal{V}'| \geq \delta n/(2\alpha)$ and $\mathcal{S}'$ is an $(\alpha, \delta/2)$-system of $\mathcal{V}'$.*

Proof. We iteratively remove every V_i such that i appears in less than $\delta n/2$ index sets, and remove the index sets in which i appears. There were n subspaces in total, so eventually we remove at most $n \cdot \delta n/2$ index sets, and we have at least $\delta n^2 - \delta n^2/2 \geq \delta n^2/2 > 0$ remaining index sets. Since there are remaining index sets, there are also remaining subspaces. Let V_{i_1} be a remaining subspace. Because i_1 appears in at least $\delta n/2$ index sets and each pair $\{i_1, i_2\}$ $(i_2 \in [n] \setminus \{i_1\})$ appears in at most α index sets, there are at least $\delta n/(2\alpha)$ remaining subspaces. Let $\mathcal{V}'$ be the list of these subspaces and $\mathcal{S}'$ be the list of the remaining index sets. Changing the indices to $1, 2, \ldots, |\mathcal{V}'|$, we see that $\mathcal{S}'$ is an $(\alpha, \delta/2)$-system of $\mathcal{V}'$. □

Lemma 3.6. *Let $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, be a list of subspaces that has an (α, δ)-system $\mathcal{S} = (S_1, S_2, \dots, S_w)$, $S_j \subseteq [n]$. Then for any linear map $P\colon \mathbb{R}^\ell \to \mathbb{R}^\ell$, $\mathcal{S}$ is also an (α, δ)-system of $\mathcal{V}' = (V_1', V_2', \dots, V_n')$, where $V_i' = P(V_i)$.*

Proof. This is trivial since if $V_{i_1} \subseteq V_{i_2} + V_{i_3}$,

$$V_{i_1}' = P(V_{i_1}) \subseteq P(V_{i_2} + V_{i_3}) = P(V_{i_2}) + P(V_{i_3}) = V_{i_2}' + V_{i_3}'. \qquad \square$$

Lemma 3.7. *Let $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, be a list of subspaces that has an (α, δ)-system, $P\colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ be any linear map, and $\mathcal{V}' = (V_1', V_2', \dots, V_{n'}')$ be the list of nonzero (not $\{\mathbf{0}\}$) subspaces among $P(V_1), P(V_2), \dots, P(V_n)$. Suppose $\mathcal{V}'$ is nonempty. Then $\mathcal{V}'$ has an (α, δ')-system, where $\delta' = \delta n/n'$.*

Proof. By Lemma 3.6, we can find an (α, δ)-system $\mathcal{S} = (S_1, S_2, \dots, S_w)$, $S_j \subseteq [n]$, of the list $(P(V_1), P(V_2), \dots, P(V_n))$. For every $i \in [n]$ with $P(V_i) = \{\mathbf{0}\}$, we remove i from all S_j's that contain i. And let $S_1', S_2', \dots, S_w'$ denote the index sets after this procedure. We claim that for every $j \in [w]$, $|S_j'| = 0$, 2 or 3 and for $|S_j'| > 0$, S_j' satisfies the second requirement in Definition 3.3.

We first show that $|S_j'| \neq 1$. Assume there is some $j \in [w]$ with $|S_j'| = 1$, and say $S_j' = \{i\}$. Then we have $P(V_i) \subseteq \{\mathbf{0}\} + \{\mathbf{0}\}$ (if the original S_j contains three elements) or $P(V_i) = \{\mathbf{0}\}$ (if the original S_j contains two elements). In both cases we have $P(V_i) = \{\mathbf{0}\}$ and i should have been removed from S_j', which leads to a contradiction.

Next, if $|S_j'| = |S_j|$, S_j' satisfies the second requirement in Definition 3.3 since $S_j' = S_j$. It remains to consider the case that $|S_j'| = 2$ and $|S_j| = 3$. Say $S_j = \{i_1, i_2, i_3\}$ and $S_j' = \{i_1, i_2\}$ (i.e., $P(V_{i_3}) = \{\mathbf{0}\}$). We have $P(V_{i_1}) \subseteq P(V_{i_2}) + \{\mathbf{0}\}$ and $P(V_{i_2}) \subseteq P(V_{i_1}) + \{\mathbf{0}\}$. It follows immediately that $P(V_{i_1}) = P(V_{i_2})$. Therefore S_j' satisfies the second requirement in Definition 3.3.

Note that for every $i \in [n]$ with $P(V_i) \neq \{\mathbf{0}\}$, i appears in $\delta n = \delta' n'$ sets S_j'. We remove empty sets from $S_1', S_2', \dots, S_w'$. One can see that the remaining list of sets (with indices in them changed to $1, 2, \dots, n'$) is an (α, δ')-system of $\mathcal{V}'$. $\square$

Theorem 3.2 (the Sylvester-Gallai theorem for subspaces) will be derived from the following Theorem 3.8, which gives an dimension upper bound for (α, δ)-systems. We defer the proof of Theorem 3.8 to Section 3.4.

Theorem 3.8. *Suppose that $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, is a list of subspaces that has an (α, δ)-system, and $k \geq \dim(V_i)$ for every $i \in [n]$. Then*

$$\dim(V_1 + V_2 + \cdots + V_n) = O(\alpha^2 k^4/\delta^2),$$

where $O(\cdot)$ hides an absolute constant independent of α, δ, k and n.

Proof of Theorem 3.2 *using Theorem* 3.8. In this proof, we use $\mathrm{span}_{\mathbb{R}}$ to denote the span using real coefficients, and $\mathrm{span}_{\mathbb{C}}$ to denote the span using complex coefficients. For every $j \in [n]$, let $\{\boldsymbol{v}_{j1}, \boldsymbol{v}_{j2}, \dots, \boldsymbol{v}_{jk}\}$ be a basis of V_j and define

$$\widehat{V}_j = \mathrm{span}_{\mathbb{R}}\Big\{\mathrm{Re}(\boldsymbol{v}_{j1}), \mathrm{Re}(\boldsymbol{v}_{j2}), \dots, \mathrm{Re}(\boldsymbol{v}_{jk}), \mathrm{Im}(\boldsymbol{v}_{j1}), \mathrm{Im}(\boldsymbol{v}_{j2}), \dots, \mathrm{Im}(\boldsymbol{v}_{jk})\Big\}.$$

Claim 3.9. $\widehat{V_j} = \{\mathrm{Re}(\boldsymbol{v}) : \boldsymbol{v} \in V_j\}$ *for every* $j \in [n]$.

Proof. For every $\widehat{\boldsymbol{v}} \in \widehat{V_j}$, there exist $\lambda_1, \lambda_2, \ldots, \lambda_k, \mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$ such that

$$\begin{aligned}\widehat{\boldsymbol{v}} &= \sum_{s=1}^{k}\Big(\lambda_s \,\mathrm{Re}(\boldsymbol{v}_{js}) + \mu_s \,\mathrm{Im}(\boldsymbol{v}_{js})\Big) = \sum_{s=1}^{k}\Big(\lambda_s \,\mathrm{Re}(\boldsymbol{v}_{js}) + \mu_s \,\mathrm{Re}(-i\boldsymbol{v}_{js})\Big) \\ &= \mathrm{Re}\Big(\sum_{s=1}^{k}(\lambda_s - i\mu_s)\boldsymbol{v}_{js}\Big).\end{aligned}$$

Since $\lambda_s - i\mu_s$ ($s \in [k]$) can take all values in $\mathbb{C}$, the claim is proved. ■

We need the following claim from [BDWY13, Lemma 6] based on [Hil73, Theorem 4]:

Claim 3.10 ([BDWY13])**.** *Given a set A of $r \geq 3$ elements, we can construct a family of $r^2 - r$ triples of elements in A with the following properties: (1) Every triple contains three distinct elements; (2) Every element of A appears in exactly $3(r-1)$ triples; (3) Every pair of two distinct elements in A is contained together in at most 6 triples.*

We say a $2k$-dimensional vector space $U \subseteq \mathbb{C}^\ell$ is *special* if U contains at least three of $V_1, V_2, \ldots, V_n$. We define the *size* of a special space as the number of V_j's contained in it. For a special space U with size $r \geq 3$, we consider the indices of the r subspaces in U, and pick $r^2 - r$ triples of these indies with the properties in Claim 3.10. Let $\mathcal{S}$ be the family of the triples picked for all special spaces.

Claim 3.11. $\mathcal{S}$ *is a* $(6, 3\delta)$*-system of* $\mathcal{V} = (\widehat{V_1}, \widehat{V_2}, \ldots, \widehat{V_n})$.

Proof. For every triple $\{j_1, j_2, j_3\} \in \mathcal{S}$, we can see that $V_{j_1}, V_{j_2}, V_{j_3}$ are contained in the same $2k$-dimensional special space. And by $V_{j_1} \cap V_{j_2} = \{\mathbf{0}\}$, the space must be $V_{j_1} + V_{j_2}$ and hence $V_{j_3} \subseteq V_{j_1} + V_{j_2}$. By Claim 3.9,

$$\widehat{V_{j_3}} = \Big\{\mathrm{Re}(\boldsymbol{v}) : \boldsymbol{v} \in V_{j_3}\Big\} \subseteq \Big\{\mathrm{Re}(\boldsymbol{u}) + \mathrm{Re}(\boldsymbol{w}) : \boldsymbol{u} \in V_{j_1}, \boldsymbol{w} \in V_{j_2}\Big\} = \widehat{V_{j_1}} + \widehat{V_{j_2}}.$$

Similarly, $\widehat{V_{j_1}} \subseteq \widehat{V_{j_2}} + \widehat{V_{j_3}}$ and $\widehat{V_{j_2}} \subseteq \widehat{V_{j_1}} + \widehat{V_{j_3}}$.

Next, since for every $j_1 \in [n]$, there are at least δn values of $j_2 \in [n] \setminus \{j_1\}$ such that there is a special space containing V_{j_1} and V_{j_2}, the number of triples in $\mathcal{S}$ that contains j_1 is

$$\sum_{\substack{\text{special space } U \\ \text{s.t. } V_{j_1} \subseteq U}} 3\big(\mathrm{size}(U) - 1\big) = 3 \sum_{\substack{\text{special space } U \\ \text{s.t. } V_{j_1} \subseteq U}} \Big|\{j_2 \neq j_1 : V_{j_2} \subseteq U\}\Big| \geq 3\delta n.$$

Finally, every pair $\{j_1, j_2\} \subseteq [n]$ appears in at most 6 triples because V_{j_1}, V_{j_2} are contained in at most one special space and $\{j_1, j_2\}$ appears at most 6 times in the triples constructed from this special space. ■

By Theorem 3.8, there is $\dim\big(\widehat{V}_1+\widehat{V}_2+\cdots+\widehat{V}_n\big) = O(6^2(2k)^4/(3\delta)^2) = O(k^4/\delta^2)$. Then noting

$$\begin{aligned} V_1+V_2+\cdots+V_n &\subseteq \underset{j\in[n],s\in[k]}{\operatorname{span}_{\mathbb{C}}}\Big\{\mathrm{Re}(\boldsymbol{v}_{js}),\mathrm{Im}(\boldsymbol{v}_{js})\Big\}, \\ \widehat{V}_1+\widehat{V}_2+\cdots+\widehat{V}_n &= \underset{j\in[n],s\in[k]}{\operatorname{span}_{\mathbb{R}}}\Big\{\mathrm{Re}(\boldsymbol{v}_{js}),\mathrm{Im}(\boldsymbol{v}_{js})\Big\}, \end{aligned}$$

we have $\dim(V_1+V_2+\cdots+V_n) \le \dim(\widehat{V}_1+\widehat{V}_2+\cdots+\widehat{V}_n) = O(k^4/\delta^2)$. □

3.3 Barthe's theorem for subspaces

In this section, we generalize a theorem of Barthe [Bar98] from the one-dimensional version (points) to higher dimension (subspaces)[2]. Using this theorem, we can find a change of basis that brings a set of subspaces to "well-separated" positions (i.e., every two subspaces in a dependent triple have large angles), which a key step in our proof of Theorem 3.8 (see Section 3.1.1 for an overview of the entire proof). The one-dimensional version of Barthe's theorem was also proved in [DSW14a], which gives the first super-quadratic lower bound for 3-query LCCs. To state the subspace version of the theorem, we need the following definition:

Definition 3.12 (admissible basis set and vector)**.** Let $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$ be a list of subspaces. An index set $H \subseteq [n]$ is called a *$\mathcal{V}$-admissible basis set* if

$$\dim\Big(\sum_{i\in H} V_i\Big) = \sum_{i\in H}\dim(V_i) = \dim\Big(\sum_{i\in[n]} V_i\Big),$$

i.e., the subspaces with indices in H span the entire $\sum_{i\in[n]} V_i$, and every subspace with index in H has intersection $\{\mathbf{0}\}$ with the span of the other subspaces with indices in H.

A *$\mathcal{V}$-admissible basis vector* is the indicator vector $\mathbf{1}_H \in \{0,1\}^n$ of some $\mathcal{V}$-admissible basis set H.

The subspace version of Barthe's theorem is as following:

Theorem 3.13. *Given a list of subspaces $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, with $V_1 + V_2 + \cdots + V_n = \mathbb{R}^\ell$ and a vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{R}^n$ in the convex hull of all $\mathcal{V}$-admissible basis vectors. Then there exists an invertible linear map $M\colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ such that*

$$\sum_{i=1}^n p_i \operatorname{Proj}_{M(V_i)} = I_{\ell\times\ell},$$

where $M(V_i)$ is the subspace obtained by applying M on V_i, and $\operatorname{Proj}_{M(V_i)}$ denotes the orthogonal projection matrix onto $M(V_i)$.

[2] Following the initial publication of this work in an earlier version of [DH16], it was brought to our attention that Bennet et al. [BCCT08] already proved a high-dimensional version of Barthe's result and, in fact, our generalization could also be derived (with some work) from their results.

The proof of the one-dimensional case in [Bar98] proceeds by defining a strictly convex function $f(t_1, \dots, t_m)$ on $\mathbb{R}^m$ and shows that the function is bounded. This means that there must exist a maximum point at which all partial derivatives of f vanish. Solving the resulting equations gives the required invertible map. We follow a similar strategy, defining an appropriate bounded function $f(t_1, \dots, t_m, R_1, \dots, R_n)$ with more variables, where the extra variables $R_1, \dots, R_n$ represent the action of the orthogonal group $\mathbf{O}(k)$ on each of the subspaces. However, in our case, we cannot show that f is strictly convex and so a maximum might not exist. Instead we show that there exists a point at which all partial derivatives are very small (smaller than any $\epsilon > 0$), which is sufficient for our purposes.

We introduce the function $f(t_1, \dots, t_m, R_1, \dots, R_n)$ in Section 3.3.1, and show that there is a point with very small partial derivatives in Section 3.3.2. Then we prove Theorem 3.13 (the subspace version of Barthe's theorem) in Section 3.3.3. Lastly, in Section 3.3.4 we state a convenient variant of the theorem that will be used later in the proof of our main result Theorem 3.8.

3.3.1 The function and basic properties

We use the notations $\mathcal{V} = (V_1, V_2, \dots, V_n)$ and $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$ that are introduced in the statement of Theorem 3.13. Let $k_1, k_2, \dots, k_n$ be the dimensions of $V_1, V_2, \dots, V_n$ respectively and $m = k_1 + k_2 + \cdots + k_n$. For every $i \in [n]$, we fix $\{\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \dots, \boldsymbol{v}_{ik_i}\}$ to be some basis of V_i (not necessarily orthonormal). A set $I \subseteq [m]$ is called a *good basis set* if

$$I = \bigcup_{i \in H} \Big\{(i,1), (i,2), \dots, (i,k_i)\Big\}$$

for some $\mathcal{V}$-admissible basis set H. We can see that for any good basis set I, the set $\{\boldsymbol{v}_{ij} : (i,j) \in I\}$ is a basis of $\mathbb{R}^\ell$. Throughout our proof, we identify $[m]$ with pairs (i,j), where $i \in [n]$, $j \in [k_i]$, and for a vector $\boldsymbol{a} \in \mathbb{R}^m$ we use a_{ij} to denote the entry at position $\sum_{i' < i} k_{i'} + j$. Let $\boldsymbol{\gamma} \in \mathbb{R}^m$ be the vector with

$$\gamma_{ij} = p_i \quad \forall i \in [n], j \in [k_i]. \tag{3.1}$$

For a list of vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_q$, we use $[\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_q]$ to denote the matrix consisting of columns $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_q$. Let $\mathbf{O}(s)$ be the group of $s \times s$ orthogonal matrices. Define $X \colon \mathbb{R}^m \times \mathbf{O}(k_1) \times \cdots \times \mathbf{O}(k_n) \to \mathbb{R}^{\ell \times \ell}$ as the following matrix valued function:

$$X(\boldsymbol{t}, R_1, \dots, R_n) = \sum_{i \in [n], j \in [k_i]} e^{t_{ij}} \boldsymbol{x}_{ij} {\boldsymbol{x}_{ij}}^{\mathsf{T}}, \tag{3.2}$$

where for every $i \in [n]$ the vectors $\boldsymbol{x}_{ij}$ are given by

$$[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \dots, \boldsymbol{x}_{ik_i}] = [\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \dots, \boldsymbol{v}_{ik_i}] \cdot R_i. \tag{3.3}$$

We note that for $i \in [n]$, $j \in [k_i]$, $\boldsymbol{x}_{ij}$ is a function of R_i and $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik_i}\}$ is also a basis of V_i. Define $f \colon \mathbb{R}^m \times \mathbf{O}(k_1) \times \cdots \times \mathbf{O}(k_n) \to \mathbb{R}$ as

$$f(\boldsymbol{t}, R_1, \ldots, R_n) = \langle \boldsymbol{\gamma}, \boldsymbol{t} \rangle - \ln \det(X).$$

The next lemma shows that the function f is bounded from above over its domain. The proof is similar to [Bar98, Proposition 3].

Lemma 3.14. *There exists a constant $C \in \mathbb{R}$ such that $f(\boldsymbol{t}, R_1, \ldots, R_n) \leq C$ for all values of $\boldsymbol{t}, R_1, \ldots, R_n$.*

Proof. In this proof, we use $\mathcal{F} = \binom{[m]}{\ell}$ to denote the family of all ℓ-element subsets of $[m]$. For $I \subseteq [m]$, we use $\mathbf{1}_I \in \{0,1\}^m$ to denote the indicator vector of I. Since $\boldsymbol{p}$ is in the convex hull of all $\mathcal{V}$-admissible basis vectors, we can write $\boldsymbol{\gamma}$ (defined in Equation (3.1)) as a convex combination of indicator vectors of good basis sets:

$$\boldsymbol{\gamma} = \sum_{I \in \mathcal{F}} \mu_I \mathbf{1}_I, \tag{3.4}$$

where $\mu_I \in [0,1]$, $\sum \mu_I = 1$ and $\mu_I \neq 0$ only if I is a good basis set.

In the proof, we will use the Cauchy-Binet formula (see [BW89, Section 4.6]) which states that for an $\ell \times m$ matrix A and an $m \times \ell$ matrix B,

$$\det(AB) = \sum_{I \in \mathcal{F}} \det(A_I) \det(B_I), \tag{3.5}$$

where A_I denotes the $\ell \times \ell$ submatrix of A consisting of columns with indices in I, and B_I denotes the $\ell \times \ell$ submatrix of B consisting of rows with indices in I.

For $I \in \mathcal{F}$, we use L_I to denote the $\ell \times \ell$ matrix consists of columns $\boldsymbol{x}_{ij}$ for all $(i,j) \in I$. By Equation (3.5),

$$\begin{aligned}
\det(X) &= \det\Bigg(\sum_{i \in [n], j \in [k_i]} e^{t_{ij}} \boldsymbol{x}_{ij} {\boldsymbol{x}_{ij}}^{\mathsf{T}} \Bigg) \\
&= \det\Big([\boldsymbol{x}_{11}, \ldots, \ldots, \ldots, \boldsymbol{x}_{nk_n}] \cdot [e^{t_{11}} \boldsymbol{x}_{11}, \ldots, \ldots, \ldots, e^{t_{nk_n}} \boldsymbol{x}_{nk_n}]^{\mathsf{T}} \Big) \\
&= \sum_{I \in \mathcal{F}} \Big(\det(L_I) \cdot \det({L_I}^{\mathsf{T}}) \prod_{(i,j) \in I} e^{t_{ij}} \Big) \\
&= \sum_{I \in \mathcal{F}} e^{\langle \boldsymbol{t}, \mathbf{1}_I \rangle} \det(L_I)^2.
\end{aligned}$$

Then by the weighted AM-GM inequality,

$$\det(X) \geq \sum_{\substack{I \subseteq \mathcal{F}: \\ \mu_I \neq 0}} \mu_I \cdot \frac{e^{\langle \boldsymbol{t}, \mathbf{1}_I \rangle}}{\mu_I} \det(L_I)^2 \geq \prod_{\substack{I \subseteq \mathcal{F}: \\ \mu_I \neq 0}} \left(\frac{e^{\langle \boldsymbol{t}, \mathbf{1}_I \rangle} \det(L_I)^2}{\mu_I} \right)^{\mu_I}.$$

Using Equation (3.4),

$$\det(X) \geq e^{\langle \boldsymbol{\gamma}, \boldsymbol{t} \rangle} \cdot \prod_{\substack{I \in \mathcal{F}: \\ \mu_I \neq 0}} \left(\frac{\det(L_I)^2}{\mu_I} \right)^{\mu_I}.$$

Take the logarithm of both sides, and we are able to cancel the variable $\boldsymbol{t}$ in f:

$$f(\boldsymbol{t}, R_1, \ldots, R_n) = \langle \boldsymbol{\gamma}, \boldsymbol{t} \rangle - \ln \det(X) \leq \sum_{\substack{I \in \mathcal{F}: \\ \mu_I \neq 0}} \mu_I \ln\left(\frac{\mu_I}{\det(L_I)^2} \right). \tag{3.6}$$

Since I is a good basis set for $\mu_I \neq 0$, the matrices L_I in the denominators have full rank. Recall that the columns $\boldsymbol{x}_{ij}$ in L_I are functions of the orthogonal matrices $R_1, \ldots, R_n$ (see Equation (3.3)). We can see that the right side of Equation (3.6) is a well-defined continuous function over the compact set $\mathbf{O}(k_1) \times \cdots \times \mathbf{O}(k_n)$. Therefore there exists a finite upper bound for f. □

3.3.2 Finding a point with small derivatives

The matrix X as defined in Equation (3.2) is always positive definite, since for any $\boldsymbol{w} \neq \mathbf{0}$,

$$\boldsymbol{w}^{\mathsf{T}} X \boldsymbol{w} = \sum_{i \in [n], j \in [k_i]} e^{t_{ij}} \langle \boldsymbol{x}_{ij}, \boldsymbol{w} \rangle^2 > 0,$$

provided that $\boldsymbol{x}_{11}, \ldots, \ldots, \boldsymbol{x}_{nk_n}$ span the entire space $\mathbb{R}^\ell$ (guaranteed by $V_1 + V_2 + \cdots + V_n = \mathbb{R}^\ell$). Let $M \colon \mathbb{R}^m \times \mathbf{O}(k_1) \times \cdots \times \mathbf{O}(k_n) \to \mathbb{R}^{\ell \times \ell}$ be any invertible matrix satisfying

$$M^{\mathsf{T}} M = X^{-1}.$$

In a later part of the proof, we will show that the linear map defined by M "almost" satisfies the requirement of Theorem 3.13 when $\boldsymbol{t}$, $R_1, \ldots, R_n$ take appropriate values, and based on this, we can show the existence of the required linear map.

We first find a value of $(R_1, \ldots, R_n)$ for every $\boldsymbol{t} \in \mathbb{R}^m$ with some specific properties, and then find an appropriate value of $\boldsymbol{t}$.

Lemma 3.15. *For every $\boldsymbol{t} \in \mathbb{R}^m$, there exists $R_1^*(\boldsymbol{t}) \in \mathbf{O}(k_1), \ldots, R_n^*(\boldsymbol{t}) \in \mathbf{O}(k_n)$ satisfying*

1. $f\big(\boldsymbol{t}, R_1^*(\boldsymbol{t}), \ldots, R_n^*(\boldsymbol{t})\big) = \max\limits_{R_1, \ldots, R_n} \big\{ f(\boldsymbol{t}, R_1, \ldots, R_n) \big\}$*;*

2. *For every $i \in [n]$, if $t_{ij} = t_{ij'}$ for some $j \neq j' \in [k_i]$, then*

   $$\langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'} \rangle = 0,$$

   where $\boldsymbol{x}_{ij}$, $\boldsymbol{x}_{ij'}$ and M denotes their values at $\big(\boldsymbol{t}, R_1^*(\boldsymbol{t}), \ldots, R_n^*(\boldsymbol{t})\big)$.

Proof. The first condition can be satisfied because of the compactness of $\mathbf{O}(k_1) \times \cdots \times \mathbf{O}(k_n)$. We will show how to change $\big(R_1^*(\boldsymbol{t}), \ldots, R_n^*(\boldsymbol{t})\big)$, which already satisfies the first condition, so that it also satisfies the second condition.

Fix an $i \in [n]$. We partition the indices of $t_{i1}, t_{i2}, \ldots, t_{ik_i}$ into equivalence classes $J_1, \ldots, J_b \subseteq [k_i]$ such that $t_{ij} = t_{ij'}$ for j, j' in the same class and $t_{ij} \neq t_{ij'}$ for j, j' in different classes. We use t_{J_r} to denote the value of t_{ij} for $j \in J_r$, and L_{J_r} to denote the matrix consisting of all columns $\boldsymbol{x}_{ij}$ with $j \in J_r$.

The terms in X that depend on R_i are

$$\sum_{r\in[b]} \left(e^{t_{J_r}} \sum_{j \in J_r} \boldsymbol{x}_{ij} {\boldsymbol{x}_{ij}}^{\mathsf{T}} \right) = \sum_{r\in[b]} \big(e^{t_{J_r}} \cdot L_{J_r} {L_{J_r}}^{\mathsf{T}} \big) = \sum_{r\in[b]} \big(e^{t_{J_r}} \cdot L_{J_r} Q_r {Q_r}^{\mathsf{T}} {L_{J_r}}^{\mathsf{T}} \big),$$

where Q_r can take any $|J_r| \times |J_r|$ orthogonal matrix. Hence if we change the variable $R_i^*(\boldsymbol{t})$ to $R_i^*(\boldsymbol{t})\cdot\operatorname{diag}(Q_1, \ldots, Q_b)$, or equivalently, change L_{J_r} to $L_{J_r}Q_r$ for every $r \in [b]$, the matrix X will not change, which means M and f will not change. Next, we find appropriate $Q_1, \ldots, Q_b$ and apply the change.

For every $r \in [b]$, we claim that there exists an orthogonal matrix Q_r such that the columns of $ML_{J_r}Q_r$ are orthogonal. In fact, take a singular value decomposition of ML_{J_r}:

$$ML_{J_r} = P\Sigma Q^{\mathsf{T}},$$

where P is an $\ell \times \ell$ orthogonal matrix, Σ is $\ell \times |J_r|$ matrix with nonzero entries only on the diagonal, and Q is an $|J_r| \times |J_r|$ orthogonal matrix. It suffices to take $Q_r = Q$.

We find such $Q_1, \ldots, Q_b$ and change $R_i^*(\boldsymbol{t})$ to $R_i^*(\boldsymbol{t}) \cdot \operatorname{diag}(Q_1, \ldots, Q_b)$. One can see that the second condition of Lemma 3.15 is satisfied while the matrix M and the value of f (which is still the maximum) are preserved. Doing this for every $i \in [n]$, we obtain the required $\big(R_1^*(\boldsymbol{t}), \ldots, R_n^*(\boldsymbol{t})\big)$. □

Next, we show that there is a $\boldsymbol{t}^* \in \mathbb{R}^m$ at which all partial derivatives are small:

Lemma 3.16. *For any $\varepsilon > 0$, there exists $\boldsymbol{t}^* \in \mathbb{R}^m$ such that for all $i \in [n]$, $j \in [k_i]$,*

$$\left| \frac{\partial f}{\partial t_{ij}} \big(\boldsymbol{t}^*, R_1^*(\boldsymbol{t}^*), \ldots, R_n^*(\boldsymbol{t}^*) \big) \right| \leq \varepsilon,$$

where for every $\boldsymbol{t} \in \mathbb{R}^m$, $\big(R_1^*(\boldsymbol{t}), \ldots, R_n^*(\boldsymbol{t})\big)$ denote an arbitrary choice of orthogonal matrices that satisfy the requirements of Lemma 3.15*.*

This lemma follows immediately from the following more general lemma, as we have shown that $f(\boldsymbol{t}, R_1, \ldots, R_n)$ is bounded from above in Lemma 3.14.

Lemma 3.17. *Let $\mathcal{A} \subseteq \mathbb{R}^h$ ($h \in \mathbb{Z}^+$) be a compact set. Suppose $f\colon \mathbb{R}^m \times \mathcal{A} \to \mathbb{R}$ and $y^*\colon \mathbb{R}^m \to \mathcal{A}$ are functions satisfying the following properties:*

1. *f is bounded from above and continuous on $\mathbb{R}^m \times \mathcal{A}$;*

2. *For every $\boldsymbol{x} \in \mathbb{R}^m$, $f\big(\boldsymbol{x}, y^*(\boldsymbol{x})\big) = \max_{y\in\mathcal{A}}\big\{f(\boldsymbol{x}, y)\big\}$;*

3. *For every fixed $y \in \mathcal{A}$, $f(\boldsymbol{x}, y)$ as a function of $\boldsymbol{x}$ is differentiable on $\mathbb{R}^m$.*

Then, for every $\varepsilon > 0$, there exists an $\boldsymbol{x}^* \in \mathbb{R}^m$ such that for every $i \in [m]$,

$$\left| \frac{\partial f}{\partial x_i} \big(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*) \big) \right| \le \varepsilon.$$

Proof. We use $f^*(\boldsymbol{x})$ to denote $f(\boldsymbol{x}, y^*(\boldsymbol{x}))$ in this proof. For the sake of contradiction, assume that for any $\boldsymbol{x} \in \mathbb{R}^m$, there is an index $i \in [m]$ such that

$$\left| \frac{\partial f}{\partial x_i} \big(\boldsymbol{x}, y^*(\boldsymbol{x}) \big) \right| > \varepsilon.$$

It follows that there exists $\boldsymbol{x}' \neq \boldsymbol{x}$ with

$$f^*(\boldsymbol{x}') - f^*(\boldsymbol{x}) \ge f\big(\boldsymbol{x}', y^*(\boldsymbol{x})\big) - f\big(\boldsymbol{x}, y^*(\boldsymbol{x})\big) \ge 0.9\varepsilon \cdot \|\boldsymbol{x}' - \boldsymbol{x}\|_2.$$

We consider the following nonempty set for $\boldsymbol{x} \in \mathbb{R}^m$:

$$\mathcal{G}(\boldsymbol{x}) = \Big\{ \boldsymbol{x}' \in \mathbb{R}^m : f^*(\boldsymbol{x}') - f^*(\boldsymbol{x}) \ge 0.9\varepsilon \cdot \|\boldsymbol{x}' - \boldsymbol{x}\|_2 > 0 \Big\}.$$

Claim 3.18. *For any $\boldsymbol{x}_0 \in \mathbb{R}^m$, $\boldsymbol{x}_1 \in \mathcal{G}(\boldsymbol{x}_0)$ and $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_1)$, we have $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_0)$.*

Proof. By the definitions of $\mathcal{G}(\boldsymbol{x}_0)$ and $\mathcal{G}(\boldsymbol{x}_1)$,

$$\begin{aligned} f^*(\boldsymbol{x}_2) - f^*(\boldsymbol{x}_0) &= \big(f^*(\boldsymbol{x}_2) - f^*(\boldsymbol{x}_1)\big) + \big(f^*(\boldsymbol{x}_1) - f^*(\boldsymbol{x}_0)\big) \\ &\ge 0.9\varepsilon \cdot \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 + 0.9\varepsilon \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|_2 \\ &\ge 0.9\varepsilon \cdot \|\boldsymbol{x}_2 - \boldsymbol{x}_0\|_2. \end{aligned}$$

We conclude the proof by noting $\boldsymbol{x}_2 \neq \boldsymbol{x}_0$, which follows from $f^*(\boldsymbol{x}_2) > f^*(\boldsymbol{x}_1) > f^*(\boldsymbol{x}_0)$. ■

Claim 3.19. *For any $\boldsymbol{x}_0 \in \mathbb{R}^m$, there exists some $\boldsymbol{x}_1 \in \mathcal{G}(\boldsymbol{x}_0)$ with the following properties: (1) $\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|_2 \ge \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ for any $\boldsymbol{x} \in \mathcal{G}(\boldsymbol{x}_0)$; (2) $f^*(\boldsymbol{x}_1) \ge f^*(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^m$ with $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 = \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|_2$.*

Proof. Fix $\boldsymbol{x}_0 \in \mathbb{R}^m$. We define

$$g(\boldsymbol{x}, y) = f(\boldsymbol{x}, y) - f^*(\boldsymbol{x}_0) - 0.9\varepsilon \cdot \|\boldsymbol{x} - \boldsymbol{x}_0\|_2,$$

and

$$\widetilde{\mathcal{G}} = \Big\{ (\boldsymbol{x}, y) \in \mathbb{R}^m \times \mathcal{A} : g(\boldsymbol{x}, y) \ge 0 \Big\} = g^{-1}\big([0, +\infty)\big).$$

Clearly, $\widetilde{\mathcal{G}} \neq \emptyset$. We show that $\widetilde{\mathcal{G}}$ is compact. Since $f(\boldsymbol{x}, y)$ is bounded from above, we have $g(\boldsymbol{x}, y) < 0$ for any $(\boldsymbol{x}, y)$ with sufficiently large $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2$. Hence $\widetilde{\mathcal{G}}$ is bounded. The function $g(\boldsymbol{x}, y)$ is continuous since $f(\boldsymbol{x}, y)$ is continuous. Hence

$\widetilde{\mathcal{G}} = g^{-1}\big([0,+\infty)\big)$ is closed. Therefore $\widetilde{\mathcal{G}}$ is compact, and there exists

$$Z = \max_{(\boldsymbol{x},y)\in\widetilde{\mathcal{G}}} \big\{\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\big\}.$$

We have $Z > 0$, because for any $\boldsymbol{x} \in \mathcal{G}(\boldsymbol{x}_0)$, there is $\big(\boldsymbol{x}, y^*(\boldsymbol{x})\big) \in \widetilde{\mathcal{G}}$ and $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 > 0$.

We consider the compact set

$$\mathcal{B} = \Big\{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}-\boldsymbol{x}_0\|_2 = Z\Big\} \times \mathcal{A},$$

and let $\big(\boldsymbol{x}_1, y^*(\boldsymbol{x}_1)\big) \in \mathcal{B}$ be any point where f is maximized over $\mathcal{B}$. To prove the claim, it suffices to show $\boldsymbol{x}_1 \in \mathcal{G}(\boldsymbol{x}_0)$. In fact, pick any $(\boldsymbol{x}, y) \in \widetilde{\mathcal{G}}$ with $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 = Z$ and there is

$$f^*(\boldsymbol{x}_1) - f^*(\boldsymbol{x}_0) \geq f(\boldsymbol{x}, y) - f^*(\boldsymbol{x}_0) \geq 0.9\varepsilon \cdot \|\boldsymbol{x}-\boldsymbol{x}_0\|_2 = 0.9\varepsilon \cdot \|\boldsymbol{x}_1-\boldsymbol{x}_0\|_2 > 0. \quad \blacksquare$$

We fix $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^m$ such that $\boldsymbol{x}_1 \in \mathcal{G}(\boldsymbol{x}_0)$, $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_1)$ and $\boldsymbol{x}_1$ satisfies the properties in Claim 3.19. By Claim 3.18, we have $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_0)$. Since $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_1)$, we have $f^*(\boldsymbol{x}_2) > f^*(\boldsymbol{x}_1)$. Combining $\boldsymbol{x}_2 \in \mathcal{G}(\boldsymbol{x}_0)$, $f^*(\boldsymbol{x}_2) > f^*(\boldsymbol{x}_1)$ with Claim 3.19, we can see $\|\boldsymbol{x}_2-\boldsymbol{x}_0\|_2 < \|\boldsymbol{x}_1-\boldsymbol{x}_0\|_2$. We consider the the compact set

$$\mathcal{C} = \Big\{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \leq \|\boldsymbol{x}_1-\boldsymbol{x}_0\|_2\Big\} \times \mathcal{A},$$

and let $\big(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*)\big) \in \mathcal{C}$ be any point where f is maximized over $\mathcal{C}$. We have $f^*(\boldsymbol{x}^*) \geq f^*(\boldsymbol{x}_2) > f^*(\boldsymbol{x}_1)$. Hence $\|\boldsymbol{x}^*-\boldsymbol{x}_0\|_2 < \|\boldsymbol{x}_1-\boldsymbol{x}_0\|_2$ by Claim 3.19. This means that $\boldsymbol{x} = \boldsymbol{x}^*$ is a local maximum of $f(\boldsymbol{x}, y^*(\boldsymbol{x}^*))$ with $y = y^*(\boldsymbol{x}^*)$ fixed. Therefore

$$\frac{\partial f}{\partial x_i}\big(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*)\big) = 0 \quad \forall i \in [m]. \qquad \square$$

3.3.3 Proof of Theorem 3.13

Before the proof, we need the following lemma:

Lemma 3.20. *Given any $\varepsilon > 0$. Suppose $\boldsymbol{t}^*, R_1^*(\boldsymbol{t}^*), \ldots, R_n^*(\boldsymbol{t}^*)$ satisfy the requirements of Lemma 3.15 and Lemma 3.16. Let M and $\boldsymbol{x}_{ij}$ $(i \in [n], j \in [k_i])$ denote their values at $\big(\boldsymbol{t}^*, R_1^*(\boldsymbol{t}^*), \ldots, R_n^*(\boldsymbol{t}^*)\big)$. Then we have $\langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle = 0$ for all $i \in [n]$ and $j \neq j' \in [k_i]$.*

Proof. In the proof we also use X to denote its value at $\big(\boldsymbol{t}^*, R_1^*(\boldsymbol{t}^*), \ldots, R_n^*(\boldsymbol{t}^*)\big)$. We fix $i \in [n]$, $j \neq j' \in [k_i]$, and prove $\langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle = 0$. If $t_{ij}^* = t_{ij'}^*$, this is guaranteed by Lemma 3.15. We only consider the case $t_{ij}^* \neq t_{ij'}^*$.

Our idea is to replace the fixed vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{x}_{ij'}$ with a pair of variables (which are linear combinations of $\boldsymbol{x}_{ij}$ and $\boldsymbol{x}_{ij'}$), and analyze the value of f. Formally, let $\theta \in \mathbb{R}$ be a parameter, and define the matrix $Q(\theta)$ as the $k_i \times k_i$ orthogonal matrix obtained by replacing the 2×2 submatrix of entries (j,j), (j,j'), (j',j), (j',j') in the

identify matrix $I_{k_i \times k_i}$ with
$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix},$$
where $\sin\theta$ is the entry (j, j') and $-\sin\theta$ is the entry (j', j). Let $[\widetilde{\boldsymbol{x}}_{i1}, \widetilde{\boldsymbol{x}}_{i2}, \ldots, \widetilde{\boldsymbol{x}}_{ik_i}] = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik_i}] \cdot Q(\theta)$. We have

$$\widetilde{\boldsymbol{x}}_{ij} = \cos\theta \cdot \boldsymbol{x}_{ij} - \sin\theta \cdot \boldsymbol{x}_{ij'}, \tag{3.7}$$
$$\widetilde{\boldsymbol{x}}_{ij'} = \sin\theta \cdot \boldsymbol{x}_{ij} + \cos\theta \cdot \boldsymbol{x}_{ij'}. \tag{3.8}$$

Define $\widetilde{R}(\theta) = R_i^*(\boldsymbol{t}) \cdot Q(\theta)$, and let $\widetilde{X}(\theta)$ be the matrix that X would be if $R_i^*(\boldsymbol{t})$ was replaced with $\widetilde{R}(\theta)$. We consider the function $\widetilde{f}\colon \mathbb{R} \to \mathbb{R}$ as defined below:

$$\begin{aligned} \widetilde{f}(\theta) &= f\big(\boldsymbol{t}^*, R_1^*(\boldsymbol{t}^*), \ldots, R_{i-1}^*(\boldsymbol{t}^*), \widetilde{R}(\theta), R_{i+1}^*(\boldsymbol{t}^*), \ldots, R_n^*(\boldsymbol{t}^*)\big) \\ &= \langle \boldsymbol{\gamma}, \boldsymbol{t}^* \rangle - \ln\det\big(\widetilde{X}(\theta)\big). \end{aligned}$$

By Lemma 3.15, $\widetilde{f}(\theta)$ has a maximum at $\theta = 0$ since $\widetilde{R}(0) = R_i^*(\boldsymbol{t})$. Next, we calculate the derivative of $\widetilde{f}$ at $\theta = 0$. We will use the following formula (see [Lax07, Chapter 9]) for an invertible matrix A and variable s:

$$\frac{d}{ds}\Big(\ln\det(A)\Big) = \operatorname{tr}\Big(A^{-1}\frac{d}{ds}A\Big). \tag{3.9}$$

Note that the only terms in $\widetilde{X}(\theta)$ (according to Equation (3.2)) that depend on θ are

$$e^{t_{ij}^*}\widetilde{\boldsymbol{x}}_{ij}\widetilde{\boldsymbol{x}}_{ij}^{\mathsf{T}} \quad \text{and} \quad e^{t_{ij'}^*}\widetilde{\boldsymbol{x}}_{ij'}\widetilde{\boldsymbol{x}}_{ij'}^{\mathsf{T}}.$$

Using Equation (3.9), we have

$$\frac{d\widetilde{f}}{d\theta}(0) = -\operatorname{tr}\bigg(X^{-1}\Big(e^{t_{ij}^*} \cdot \frac{d}{d\theta}\bigg|_{\theta=0} \widetilde{\boldsymbol{x}}_{ij}\widetilde{\boldsymbol{x}}_{ij}^{\mathsf{T}} + e^{t_{ij'}^*} \cdot \frac{d}{d\theta}\bigg|_{\theta=0} \widetilde{\boldsymbol{x}}_{ij'}\widetilde{\boldsymbol{x}}_{ij'}^{\mathsf{T}}\Big)\bigg). \tag{3.10}$$

By Equation (3.7),

$$\begin{aligned} \frac{d}{d\theta}\bigg|_{\theta=0} \widetilde{\boldsymbol{x}}_{ij}\widetilde{\boldsymbol{x}}_{ij}^{\mathsf{T}} &= \frac{d}{d\theta}\bigg|_{\theta=0} (\cos\theta \cdot \boldsymbol{x}_{ij} - \sin\theta \cdot \boldsymbol{x}_{ij'})(\cos\theta \cdot \boldsymbol{x}_{ij} - \sin\theta \cdot \boldsymbol{x}_{ij'})^{\mathsf{T}} \\ &= \frac{d}{d\theta}\bigg|_{\theta=0} \Big((\cos\theta)^2 \cdot \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{T}} + (\sin\theta)^2 \cdot \boldsymbol{x}_{ij'}\boldsymbol{x}_{ij'}^{\mathsf{T}} \\ &\qquad\qquad - \sin\theta\cos\theta \cdot \big(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij'}^{\mathsf{T}} + \boldsymbol{x}_{ij'}\boldsymbol{x}_{ij}^{\mathsf{T}}\big)\Big) \\ &= -\big(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij'}^{\mathsf{T}} + \boldsymbol{x}_{ij'}\boldsymbol{x}_{ij}^{\mathsf{T}}\big). \end{aligned}$$

Hence

$$
\begin{aligned}
-\operatorname{tr}\left(X^{-1} \cdot e^{t^*_{ij}} \cdot \frac{d}{d\theta}\bigg|_{\theta=0} \widetilde{\boldsymbol{x}}_{ij}\widetilde{\boldsymbol{x}}_{ij}{}^\mathsf{T}\right) &= e^{t^*_{ij}} \cdot \operatorname{tr}\Big(M^\mathsf{T} M \cdot \big(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij'}{}^\mathsf{T} + \boldsymbol{x}_{ij'}\boldsymbol{x}_{ij}{}^\mathsf{T}\big)\Big) \\
&= e^{t^*_{ij}} \cdot \operatorname{tr}\Big(M \cdot \big(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij'}{}^\mathsf{T} + \boldsymbol{x}_{ij'}\boldsymbol{x}_{ij}{}^\mathsf{T}\big) \cdot M^\mathsf{T}\Big) \\
&= 2e^{t^*_{ij}} \cdot \langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle.
\end{aligned}
$$

Similarly,

$$
-\operatorname{tr}\left(X^{-1} \cdot e^{t^*_{ij}} \cdot \frac{d}{d\theta}\bigg|_{\theta=0} \widetilde{\boldsymbol{x}}_{ij'}\widetilde{\boldsymbol{x}}_{ij'}{}^\mathsf{T}\right) = -2e^{t^*_{ij'}} \cdot \langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle.
$$

Plugging these into Equation (3.10), we have

$$
\frac{d\widetilde{f}}{d\theta}(0) = 2\big(e^{t^*_{ij}} - e^{t^*_{ij'}}\big) \cdot \langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle.
$$

Since $\theta = 0$ is a maximum for $\widetilde{f}(\theta)$, the above derivative must be 0. Thus we have proved $\langle M\boldsymbol{x}_{ij}, M\boldsymbol{x}_{ij'}\rangle = 0$ for the case $t^*_{ij} \neq t^*_{ij'}$. □

Finally we are able to prove Theorem 3.13:

Proof of Theorem 3.13. Our proof consists of two steps. With slight abuse of notation, we will use a matrix to denote the linear map represented by this matrix.

Step 1: Fix some $\varepsilon > 0$. We obtain $\boldsymbol{t}^*, R^*_1(\boldsymbol{t}^*), \ldots, R^*_n(\boldsymbol{t}^*)$ according to Lemma 3.15 and Lemma 3.16. In this step, we use X, M and $\boldsymbol{x}_{ij}$ ($i \in [n], j \in [k_i]$) to denote their values at $\big(\boldsymbol{t}^*, R^*_1(\boldsymbol{t}^*), \ldots, R^*_n(\boldsymbol{t}^*)\big)$. We will show that M satisfies the requirement of Theorem 3.13 "approximately".

For $i \in [n]$, $j \in [k_i]$, we define

$$
\boldsymbol{u}_{ij} = \frac{M\boldsymbol{x}_{ij}}{\|M\boldsymbol{x}_{ij}\|_2}
$$

and

$$
\varepsilon_{ij} = \frac{\partial f}{\partial t_{ij}}\big(\boldsymbol{t}^*, R^*_1(\boldsymbol{t}^*), \ldots, R^*_n(\boldsymbol{t}^*)\big).
$$

By Lemma 3.20, $\{\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \ldots, \boldsymbol{u}_{ik_i}\}$ is an orthonormal basis of $M(V_i)$. And by the choice of $\boldsymbol{t}^*$, we have $\varepsilon_{ij} \in [-\varepsilon, \varepsilon]$. Using Equation (3.9),

$$
\begin{aligned}
\varepsilon_{ij} &= p_i - \operatorname{tr}\big(X^{-1} \cdot e^{t^*_{ij}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}{}^\mathsf{T}\big) \\
&= p_i - e^{t^*_{ij}} \cdot \operatorname{tr}\big(M \cdot \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}{}^\mathsf{T} \cdot M^\mathsf{T}\big) \\
&= p_i - e^{t^*_{ij}} \cdot \|M\boldsymbol{x}_{ij}\|_2^2.
\end{aligned}
$$

Hence

$$\begin{aligned}\sum_{i\in[n],j\in[k_i]}(p_i-\varepsilon_{ij})\cdot \boldsymbol{u}_{ij}{\boldsymbol{u}_{ij}}^{\mathsf{T}} &= \sum_{i\in[n],j\in[k_i]} e^{t^*_{ij}}\cdot\|M\boldsymbol{x}_{ij}\|_2^2\cdot\left(\frac{M\boldsymbol{x}_{ij}}{\|M\boldsymbol{x}_{ij}\|_2}\right)\left(\frac{M\boldsymbol{x}_{ij}}{\|M\boldsymbol{x}_{ij}\|_2}\right)^{\mathsf{T}}\\ &= \sum_{i\in[n],j\in[k_i]} e^{t^*_{ij}}\cdot M\boldsymbol{x}_{ij}{\boldsymbol{x}_{ij}}^{\mathsf{T}}M^{\mathsf{T}}\\ &= MXM^{\mathsf{T}}\\ &= I_{\ell\times\ell},\end{aligned}$$

where we used the fact $X = M^{-1}(M^{\mathsf{T}})^{-1}$, which follows from $M^{\mathsf{T}}M = X^{-1}$. Since $\{\boldsymbol{u}_{i1},\boldsymbol{u}_{i2},\ldots,\boldsymbol{u}_{ik_i}\}$ is an orthonormal basis of $M(V_i)$,

$$\begin{aligned}\left\|\Big(\sum_{i=1}^{n} p_i \operatorname{Proj}_{M(V_i)}\Big)-I_{\ell\times\ell}\right\|_2 &= \left\|\Big(\sum_{i=1}^{n} p_i\sum_{j=1}^{k_i}\boldsymbol{u}_{ij}{\boldsymbol{u}_{ij}}^{\mathsf{T}}\Big)-I_{\ell\times\ell}\right\|_2\\ &= \left\|\sum_{i\in[n],j\in[k_i]}\varepsilon_{ij}\cdot\boldsymbol{u}_{ij}{\boldsymbol{u}_{ij}}^{\mathsf{T}}\right\|_2\\ &\le \varepsilon\sum_{i\in[n],j\in[k_i]}\|\boldsymbol{u}_{ij}{\boldsymbol{u}_{ij}}^{\mathsf{T}}\|_2\\ &\le \varepsilon m,\end{aligned}$$

where $\|\cdot\|_2$ denotes the spectral norm.

Step 2: Let $\varepsilon\to 0$. Note that if we replace M with $M/\|M\|_2$ (which has spectral norm 1), the subspaces $M(V_i)$ are not changed. Our previous arguments in this section have shown the following: For any $\varepsilon>0$, there is an invertible matrix M with $\|M\|_2=1$ that satisfies

$$\left\|\Big(\sum_{i=1}^{n} p_i \operatorname{Proj}_{M(V_i)}\Big)-I_{\ell\times\ell}\right\|_2 \le \varepsilon m, \tag{3.11}$$

and for every $i\in[n]$, $M(V_i)$ has an orthonormal basis $\{\boldsymbol{u}_{i1},\boldsymbol{u}_{i2},\ldots,\boldsymbol{u}_{ik_i}\}$ with $\boldsymbol{u}_{ij}=M\boldsymbol{x}_{ij}/\|M\boldsymbol{x}_{ij}\|_2$ $(j\in[k_i])$, where $[\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\ldots,\boldsymbol{x}_{k_i}]=[\boldsymbol{v}_{i1},\boldsymbol{v}_{i2},\ldots,\boldsymbol{v}_{k_i}]\cdot R_i$ for some $R_i\in\mathbf{O}(k_i)$.

We define a constant

$$C=\min_{i\in[n],j\in[k_i]}\left\{\frac{1}{\sqrt{k_i}\cdot\|\boldsymbol{v}_{ij}\|_2}\right\}.$$

One can verify that $1/\|M\boldsymbol{x}_{ij}\|_2\ge C$ for all $i\in[n]$, $j\in[k_i]$, $R_i\in\mathbf{O}(k_i)$ and $\|M\|_2=1$. We see that the matrix $[\boldsymbol{u}_{11},\ldots,\ldots,\boldsymbol{u}_{nk_n}]$ is contained in the following two sets:

$$\mathcal{A} = \Big\{ [\boldsymbol{u}'_{11}, \ldots, \ldots, \boldsymbol{u}'_{nk_n}] \in \mathbb{R}^{\ell \times m} \\ : \|\boldsymbol{u}'_{ij}\|_2 = 1, \langle \boldsymbol{u}'_{ij}, \boldsymbol{u}'_{ij'} \rangle = 0, \forall i \in [n], j \in [k_i], j' \in [k_i] \setminus \{j\} \Big\}$$

and

$$\mathcal{B} = \Big\{ M' \cdot [\boldsymbol{v}_{11}, \ldots, \ldots, \boldsymbol{v}_{nk_n}] \cdot \operatorname{diag}\big(R_1, \ldots, R_n\big) \cdot \operatorname{diag}(C_{11}, \ldots, \ldots, C_{nk_n}) \\ : M' \in \mathbb{R}^{\ell \times \ell}, \|M'\|_2 = 1, R_i \in \mathbf{O}(k_i), C_{ij} \in [C, +\infty), \forall i \in [n], j \in [k_i] \Big\}.$$

Note that we do not require M' in the definition of $\mathcal{B}$ to be invertible. Since $\mathcal{A}$ is bounded and both $\mathcal{A}$, $\mathcal{B}$ are closed, the intersection $\mathcal{A} \cap \mathcal{B}$ must be compact. Hence there exists

$$[\boldsymbol{u}^*_{11}, \ldots, \ldots, \boldsymbol{u}^*_{nk_n}] \in \mathcal{A} \cap \mathcal{B}$$

such that for any $\varepsilon > 0$, there is $[\boldsymbol{u}_{11}, \ldots, \ldots, \boldsymbol{u}_{nk_n}]$ satisfying all above requirements and

$$\Big\| [\boldsymbol{u}_{11}, \ldots, \ldots, \boldsymbol{u}_{nk_n}] - [\boldsymbol{u}^*_{11}, \ldots, \ldots, \boldsymbol{u}^*_{nk_n}] \Big\|_2 < \varepsilon.$$

By the definition of set $\mathcal{B}$, there are $M' \in \mathbb{R}^{\ell \times \ell}$ with $\|M'\|_2 = 1$ and $R_i \in \mathbf{O}(k_i)$, $C_{ij} \in [C, +\infty)$ $(i \in [n], j \in [k_i])$ such that

$$[\boldsymbol{u}^*_{11}, \ldots, \ldots, \boldsymbol{u}^*_{nk_n}] = M' \cdot [\boldsymbol{v}_{11}, \ldots, \ldots, \boldsymbol{v}_{nk_n}] \\ \cdot \operatorname{diag}\big(R_1, \ldots, R_n\big) \cdot \operatorname{diag}(C_{11}, \ldots, \ldots, C_{nk_n}).$$

Then by the definition of set $\mathcal{A}$, we see that $[\boldsymbol{u}^*_{i1}, \boldsymbol{u}^*_{i2}, \ldots, \boldsymbol{u}^*_{ik_i}]$ is an orthonormal basis of $M'(V_i)$ for every $i \in [n]$. Using Inequality (3.11), we have

$$\sum_{i=1}^{n} p_i \operatorname{Proj}_{M'(V_i)} = \sum_{i=1}^{n} p_i \sum_{j=1}^{k_i} \boldsymbol{u}^*_{ij} {\boldsymbol{u}^*_{ij}}^{\mathsf{T}} = \lim_{\varepsilon \to 0} \left(\sum_{i=1}^{n} p_i \sum_{j=1}^{k_i} \boldsymbol{u}_{ij} {\boldsymbol{u}_{ij}}^{\mathsf{T}} \right) = I_{\ell \times \ell}, \quad (3.12)$$

where we used the fact that $\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$ is a continuous function of vector $\boldsymbol{u}$. To show that M' satisfies the requirement of Theorem 3.13, it remains to show that M' is invertible. Assume the opposite, i.e., there exists a nonzero vector $\boldsymbol{w} \in \mathbb{R}^{\ell}$ with $\boldsymbol{w}^{\mathsf{T}} M' = \mathbf{0}^{\mathsf{T}}$. Then there is $\boldsymbol{w}^T \boldsymbol{u}^*_{ij} = 0$ for all $i \in [n]$, $j \in [k_i]$, which contradicts Equation (3.12). Thus Theorem 3.13 is proved. □

3.3.4 A convenient form of Theorem 3.13

We give Theorem 3.22 below which is implied by Theorem 3.13 and is the form that will be used in our proof of Theorem 3.8. Before stating the theorem, we need to define *admissible sets* and *admissible vectors*, which have weaker requirements than admissible basis sets and admissible basis vectors (Definition 3.12) as they are not required to span the entire arrangement.

Definition 3.21 (admissible set and vector)**.** Let $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, be a list of subspaces. An index set $H \subseteq [n]$ is called a $\mathcal{V}$*-admissible set* if

$$\dim\Big(\sum_{i \in H} V_i\Big) = \sum_{i \in H} \dim(V_i),$$

i.e., every subspace with index in H has intersection $\{\mathbf{0}\}$ with the span of the other subspaces with indices in H.

A $\mathcal{V}$*-admissible vector* is the indicator vector $\mathbf{1}_H \in \{0,1\}^n$ of some $\mathcal{V}$-admissible set H.

Theorem 3.22. *Given a list of subspaces $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, and a vector $\boldsymbol{p} \in \mathbb{R}^n$ in the convex hull of all $\mathcal{V}$-admissible vectors. Then there exists an invertible linear map $M \colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ such that for any unit vector $\boldsymbol{w} \in \mathbb{R}^\ell$,*

$$\sum_{i=1}^{n} p_i \| \operatorname{Proj}_{M(V_i)}(\boldsymbol{w}) \|_2^2 \le 1,$$

where $M(V_i)$ is the subspace obtained by applying M on V_i, and $\operatorname{Proj}_{M(V_i)}(\boldsymbol{w})$ *is the projection of $\boldsymbol{w}$ onto $M(V_i)$.*

Note that with a slight abuse of notation we use $\operatorname{Proj}_{M(V_i)}$ to denote both the projection matrix and the projection map.

Proof. We construct a list of subspaces $\mathcal{V}'$ and a vector $\boldsymbol{p}' = (p'_1, p'_2, \dots, p'_{|\mathcal{V}'|})$ that satisfy the conditions of Theorem 3.13.

Let $V = V_1 + V_2 + \cdots + V_n$, $d = \dim(V)$, $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_d\}$ be some orthonormal basis of V, $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_d\}$ be the standard basis of $\mathbb{R}^d$, and $P \colon V \to \mathbb{R}^d$ be the linear map such that $P(\boldsymbol{b}_i) = \boldsymbol{e}_i$ for every $i \in [d]$. We use $P(\mathcal{V})$ to denote the list of subspaces $(P(V_1), P(V_2), \dots, P(V_2))$. Appending the one-dimensional spaces $\operatorname{span}\{\boldsymbol{e}_i\}$ $(i \in [d])$ to $P(\mathcal{V})$, we define

$$\mathcal{V}' = \big(P(V_1), P(V_2), \dots, P(V_n), \operatorname{span}\{\boldsymbol{e}_1\}, \operatorname{span}\{\boldsymbol{e}_2\}, \dots, \operatorname{span}\{\boldsymbol{e}_d\}\big).$$

For a $\mathcal{V}$-admissible set $H \subseteq [n]$, we can see that it is also $P(\mathcal{V})$-admissible and $\mathcal{V}'$-admissible. Moreover, there exists some $G \subseteq \{n+1, n+2, \dots, n+d\}$ such that $H' = H \cup G$ is a $\mathcal{V}'$-admissible basis set. Assume

$$\boldsymbol{p} = \sum_{\mathcal{V}\text{-admissible } H} \mu_H \mathbf{1}_H,$$

where $\mu_H \in [0,1]$ and $\sum \mu_H = 1$. We define

$$\boldsymbol{p}' = \sum_{\mathcal{V}\text{-admissible } H} \mu_H \mathbf{1}_{H'},$$

where H' is the $\mathcal{V}'$-admissible basis set extended from H as defined above. We have $\boldsymbol{p}' \in \mathbb{R}^{n+d}$ and $p'_i = p_i$ for all $i \in [n]$.

Apply Theorem 3.13 on $\mathcal{V}'$ and $\boldsymbol{p}'$. There exists an invertible linear map $M' \colon \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\sum_{i=1}^{n} p_i \operatorname{Proj}_{M'(P(V_i))} + \sum_{i=1}^{d} p'_{n+i} \operatorname{Proj}_{M'(\operatorname{span}\{\boldsymbol{e}_i\})} = I_{d\times d}.$$

For every unit vector $\boldsymbol{w}' \in \mathbb{R}^d$, we have

$$\begin{aligned}
1 &= \boldsymbol{w}'^{\mathsf{T}} \cdot I_{d\times d} \cdot \boldsymbol{w}' \\
&= \boldsymbol{w}'^{\mathsf{T}} \cdot \left(\sum_{i=1}^{n} p_i \operatorname{Proj}_{M'(P(V_i))} + \sum_{i=1}^{d} p'_{n+i} \operatorname{Proj}_{M'(\operatorname{span}\{\boldsymbol{e}_i\})} \right) \cdot \boldsymbol{w}' \\
&= \sum_{i=1}^{n} p_i \left\| \operatorname{Proj}_{M'(P(V_i))}(\boldsymbol{w}') \right\|_2^2 + \sum_{i=1}^{d} p'_{n+i} \left\| \operatorname{Proj}_{M'(\operatorname{span}\{\boldsymbol{e}_i\})}(\boldsymbol{w}') \right\|_2^2 \\
&\ge \sum_{i=1}^{n} p_i \left\| \operatorname{Proj}_{M'(P(V_i))}(\boldsymbol{w}') \right\|_2^2.
\end{aligned}$$

For every unit vector $\boldsymbol{w} \in V$, since $P(\boldsymbol{w})$ is a unit vector in $\mathbb{R}^d$,

$$\sum_{i=1}^{n} p_i \left\| \operatorname{Proj}_{P^{-1}(M'(P(V_i)))}(\boldsymbol{w}) \right\|_2^2 = \sum_{i=1}^{n} p_i \left\| \operatorname{Proj}_{M'(P(V_i))}(P(\boldsymbol{w})) \right\|_2^2 \le 1,$$

where $P^{-1} \colon \mathbb{R}^d \to V$ is the inverse of P with $P^{-1}(\boldsymbol{e}_i) = \boldsymbol{b}_i$ for every $i \in [d]$. Noting that $P^{-1}(M'(P(V_i))) \subseteq V$ for every $i \in [n]$, the above inequality also holds for unit vectors $\boldsymbol{w} \notin V$, as their projections onto $P^{-1}(M'(P(V_i)))$ are even shorter. We extend the $V \to V$ invertible linear map $P^{-1} \circ M' \circ P$ to an $\mathbb{R}^\ell \to \mathbb{R}^\ell$ invertible linear map M, and the theorem is proved. □

3.4 Proof of the main theorem

We derive Theorem 3.8 from the following Theorem 3.23 with a recursive argument, and then prove Theorem 3.23.

Theorem 3.23. *Suppose that $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, is a list of subspaces that has an (α, δ)-system. Let $k \ge \dim(V_i)$ for every $i \in [n]$ and $d = \dim(V_1 + V_2 + \cdots + V_n)$. Then for any $\beta \in (0, 1)$, at least one of the following two cases holds:*

1. *$d \le 120\alpha k^3/(\beta\delta)$;*

2. *There are $q \ge \delta n/(10\alpha)$ subspaces $V_{i_1}, V_{i_2}, \ldots, V_{i_q}$ and nonzero vectors $\boldsymbol{z}_1 \in V_{i_1}, \boldsymbol{z}_2 \in V_{i_2}, \ldots, \boldsymbol{z}_q \in V_{i_q}$ such that $\operatorname{rank}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q\} \le \beta d$.*

Proof of Theorem 3.8 *using Theorem* 3.23. Initially, let $n_0 = n$, $\delta_0 = \delta$, $d_0 = d$, $V_i^{(0)} = V_i$ for $i \in [n_0]$ and $\mathcal{V}^{(0)} = \big(V_1^{(0)}, V_2^{(0)}, \ldots, V_{n_0}^{(0)}\big) = \mathcal{V}$. We repeatedly apply Theorem 3.23 with parameters n_t, δ_t, d_t and $\mathcal{V}^{(t)} = \big(V_1^{(t)}, V_2^{(t)}, \ldots, V_{n_t}^{(t)}\big)$ for $t \in \mathbb{N}$ (α, k and β are fixed).

Suppose that we are at the tth step, $\mathcal{V}^{(t)} = \big(V_1^{(t)}, V_2^{(t)}, \ldots, V_{n_t}^{(t)}\big)$ is a list of subspaces that has an (α, δ_t)-system, $\dim\big(V_i^{(t)}\big) \leq k$ for all $i \in [n_t]$, and $d_t = \dim\big(V_1^{(t)} + V_2^{(t)} + \cdots + V_{n_t}^{(t)}\big)$. By Theorem 3.23, there are two cases:

1. $d_t \leq 120\alpha k^3/(\beta\delta_t)$. In this case, we do nothing and terminate.
2. There are $q \geq \delta_t n_t/(10\alpha)$ nonzero vectors $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q$ from different subspaces with $\operatorname{rank}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q\} \leq \beta d_t$.

   In this case, we find a linear map $P\colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ with kernel $\operatorname{span}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q\}$, and define $\mathcal{V}^{(t+1)} = \big(V_1^{(1)}, V_2^{(2)}, \ldots, V_{n_{t+1}}^{(t+1)}\big)$ as the list of nonzero (not $\{\mathbf{0}\}$) subspaces among $P\big(V_1^{(t)}\big), P\big(V_2^{(t)}\big), \ldots, P\big(V_{n_t}^{(t)}\big)$. We note that

$$\begin{aligned} d_{t+1} &= \dim\big(V_1^{(t+1)} + V_2^{(t+1)} + \cdots + V_{n_{t+1}}^{(t+1)}\big) \\ &= d_t - \operatorname{rank}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q\} \geq (1-\beta) d_t > 0. \end{aligned} \tag{3.13}$$

   Hence the list $\mathcal{V}^{(t+1)}$ is nonempty. By Lemma 3.7, $\mathcal{V}^{(t+1)}$ has an (α, δ_{t+1})-system for $\delta_{t+1} = \delta_t n_t / n_{t+1}$. Repeat the arguments for $t+1$.

Now we analyze the above procedure. By $\delta n = \delta_0 n_0 = \delta_1 n_1 = \delta_2 n_2 = \cdots$, the number of subspaces with a vector $\boldsymbol{z}_i$ ($i \in [q]$) mapped to $\{\mathbf{0}\}$ at each step is $q \geq \delta n/(10\alpha)$. Hence for every $t \in \mathbb{N}$ we have

$$\sum_{i=1}^{n_t} \dim\big(V_i^{(t)}\big) \leq \sum_{i=1}^{n} \dim(V_i) - t \cdot \frac{\delta n}{10\alpha} \leq kn - \frac{\delta t n}{10\alpha}.$$

Since the left side is at least $n_t > 0$, we have

$$t \leq \frac{kn}{\delta n/(10\alpha)} = \frac{10\alpha k}{\delta}.$$

Therefore the procedure terminates in $10\alpha k/\delta$ steps. Suppose $t^* \leq 10\alpha k/\delta$ is the last step, i.e., the case $d_{t^*} \leq 120\alpha k^3/(\beta\delta_{t^*})$ holds for t^*. By Inequality (3.13),

$$\frac{120\alpha k^3}{\beta\delta_{t^*}} \geq d_{t^*} \geq (1-\beta)^{t^*} d \quad \Longrightarrow \quad d \leq \frac{1}{(1-\beta)^{10\alpha k/\delta}} \cdot \frac{120\alpha k^3}{\beta\delta}, \tag{3.14}$$

where we used the fact $\delta_{t^*} \geq \delta$, which follows from $\delta_{t^*} n_{t^*} = \delta n$ and $n_{t^*} \leq n$.

In Inequality (3.14), we set

$$\beta = \min\left\{\frac{1}{2}, \frac{\delta}{\alpha k}\right\}. \tag{3.15}$$

One can see

$$(1-\beta)^{\alpha k/\delta} \geq \frac{1}{4}$$

by considering the following two cases:

1. If $\delta/(\alpha k) < 1/2$, we have $\beta = \delta/(\alpha k)$ and $(1-\beta)^{\alpha k/\delta} = \big(1-\delta/(\alpha k)\big)^{\alpha k/\delta} \geq 1/4$. The last step is seen by noting that $(1-x)^{1/x}$ is a decreasing function.
2. If $\delta/(\alpha k) \geq 1/2$, we have $\beta = 1/2$ and $(1-\beta)^{\alpha k/\delta} \geq (1/2)^{\alpha k/\delta} \geq 1/4$.

Therefore

$$d \leq 4^{10} \cdot \frac{120\alpha k^3}{\beta\delta} = O(\alpha^2 k^4/\delta^2),$$

where we used the fact $1/\beta = O(\alpha k/\delta)$, which follows from $\delta/\alpha \leq 3/2$ (Lemma 3.4) and Equation (3.15). □

3.4.1 Proof of Theorem 3.23 – a special case

In this subsection, we consider the case that the subspaces $V_1, V_2, \ldots, V_n$ are "well-separated". Formally, we give the following definition:

Definition 3.24 (τ-separated spaces)**.** We say two subspaces $V, V' \subseteq \mathbb{R}^\ell$ are τ-*separated* ($0 < \tau \leq 1$) if there is $|\langle \boldsymbol{u}, \boldsymbol{u}' \rangle| \leq 1-\tau$ for any two unit vectors $\boldsymbol{u} \in V$ and $\boldsymbol{u}' \in V'$.

We state two simple lemmas about τ-separated spaces:

Lemma 3.25. *Given two τ-separated subspaces $V, V' \subseteq \mathbb{R}^\ell$. Let $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k_1}\}$ and $\{\boldsymbol{u}'_1, \boldsymbol{u}'_2, \ldots, \boldsymbol{u}'_{k_2}\}$ be any orthonormal bases of V and V' respectively. For any unit vector $\boldsymbol{u} \in V + V'$, if we write $\boldsymbol{u}$ as a linear combination of the bases vectors:*

$$\boldsymbol{u} = \lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2 + \cdots + \lambda_{k_1} \boldsymbol{u}_{k_1} + \mu_1 \boldsymbol{u}'_1 + \mu_2 \boldsymbol{u}'_2 + \cdots + \mu_{k_2} \boldsymbol{u}'_{k_2},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{k_1}, \mu_1, \mu_2, \ldots, \mu_{k_2} \in \mathbb{R}$, then we have

$$\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_1}^2 + \mu_1^2 + \mu_2^2 + \cdots + \mu_{k_2}^2 \leq \frac{1}{\tau}.$$

Proof. Let $\boldsymbol{v} = \lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2 + \cdots + \lambda_{k_1} \boldsymbol{u}_{k_1} \in V$ and $\boldsymbol{w} = \mu_1 \boldsymbol{u}'_1 + \mu_2 \boldsymbol{u}'_2 + \cdots + \mu_{k_2} \boldsymbol{u}'_{k_2} \in V'$. Since V and V' are τ-separated,

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle \geq -(1-\tau) \cdot \|\boldsymbol{v}\|_2 \cdot \|\boldsymbol{w}\|_2 \geq -\frac{1-\tau}{2} \cdot (\|\boldsymbol{v}\|_2^2 + \|\boldsymbol{w}\|_2^2).$$

It follows that

$$1 = \|\boldsymbol{u}\|_2^2 = \|\boldsymbol{v}+\boldsymbol{w}\|_2^2 = \|\boldsymbol{v}\|_2^2 + \|\boldsymbol{w}\|_2^2 + 2\langle \boldsymbol{v}, \boldsymbol{w} \rangle \geq \tau \cdot (\|\boldsymbol{v}\|_2^2 + \|\boldsymbol{w}\|_2^2).$$

The lemma is proved by noting $\|\boldsymbol{v}\|_2^2 = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_1}^2$ and $\|\boldsymbol{w}\|_2^2 = \mu_1^2 + \mu_2^2 + \cdots + \mu_{k_2}^2$. □

Lemma 3.26. *Given two subspaces $V, V' \subseteq \mathbb{R}^\ell$ that are not τ-separated. For any orthonormal basis $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k_1}\}$ of V, there exists $j \in [k_1]$ with*

$$\| \operatorname{Proj}_{V'}(\boldsymbol{u}_j)\|_2^2 \geq \frac{(1-\tau)^2}{k_1},$$

where $\operatorname{Proj}_{V'}(\boldsymbol{u}_j)$ is the projection of $\boldsymbol{u}_j$ onto V'.

Proof. Let $\boldsymbol{u} \in V$, $\boldsymbol{u}' \in V'$ be unit vectors such that $|\langle \boldsymbol{u}, \boldsymbol{u}'\rangle| > 1-\tau$. Then $\| \operatorname{Proj}_{V'}(\boldsymbol{u})\|_2 \geq |\langle \boldsymbol{u}, \boldsymbol{u}'\rangle| > 1-\tau$. Suppose $\boldsymbol{u} = \lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2 + \cdots + \lambda_{k_1}\boldsymbol{u}_{k_1}$, where $\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_1}^2 = 1$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}(1-\tau)^2 < \| \operatorname{Proj}_{V'}(\boldsymbol{u})\|_2^2 &\leq \Big(\sum_{j=1}^{k_1} |\lambda_j| \cdot \| \operatorname{Proj}_{V'}(\boldsymbol{u}_j)\|_2\Big)^2 \\ &\leq \Big(\sum_{j=1}^{k_1} \lambda_j^2\Big)\Big(\sum_{j=1}^{k_1} \| \operatorname{Proj}_{V'}(\boldsymbol{u}_j)\|_2^2\Big) \\ &= \sum_{j=1}^{k_1} \| \operatorname{Proj}_{V'}(\boldsymbol{u}_j)\|_2^2.\end{aligned}$$

Therefore there exists $j \in [k_1]$ with $\| \operatorname{Proj}_{V'}(\boldsymbol{u}_j)\|_2^2 \geq (1-\tau)^2/k_1$. □

The following theorem handles the "well-separated" case of Theorem 3.23.

Theorem 3.27. *Suppose $\mathcal{V} = (V_1, V_2, \ldots, V_n)$, $V_i \in \mathbb{R}^\ell$, is a list of subspaces that has an (α, δ)-system $\mathcal{S} = (S_1, S_2, \ldots, S_w)$, $S_j \subseteq [n]$. Let $k \geq \dim(V_i)$ for every $i \in [n]$ and $d = \dim(V_1 + V_2 + \cdots + V_n)$. If for every $j \in [w]$ and $\{i_1, i_2\} \subseteq S_j$, the subspaces V_{i_1} and V_{i_2} are τ-separated, then $d \leq \alpha k/(\tau\delta)$.*

Proof. Let $k_1, k_2, \ldots, k_n$ be the dimensions of $V_1, V_2, \ldots, V_n$ respectively and $m = k_1 + k_2 + \cdots + k_n$. For every $i \in [n]$, we fix $B_i = \{\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \ldots, \boldsymbol{u}_{ik_i}\}$ to be some orthonormal basis of V_i. We use L to denote the $\ell \times m$ matrix consists of columns $\boldsymbol{u}_{11}, \ldots, \ldots, \boldsymbol{u}_{nk_n}$. For $s \in [m]$, we use $\psi(s) \in [n]$ to denote the integer satisfying

$$k_1 + k_2 + \cdots + k_{\psi(s)-1} < s \leq k_1 + k_2 + \cdots + k_{\psi(s)-1} + k_{\psi(s)}.$$

In other words, the sth column of L is a vector in $B_{\psi(s)}$. We will prove an upper bound for $d = \operatorname{rank}(L)$ by constructing a high rank $m \times m$ matrix Y such that $LY = 0$.

Claim 3.28. *For $s \in [m]$ and an index set S_j ($j \in [w]$) that contains $\psi(s)$, there exists a vector $\boldsymbol{c} \in \mathbb{R}^m$ such that $L\boldsymbol{c} = \boldsymbol{0}$, $c_s = 1$, $\sum_{t \neq s} c_t^2 \leq 1/\tau$, and for $t \in [m] \setminus \{s\}$, $c_t \neq 0$ only if $\psi(t) \in S_j \setminus \{\psi(s)\}$.*

Proof. Let $\boldsymbol{u}$ denote the sth column of L. There are two cases as follows:

Case 1: $|S_j| = 2$**.** Say $S_j = \{\psi(s), i\}$. We have $V_{\psi(s)} = V_i$. Hence there exist coefficients $\lambda_1, \lambda_2, \ldots, \lambda_{k_i} \in \mathbb{R}$ with $\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_i}^2 = 1$ such that

$$\boldsymbol{u} - \lambda_1 \boldsymbol{u}_{i1} - \lambda_2 \boldsymbol{u}_{i2} - \cdots - \lambda_{k_i} \boldsymbol{u}_{ik_i} = \boldsymbol{0}.$$

We can obtain from this equation a vector $\boldsymbol{c} \in \mathbb{R}^m$ such that $L\boldsymbol{c} = \boldsymbol{0}$, $c_s = 1$, $c_t \neq 0$ only if $t = s$ or $\psi(t) = i$, and $\sum_{t \neq s} c_t^2 = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_2}^2 = 1 \leq 1/\tau$.

Case 2: $|S_j| = 3$**.** Say $S_j = \{\psi(s), i, i'\}$. We have $V_{\psi(s)} \subseteq V_i + V_{i'}$. Hence there exist coefficients $\lambda_1, \lambda_2, \ldots, \lambda_{k_i}, \mu_1, \mu_2, \ldots, \mu_{k_{i'}} \in \mathbb{R}$ such that

$$\boldsymbol{u} - \lambda_1 \boldsymbol{u}_{i1} - \lambda_2 \boldsymbol{u}_{i2} - \cdots - \lambda_{k_i} \boldsymbol{u}_{ik_i} - \mu_1 \boldsymbol{u}_{i'1} - \mu_2 \boldsymbol{u}_{i'2} - \cdots - \mu_{k_{i'}} \boldsymbol{u}_{i'k_{i'}} = \boldsymbol{0}.$$

We can obtain from this equation a vector $\boldsymbol{c} \in \mathbb{R}^m$ such that $L\boldsymbol{c} = \boldsymbol{0}$, $c_s = 1$, $c_t \neq 0$ only if $t = s$ or $\psi(t) \in \{i, i'\}$, and by Lemma 3.25, $\sum_{t \neq s} c_t^2 = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{k_i}^2 + \mu_1^2 + \mu_2^2 + \cdots + \mu_{k_{i'}}^2 \leq 1/\tau$. ■

Claim 3.29. *For every $s \in [m]$, there exists a vector $\boldsymbol{y} \in \mathbb{R}^m$ satisfying $L\boldsymbol{y} = \boldsymbol{0}$, $y_s = \lceil \delta n \rceil$, and $\sum_{t \neq s} y_t^2 \leq \alpha \lceil \delta n \rceil / \tau$.*

Proof. There are at least δn index sets S_j that contain $\psi(s)$. Let $J \subseteq [w]$, $|J| = \lceil \delta n \rceil$, be such that $\psi(s) \in S_j$ and for every $j \in J$. Using Claim 3.28, we find a vector $\boldsymbol{c}_j = (c_{j1}, c_{j_2}, \ldots, c_{jm})^\mathsf{T} \in \mathbb{R}^m$ for each $j \in J$. Let

$$\boldsymbol{y} = \sum_{j \in J} \boldsymbol{c}_j.$$

Clearly, we have $L\boldsymbol{y} = \boldsymbol{0}$ and $y_s = \lceil \delta n \rceil$. It remains to consider $\sum_{t \neq s} y_t^2$. For $t \in [m] \setminus \{s\}$ with $\psi(t) = \psi(s)$, we have $c_{jt} = 0$ for every $j \in J$. For $t \in [m]$ with $\psi(t) \neq \psi(s)$, since there are at most α index sets that contain $\{\psi(s), \psi(t)\}$, the number of nonzero elements in $\{c_{jt}\}_{j \in J}$ is at most α. Hence

$$\sum_{t \neq s} y_t^2 = \sum_{t \neq s} \Big(\sum_{j \in J} c_{jt} \Big)^2 \leq \sum_{t \neq s} \Big(\alpha \cdot \sum_{j \in J} c_{jt}^2 \Big) = \alpha \sum_{j \in J} \Big(\sum_{t \neq s} c_{jt}^2 \Big) \leq \frac{\alpha \lceil \delta n \rceil}{\tau}. \qquad ■$$

Using Claim 3.29, we find vectors $\boldsymbol{y}_s \in \mathbb{R}^m$ for every $s \in [m]$. Define Y to be the matrix consists of columns $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m$. Clearly, we have $LY = 0$. By Lemma 2.17,

$$\operatorname{rank}(Y) \geq \frac{\operatorname{tr}(Y)^2}{\|Y\|_F^2} \geq \frac{(m \cdot \lceil \delta n \rceil)^2}{m \cdot \lceil \delta n \rceil^2 + m \cdot \alpha \lceil \delta n \rceil / \tau} = \frac{m}{1 + \alpha / (\tau \lceil \delta n \rceil)} \geq m - \frac{\alpha m}{\tau \lceil \delta n \rceil}.$$

Noting that $m \leq kn$, we have

$$d = \operatorname{rank}(L) \leq \frac{\alpha m}{\tau \lceil \delta n \rceil} \leq \frac{\alpha k}{\tau \delta}. \qquad □$$

3.4.2 Proof of Theorem 3.23 – general case

In this subsection we prove Theorem 3.23 for all cases.

Lemma 3.30. *Suppose $\mathcal{V} = (V_1, V_2, \dots, V_n)$, $V_i \subseteq \mathbb{R}^\ell$, is a list of subspaces that has an (α, δ)-system. Let $k \geq \dim(V_i)$ for every $i \in [n]$ and $d = \dim(V_1 + V_2 + \cdots + V_n)$.*

Assume $\beta \in (0, 1)$ is such that the second case of Theorem 3.23 does not hold. That is, for any $q \geq \delta n/(10\alpha)$ subspaces $V_{i_1}, V_{i_2}, \dots, V_{i_q}$ and nonzero vectors $\boldsymbol{z}_1 \in V_{i_1}, \boldsymbol{z}_2 \in V_{i_2}, \dots, \boldsymbol{z}_q \in V_{i_q}$, there is $\operatorname{rank}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_q\} > \beta d$.

Then there exists $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$ in the convex hull of all $\mathcal{V}$-admissible vectors and an index set $I \subseteq [n]$ of size $|I| \geq (1 - \delta/(10\alpha))n$ such that $p_i \geq \beta d/(kn)$ for all $i \in I$.

Note that since $\delta/(10\alpha) < 1$ by $\delta/\alpha \leq 3/2$ (Lemma 3.4), the requirement on $|I|$ in the lemma is non-trivial.

Proof. We show the following claim first:

Claim 3.31. *For any $E \subseteq [n]$ with $|E| \geq \delta n/(10\alpha)$, we can find a $\mathcal{V}$-admissible set $H \subseteq E$ with $|H| \geq \beta d/k$.*

Proof. We construct H in the following way: Initially let $H = \emptyset$. Then repeatedly find an $i_0 \in E \setminus H$ with $V_{i_0} \cap \left(\sum_{i \in H} V_i\right) = \{\mathbf{0}\}$ and add i_0 to H, until such an i_0 does not exist.

We claim that $|H| \geq \beta d/k$. Note that when the above procedure ends, for every $i_0 \in E \setminus H$, there exists a nonzero vector $\boldsymbol{z}_{i_0}$ in both V_{i_0} and $\sum_{i \in H} V_i$. We use the condition of Lemma 3.30 for the $|E| \geq \delta n/(10\alpha)$ subspaces that have indices in E. Pick an arbitrary vector from each V_{i_0} with $i_0 \in H$, and pick $\boldsymbol{z}_{i_0}$ from each V_{i_0} with $i_0 \in E \setminus H$. Then these vectors have rank at least βd. On the other hand, these vectors are contained in the space $\sum_{i \in H} V_i$, which has dimension at most $k|H|$. Therefore $k|H| \geq \beta d$ and $|H| \geq \beta d/k$. ■

Using Claim 3.31 repeatedly, we find $\mathcal{V}$-admissible sets $H_1 \in [n]$ with $|H_1| \geq \beta d/k$, $H_2 \in [n] \setminus H_1$ with $|H_2| \geq \beta d/k$, $H_3 \in [n] \setminus (H_1 \cup H_2)$ with $|H_3| \geq \beta d/k$, and so on. We do this until there are less than $\delta n/(10\alpha)$ indices left. Let t be the total number of $\mathcal{V}$-admissible sets in this list. We have

$$t \leq \frac{n}{\beta d/k} = \frac{nk}{\beta d}.$$

Define $I = H_1 \cup H_2 \cup \cdots \cup H_t$. There is

$$|I| \geq n - \frac{\delta n}{10\alpha} = \left(1 - \frac{\delta}{10\alpha}\right)n.$$

Define

$$\boldsymbol{p} = \frac{1}{t} \cdot \sum_{i=1}^{t} \mathbf{1}_{H_i},$$

where $\mathbf{1}_{H_i} \in \{0,1\}^n$ is the indicator vector of H_i. Clearly, $\boldsymbol{p}$ is in the convex hull of $\mathcal{V}$-admissible vectors, and for every $i \in I$,

$$p_i \geq \frac{1}{t} \geq \frac{\beta d}{nk}. \qquad \square$$

Finally, we are able to prove Theorem 3.23:

Proof. For the sake of contradiction, we assume that neither case of Theorem 3.23 holds.

By Lemma 3.30, there is a vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ in the convex hull of all $\mathcal{V}$-admissible vectors and an index set $I \subseteq [n]$ of size $|I| \geq (1 - \delta/(10\alpha))n$ such that $p_i \geq \beta d/(kn)$ for all $i \in I$. We apply Theorem 3.22 with $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$, and obtain an invertible linear map $M \colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ such that for any unit vector $\boldsymbol{w} \in \mathbb{R}^\ell$,

$$\sum_{i=1}^{n} p_i \| \operatorname{Proj}_{V'_i}(\boldsymbol{w}) \|_2^2 \leq 1,$$

where V'_i denotes $M(V_i)$. Since $p_i \geq \beta d/(kn)$ for every $i \in I$, we have

$$\sum_{i \in I} \| \operatorname{Proj}_{V'_i}(\boldsymbol{w}) \|_2^2 \leq \frac{kn}{\beta d}. \tag{3.16}$$

We will reduce the problem to the "well-separated" case discussed in the previous subsection. We say a pair $\{i_1, i_2\} \subseteq [n]$ is *bad* if V'_{i_1}, V'_{i_2} are not (1/2)-separated. Let $\mathcal{S} = (S_1, S_2, \ldots, S_w)$ be the (α, δ)-system of $\mathcal{V}$. By Lemma 3.6, $\mathcal{S}$ is also an (α, δ)-system of $\mathcal{V}' = (V'_1, V'_2, \ldots, V'_n)$. Next, we estimate the number of sets among $S_1, S_2, \ldots, S_w$ that contain a bad pair.

Claim 3.32. *For every $i_0 \in I$, there are at most $\delta n/(30\alpha)$ values of $i \in I$ such that the pair $\{i_0, i\}$ is bad.*

Proof. Let k_0 be the dimension of V'_{i_0} and $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k_0}\}$ be an orthonormal basis of V'_{i_0}. For any $i \in I$ such that V'_{i_0} and V'_i are not (1/2)-separated, by Lemma 3.26, there must be $j \in [k_0]$ with

$$\| \operatorname{Proj}_{V'_i}(\boldsymbol{u}_j) \|_2^2 \geq \frac{1}{4k_0} \geq \frac{1}{4k}. \tag{3.17}$$

Set $\boldsymbol{w} = \boldsymbol{u}_j$ in Inequality (3.16). We have

$$\sum_{i \in I} \| \operatorname{Proj}_{V'_i}(\boldsymbol{u}_j) \|_2^2 \leq \frac{kn}{\beta d}.$$

Hence there are at most

$$\frac{kn}{\beta d} \Big/ \frac{1}{4k} = \frac{4k^2 n}{\beta d}$$

values of $i \in I$ satisfying Inequality (3.17) for every fixed $j \in [k_0]$. It follows that the number of bad pairs $\{i_0, i\}$ with $i \in I$ is

$$k_0 \cdot \frac{4k^2 n}{\beta d} \leq \frac{4k^3 n}{\beta d} \leq \frac{\delta n}{30\alpha}.$$

In the last step, we used the assumption $d > 120\alpha k^3/(\beta\delta)$. ■

Using this claim, the total number of bad pairs in $[n] \times [n]$ is at most

$$\big|([n] \setminus I) \times [n]\big| + |I| \cdot \frac{\delta n}{30\alpha} \leq \frac{\delta n^2}{10\alpha} + \frac{\delta n^2}{30\alpha} = \frac{2\delta n^2}{15\alpha}.$$

We remove every $S_j \in \mathcal{S}$ that contains a bad pair, and use $\mathcal{S}'$ to denote the list of the remaining sets. Since each pair appears at most α times, we have removed at most $2\delta n^2/15$ sets. Noting that originally we have $|\mathcal{S}| \geq \delta n^2/3$ by Lemma 3.4,

$$|\mathcal{S}'| \geq \frac{\delta n^2}{3} - \frac{2\delta n^2}{15} \geq \frac{\delta n^2}{5}.$$

By Lemma 3.5, there is a sublist $\mathcal{V}'' = (V_1'', V_2'', \ldots, V_{n'}'') = (V_{i_1}', V_{i_2}', \ldots, V_{i_{n'}}')$ of $\mathcal{V}'$ and a sublist $\mathcal{S}''$ of $\mathcal{S}'$ such that $n' \geq \delta n/(10\alpha)$ and $\mathcal{S}''$ is an $(\alpha, \delta/10)$-system of $\mathcal{V}''$.

Now we Theorem 3.27 (the "well-separated" case) on $\mathcal{V}''$ and $\mathcal{S}''$,

$$\dim\big(V_{i_1}' + V_{i_2}' + \cdots + V_{i_{n'}}'\big) \leq \frac{\alpha k}{(1/2) \cdot \delta/10} = \frac{20\alpha k}{\delta} \leq \beta d.$$

In the last step, we used the assumption $d > 120\alpha k^3/(\beta\delta)$. Since the linear map M is invertible, there is $\dim(V_{i_1} + V_{i_2} + \cdots + V_{i_{n'}}) = \dim(V_{i_1}' + V_{i_2}' + \cdots + V_{i_{n'}}') \leq \beta d$. Recall $n' \geq \delta n/(10\alpha)$. We see that the second case of Theorem 3.23 holds, which contradicts our assumption. Thus Theorem 3.23 is proved. □

Chapter 4

Field Size Lower Bounds for Maximally Recoverable Codes

In this chapter, we prove a super-polynomial lower bound on the alphabet size of maximally recoverable codes. We first define the model formally and state our results in Section 4.1. Our proof will be in the next two sections. We reduce the lower bound problem to a graph labeling problem in Section 4.2. And finally we solve the graph problem in Section 4.3. The results in this chapter are also included in [GHK$^+$17].

4.1 Topologies and maximally recoverable codes

Many storage systems in practice (e.g., [HSX$^+$12, MLR$^+$14]) have the following rectangular layout: Data are stored as an $m \times n$ matrix such that every entry of the matrix is one symbol of the data. For every column, the last a ($0 \leq a < m$) symbols are parities (linear combinations) of the previous $m - a$ symbols in the column. Parities in different columns are calculated in the same way, i.e., all columns are codewords of the same code. Similarly, the last b ($0 \leq b < n$) symbols of every row are parities of that row. In addition to these row and column parities, there are also h ($0 \leq h < (m-a)(n-b)$) global parities that depend on all data symbols. See Figure 4.1 for an example.

To ease the proofs, we will define the model formally in Definition 4.1 using parity check equations instead of specifying the numbers and locations of the parities. We will show that this will not make a difference as we can treat an arbitrary set of the symbols as the parities, provided that the code has the desired property (being maximally recoverable) and h is within a reasonable range.

With abuse of notation, in this chapter we will use a code to denote the vector space of all its codewords, and vice versa. For convenience, we will also treat $m \times n$ matrices as vectors of length mn. For a vector $\boldsymbol{v}$ and a subset of its coordinates S, $\boldsymbol{v}$ *restricted to* S (denoted by $\boldsymbol{v}|_S$) is the subvector of $\boldsymbol{v}$ with only coordinates in S. For a linear code C and a subset of the coordinates S, C *restricted to* S (denoted by $C|_S$) is the code $\{\boldsymbol{v}|_S : \boldsymbol{v} \in C\}$.

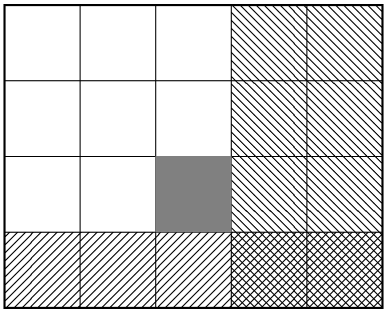

Figure 4.1: Illustration of a code with the described rectangular layout using a 4 × 5 grid, where the parameters are $m = 4$, $n = 5$, $a = 1$, $b = 2$, $h = 1$. The hatched cells correspond to the row and column parities. There are two cells at the bottom right corner that are both row and column parities. The gray cell correspond to the global parity.

Definition 4.1 (topology, row code, column code)**.** Let $m, n \in \mathbb{Z}^+$, $0 \le a < m$, $0 \le b < n$, and $0 \le h < (m-a)(n-b)$ be integers. We use the term *topology* $T_{m\times n}(a, b, h)$ to refer to the type of codes with $m \times n$ symbols, a parity check equations per column, b parity check equations per row, and h global parity check equations that depend on all symbols.

For a finite field $\mathbb{F}$ and a linear code $C \subseteq \mathbb{F}^{m\times n}$, We say that C *instantiates the topology* $T_{m\times n}(a, b, h)$ if there exist vectors $\boldsymbol{\alpha}^{(k)} = \big\{\alpha_i^{(k)}\big\}_{i\in[m]} \in \mathbb{F}^m$ for every $k \in [a]$, $\boldsymbol{\beta}^{(k)} = \big\{\beta_j^{(k)}\big\}_{j\in[n]} \in \mathbb{F}^n$ for every $k \in [b]$, and $\boldsymbol{\gamma}^{(k)} = \big\{\gamma_{i,j}^{(k)}\big\}_{i\in[m],j\in[n]} \in \mathbb{F}^{m\times n}$ for every $k \in [h]$ such that C is the linear space defined by the linear constraints below: For $\boldsymbol{v} \in \mathbb{F}^{m\times n}$, $\boldsymbol{v} \in C$ if and only if

1. For every column $\boldsymbol{c}$ of $\boldsymbol{v}$ and every $k \in [a]$, $\langle \boldsymbol{c}, \boldsymbol{\alpha}^{(k)} \rangle = 0$;

2. For every row $\boldsymbol{r}$ of $\boldsymbol{v}$ and every $k \in [b]$, $\langle \boldsymbol{r}, \boldsymbol{\beta}^{(k)} \rangle = 0$;

3. For every $k \in [h]$, $\langle \boldsymbol{v}, \boldsymbol{\gamma}^{(k)} \rangle = 0$.

For the above code C, we define its *column code* as $\big\{\boldsymbol{v} \in \mathbb{F}^m : \langle \boldsymbol{v}, \boldsymbol{\alpha}^{(k)} \rangle = 0, \forall k \in [a]\big\}$, and define its *row code* as $\big\{\boldsymbol{v} \in \mathbb{F}^n : \langle \boldsymbol{v}, \boldsymbol{\beta}^{(k)} \rangle = 0, \forall k \in [b]\big\}$.

In the above definition, every column is a codeword of the column code, and every row is a codeword of the row code. We note that the definition does not involve the parities mentioned in the rectangular layout model. In fact, it is possible that the last a entries of $\boldsymbol{\alpha}^{(k)}$ are all zeros, in which case one can no longer consider the last a symbols in a column as parities. The same applies to $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$.

Next, we define recoverable patterns and maximally recoverable codes. Intuitively, a *recoverable pattern* for a topology is a subset of all data symbols that can be recovered from erasure in some (not necessarily every) code that instantiates the topology. And *maximally recoverable codes* for a topology are codes that allow recoveries of all recoverable patterns.

Definition 4.2 (recoverable pattern)**.** Let C be a code that instantiates $T_{m\times n}(a,b,h)$ and $E \subseteq [m]\times[n]$. We say that E is a *recoverable pattern for the code C* if for vector $\boldsymbol{v} \in C$, the variables $\boldsymbol{v}|_E$ are uniquely determined by $\boldsymbol{v}|_{([m]\times[n])\setminus E}$ (i.e., $\boldsymbol{v}|_E$ can be represented as a linear transformation of $\boldsymbol{v}_{([m]\times[n])\setminus E}$ by solving the linear constraints in Definition 4.1).

We say that E is a *recoverable pattern for the topology $T_{m\times n}(a,b,h)$*, if there exists a code C instantiating $T_{m\times n}(a,b,h)$ such that E is a recoverable pattern for C.

Definition 4.3 (maximally recoverable code)**.** A code C that instantiates the topology $T_{m\times n}(a,b,h)$ is *maximally recoverable (MR)*, if for every $E \subseteq [m]\times[n]$ that is a recoverable pattern for the topology $T_{m\times n}(a,b,n)$, E is a recoverable pattern for the code C.

We note that MR codes exist for all values of m,n,a,b,h, if the field size is sufficiently large [GHJY14]. (In fact, [GHJY14] has shown MR codes exist for models more general than the rectangular topologies that we are considering.)

Recall that *maximum distance separable (MDS)* codes are linear codes that attain the maximum distance $\mathsf{length} - \mathsf{dimension} + 1$. It is well known that a linear code of dimension d is MDS if and only if every set of d coordinates is an *information set*, which is a set of coordinates that can take any values and any assignment of these coordinates uniquely determines the entire codeword [MS77].

Lemma 4.4. *For an MR code C that instantiates $T_{m\times n}(a,b,h)$, we have*

1. *For any $U \subseteq [m]$ of size $|U| = m-a$, $V \subseteq [n]$ of size $|V| = n-b$, and $H \subseteq U\times V$ of size $|H| = h$, $(U\times V)\setminus H$ is an information set of the code C.*
2. *Assume*
$$h \le (m-a)(n-b) - \max\{m-a, n-b\}.$$
Then the column code of C is an MDS code of length m and dimension $m-a$, and the row code is an MDS code of length n and dimension $n-b$.

Proof. We proceed item by item.

1. Let I denote $(U\times V)\setminus H$. We first show that $([m]\times[n])\setminus I$ is a recoverable pattern for C. Since C is maximally recoverable, it suffices to construct a code C' also instantiating $T_{m\times n}(a,b,h)$ such that $([m]\times[n])\setminus I$ is a recoverable pattern for C'. To define C', we pick $\mathbb{F}$, $\boldsymbol{\alpha}^{(k)}$, $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ in Definition 4.1 in such a way that the column code of C' is an MDS code of length m and dimension $m-a$, the row code of C' is an MDS code of length n and dimension $n-b$, and $C'|_{U\times V}$ is an MDS code of length $(m-a)(n-b)$ and dimension $(m-a)(n-b)-h = |I|$. It is easy to see that the values in I uniquely determine the rest of the codeword. Hence $([m]\times[n])\setminus I$ is a recoverable pattern for C', and it is also a recoverable pattern for C.

   It remains to show that the dimension of C is at least $|I| = (m-a)(n-b)-h$. We count the total number of parity check constraints for C. Let $S \subseteq [m]$

($|S| \geq m-a$) be an information set of the column code. We see that it is sufficient to count $m-|S|$ linear constraints for the column code. (If $a > m-|S|$, there must be redundancies among $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(a)}$ and we do not need a constraints.) We then consider the row code constraints. It is easy to see that the rows with indices in $[m] \setminus S$ are automatically codewords of the row code if the rows with indices in S are. So we count the row code linear constraints for only $|S|$ rows. In this way, the total number of constraints is

$$\begin{aligned}(m-|S|) \cdot n + b \cdot |S| + h &= mn - |S| \cdot (n-b) + h \\ &\leq mn - (m-a)(n-b) + h. \qquad (4.1)\end{aligned}$$

Hence the dimension of C is at least $(m-a)(n-b)-h$. Combining with the previous paragraph, we see that the dimension must be exactly $(m-a)(n-b)-h$ and I is an information set.

2. We only prove the claim for the column code. From the proof of the previous item, equality must hold in Inequality (4.1), and the dimension of the column code must be $|S| = m-a$. Assume that there is a $U \subseteq [m]$ of size $|U| = m-a$ that is not an information set of the column code. There must exist a linear dependency between the entries in U. Pick an arbitrary $V \subseteq [n]$ of size $|V| = n-b$ and $H \subseteq U \times V$ of size $|H| = h$ such that $(U \times V) \setminus H$ contains a complete column of $U \times V$. On one hand $(U \times V) \setminus H$ cannot be an information set of C as the entries of the column are linearly dependent. On the other hand $(U \times V) \setminus H$ is an information set of C by the previous item. We arrived at a contradiction. Therefore the column code must be MDS. □

By this lemma, we can see that under the mild assumption $h \leq (m-a)(n-b) - \max\{m-a, n-b\}$, the model in Definition 4.1 is equivalent to the storage model with the parities described at the beginning of this section. In fact, by Lemma 4.4 we can assume that the last a symbols of every column are the column parities, the last b symbols of every row are the row parities, and we can pick an arbitrary set of other h symbols as the global parities.

In this chapter, we consider lower bounds on the field sizes of MR codes. Intuitively, it might be easier to prove lower bounds for simple topologies with small a and b. And the following lemma provides a reduction to these simple topologies:

Lemma 4.5. *Suppose $C \subseteq \mathbb{F}^{m \times n}$ is an MR code that instantiates the topology $T_{m \times n}(a, b, h)$, and the condition*

$$h \leq (m-a)(n-b) - \max\{m-a, n-b\}$$

in Lemma 4.4 is satisfied. Then for any $0 \leq a' \leq a$ and $0 \leq b' \leq b$, there exists an MR code $C' \subseteq \mathbb{F}^{m' \times n'}$ that instantiates the topology $T_{m' \times n'}(a', b', h)$ over the same field, where $m' = m - a + a'$ and $n' = n - b + b'$.

Proof. By Lemma 4.4, we can consider an arbitrary set I of $(m-a)(n-b)-h$ coordinates in $[m-a] \times [n-b]$ as the original data symbols, and the other coordinates in $[m] \times [n]$ as the parities. Define C' as

$$C' = C|_{[m'] \times [n']},$$

i.e., the code obtained by restricting C to the coordinates $[m'] \times [n']$. We can see that C' instantiates the topology $T_{m' \times n'}(a', b', h)$, as $[m'] \times [n']$ contains I and the entries in $([m'] \times [n']) \setminus I$ are parities of the entries in I.

It remains to show that C' is MR. Suppose that $E' \subseteq [m'] \times [n']$ is an recoverable pattern for $T_{m' \times n'}(a', b', h)$, i.e., E' is recoverable for some code $\widetilde{C}'$ that instantiates $T_{m' \times n'}(a', b', h)$. We define $E \subseteq [m] \times [n]$ as the set of coordinates obtained by adding the last $m - m' = a - a'$ rows and $n - n' = b - b'$ columns of $[m] \times [n]$ to E', and define $\widetilde{C}$ as a code that instantiates $T_{m \times n}(a, b, h)$ obtained from $\widetilde{C}'$ by adding $a - a'$ parities to the column code and $b - b'$ parities to the row code, where the parities are arbitrary linear combinations of the existing entries in the column or row. Since E' is recoverable for $\widetilde{C}'$, we see that E is recoverable for $\widetilde{C}$, as one can first recover all entries in $E' \subseteq [m'] \times [n']$ and then calculate the newly added parities $([m] \times [n]) \setminus ([m'] \times [n'])$. Recall that the code C is MR. E must also be recoverable for C. It follows that E' is recoverable for C' since $([m] \times [n]) \setminus E$ and $([m'] \times [n']) \setminus E'$ are the same set of coordinates which uniquely determines a codeword for C and C'. Therefore C' is an MR code. □

4.1.1 Our results

We restrict out attention to characteristic-two fields, as these are the type of fields that are most widely considered in applications and research. Our main result in this chapter is a super-polynomial (of code length) lower bound on the field sizes of all MR codes instantiating $T_{m \times n}(a, b, h)$, where $a, b, h \geq 1$. Previously, there was not even a super-linear lower bound known for any topology (see [Bal12, GHJY14] for a simple linear lower bound for $h \geq 2$).

The following theorem is for the case $a = b = 1$:

Theorem 4.6. *Suppose $C \subseteq \mathbb{F}_q^{m \times n}$ is a maximally recoverable code that instantiates the topology $T_{m \times n}(1, 1, h)$, where $h \geq 1$ and $q = 2^t$ for some positive integer t. Then*

$$t = \Omega\Big(\big(\log \frac{\min\{m, n\}}{h}\big)^2\Big),$$

where $\Omega(\cdot)$ hides an absolute constant independent of h, m, n.

Following this theorem and using Lemma 4.5 with $a' = b' = 1$, one can easily derive a lower bound for general $a, b \geq 1$ as below. (We assume that the required condition on h in Lemma 4.5 is satisfied since otherwise the lower bound will become trivial.)

Corollary 4.7. *Suppose $C \subseteq \mathbb{F}_q^{m \times n}$ is a maximally recoverable code that instantiates the topology $T_{m \times n}(a, b, h)$, where $a, b, h \geq 1$ and $q = 2^t$ for some positive integer t.*

Then

$$t = \Omega\Big(\big(\log \frac{\min\{m-a+1, n-b+1\}}{h}\big)^2\Big),$$

where $\Omega(\cdot)$ *hides an absolute constant independent of* a, b, h, m, n*.*

This lower bound shows that MR codes with polynomial field sizes can only exist if one of a, b, h is 0. And it remains an open problem to prove non-trivial lower bounds for these cases. Among them, the topology $T_{m\times n}(1, 0, h)$ is well studied in the literature and widely used in practice (see also Appendix A for discussions about this topology), and it is of vital interest to prove super-linear lower bounds for general h.

Another open problem is to improve our existing lower bounds. For example, it not known if one can improve Theorem 4.6 to $t = \Omega\big(\log(m/h) \cdot \log(n/h)\big)$ for the case $m \neq n$. And for the simple topology $T_{n\times n}(1, 1, 1)$ (for which our result gives the "best" lower bound), there is a huge gap between our lower bound $t = \Omega\big((\log n)^2\big)$ and the best known constructions with $t = \Theta(n \log n)$.

A key step in proving Theorem 4.6 is the following lemma on graph edge labeling, which might be of independent interest:

Lemma 4.8. *Consider the complete bipartite graph* $K_{w,w}$ *and identify the edges with* $[w] \times [w]$*. Let* $\ell\colon [w] \times [w] \to \mathbb{F}_q$ *be a labeling of the edges such that for any simple cycle* $\mathcal{C} \subseteq [w] \times [w]$*,*

$$\sum_{e\in\mathcal{C}} \ell(e) \neq 0,$$

where $q = 2^t$ *for some positive integer* t*. Then* $t = \Omega\big((\log w)^2\big)$*.*

This can be viewed as an instance of *the critical problem* posed by Crapo and Rota in 1970 [CR70], where the goal is to find the largest dimension of a linear subspace of $\mathbb{F}_2^N$ that does not intersect a given set of vectors. Let $N = w^2$ and identify $[N]$ with the edges of $K_{w,w}$. Consider the labels as vectors in $\mathbb{F}_2^t$. We denote the vector consists of the ith bits of the labels of all edges by $\boldsymbol{u}_i \in \mathbb{F}_2^N$ ($i \in [t]$), and let V be the orthogonal complement of $\mathrm{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_t\}$. We can see that V consists of indicators of all $S \subseteq [N]$ such that $\sum_{e\in S} \ell(e) = 0$. And finding a labeling that satisfies the condition in Lemma 4.8 is equivalent to finding a subspace $V \subseteq \mathbb{F}_2^N$ that does not intersect the set of indicators of all simple cycles.

A similar edge labeling problem was also studied in a recent work [FGT16] in the context of derandomizing parallel algorithms for perfect matchings. The authors also considered an edge labeling problem where simple cycles carry nonzero sums. Some of the key differences from our setting are: Our techniques work for a special type of graphs (e.g., the complete bipartite graph) while [FGT16] considers general bipartite graphs; We need a single assignment while [FGT16] may have multiple assignments; We work over characteristic-two fields while [FGT16] work over characteristic-zero.

In the next section, we will reduce the MR code lower bound Theorem 4.6 to the graph labeling problem Lemma 4.8. After that, we will prove Lemma 4.8 in Section 4.3.

4.2 Reduction to the graph labeling problem

In this section, we will consider a code that instantiates the the topology $T_{m\times n}(1,1,h)$, where $h \geq 1$, and prove the lower bound as stated in Theorem 4.6 using the lower bound for graph labeling in Lemma 4.8. We assume that the condition in Lemma 4.4

$$h \leq (m-1)(n-1) - \max\{m-1, n-1\}$$

is satisfied, because otherwise the lower bound in Theorem 4.6 will become trivial. Recall that to specify a code that instantiates the topology $T_{m\times n}(1,1,h)$, one needs to specify the follows: the underlying field $\mathbb{F}_q$ (where q is a power of 2), the vector $\boldsymbol{\alpha}^{(1)} \in \mathbb{F}_q^m$, the vector $\boldsymbol{\beta}^{(1)} \in \mathbb{F}_q^n$, and the vectors $\boldsymbol{\gamma}^{(k)} \in \mathbb{F}_q^{m\times n}$ $(k \in h)$. If a code is MR, by Lemma 4.4, its column and row codes are MDS. Hence the entries in $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ are nonzero. If we replace $\gamma_{i,j}^{(k)}$ with $\gamma_{i,j}^{(k)}/\big(\alpha_i^{(1)}\beta_j^{(1)}\big)$ for all $i \in [m], j \in [n]$ and replace $\boldsymbol{\alpha}^{(1)}$, $\boldsymbol{\beta}^{(1)}$ with the all-ones vector, the code is still MR. Without loss of generality, we assume that $\boldsymbol{\alpha}^{(1)}$, $\boldsymbol{\beta}^{(1)}$ are the all-ones vector in this section. Under this assumption, a code that instantiates $T_{m\times n}(1,1,h)$ is uniquely determined by the field $\mathbb{F}_q$ and the vectors $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(h)}$.

We identify the code coordinates $[m] \times [n]$ with the edges of the complete bipartite graph $K_{m,n}$ in the natural way. That is, every vertex on the left side of $K_{m,n}$ corresponds to a row, every vertex on the right side of $K_{m,n}$ corresponds to a column, and every edge of $K_{m,n}$ corresponds to a code coordinate.

For a vertex v of $K_{m,n}$, we use $\Gamma(v) \subseteq [m] \times [n]$ to denote the set of edges incident to v ($\Gamma(v)$ contains either m or n edges). Let H denote the parity check matrix that defines the code. The columns of H correspond to the coordinates of the code, which have been identified with the edges of $K_{m,n}$. For the rows of H, we see the matrix H as two parts: (1) The *top part* consists of $m+n$ rows corresponding to the vertices of $K_{m,n}$. And for each vertex v, the corresponding row is the indicator vector $\mathbf{1}_{\Gamma(v)} \in \{0,1\}^{mn}$ of the edges incident to v; (2) The *bottom part*, which consists of h rows $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(h)}$, is for the global constraints of the code.

For a set of code coordinates $E \subseteq [m] \times [n]$, we consider E as a subset of the edges of $K_{m,n}$, and use $G_E = (L_E, R_E, E)$ to denote the subgraph of $K_{m,n}$ that only contains the edges in E and the vertices incident to E, where L_E and R_E denote the sets of vertices on the left side and right side respectively. We note that for $E' \subseteq E \subseteq [m] \times [n]$, with slight abuse of notation, in the remaining proofs we will use $\mathbf{1}_{E'}$ to denote the indicator of E' that has length either mn or $|E|$, depending on the concerned ground set is $[m] \times [n]$ or E.

Next, we characterize all recoverable patterns for $T_{m\times n}(1,1,h)$. For every $E \in [m] \times [n]$, one can see that E is recoverable if and only if the submatrix of H consisting of the columns corresponding to E has rank $|E|$. We will use $H|_E$ to denote this submatrix.

Lemma 4.9. *For every $E \subseteq [m] \times [n]$, E is a recoverable pattern for the topology $T_{m\times n}(1,1,h)$ if and only if*

$$|E| \leq |L_E| + |R_E| - c + h,$$

where c denotes the number of connected components in G_E. Moreover, if the equality $|E| = |L_E| + |R_E| - c + h$ holds and E is a recoverable pattern for the code defined by the field $\mathbb{F}_q$ and the vectors $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(h)}$, then the bottom part of $H|_E$ has rank h, i.e.,

$$\operatorname{rank}\bigl\{\boldsymbol{\gamma}^{(1)}|_E, \boldsymbol{\gamma}^{(2)}|_E, \ldots, \boldsymbol{\gamma}^{(h)}|_E\bigr\} = h.$$

Proof. We prove by calculating $\operatorname{rank}(H|_E)$. We first show the following claim:

Claim 4.10. *For any $E' \subseteq [m] \times [n]$ such that the graph $G_{E'}$ is connected, the rank of the top part of $H|_{E'}$ is exactly $|L_{E'}| + |R_{E'}| - 1$.*

Proof. In the top part of $H|_{E'}$, every column corresponds to an edge $e \in E'$, and every nonzero row is the indicator vector $\mathbf{1}_{\Gamma(v)\cap E'} \in \{0,1\}^{|E'|}$ of the edges incident to some vertex $v \in L_{E'} \cup R_{E'}$. For any $e \in E'$ and its two incident vertices $v_1 \in L_{E'}$, $v_2 \in R_{E'}$, we see that there are exactly two 1's in column corresponding to e, and they are at the rows corresponding to v_1 and v_2 respectively. It follows that the rows of the top part of $H|_{E'}$ sum to zero, i.e.,

$$\sum_{v\in L_{E'}} \mathbf{1}_{\Gamma(v)\cap E'} + \sum_{v\in R_{E'}} \mathbf{1}_{\Gamma(v)\cap E'} = \mathbf{0}. \tag{4.2}$$

Hence the rank of the top part of $H|_{E'}$ (which contains $|L_{E'}| + |R_{E'}|$ nonzero rows) is at most $|L_{E'}| + |R_{E'}| - 1$. To finish the proof, we need to show that Equation (4.2) is the only way (up to a multiplicative factor in $\mathbb{F}_q$) to linearly combine the nonzero rows in the top part of $H|_{E'}$ and get zero. In fact, in any linear combination of these rows that equals zero, if we take the row $\mathbf{1}_{\Gamma(v)\cap E'}$ corresponding to a vertex $v \in L_{E'} \cup R_{E'}$ with some nonzero coefficient, we must also take the row $\mathbf{1}_{\Gamma(v')\cap E'}$ for any neighbor v' of v with the same coefficient in order to cancel the two 1's in the column corresponding to the edge between v and v'. Since $G_{E'}$ is connected, we end up taking all nonzero rows with the same coefficient and getting the same linear combination as Equation (4.2). ■

Using this claim, we can see that the rank of the top part of $H|_E$ is exactly $|L_E|+|R_E|-c$. This is because in the top part of $H|_E$, the submatrices corresponding to the c connected components have disjoint columns and rows, and the rank of the top part is the sum of the ranks of these submatrices.

It is easy to see that there exists a field $\mathbb{F}_q$ and vectors $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(h)}$ such that the bottom part of $H|_E$ has rank h and the h rows of the bottom part are linearly independent of the rows of the top part. Therefore the maximum rank of $H|_E$ is $|L_E|+|R_E|-c+h$, and E is a recoverable pattern if and only if $|E| \le |L_E|+|R_E|-c+h$. And if $|E| = |L_E|+|R_E|-c+h$, E is recoverable if and only if $H|_E$ attains its maximum value, which happens only when the bottom part of $H|_E$ has rank h. □

Now we prove Theorem 4.6:

Proof of Theorem 4.6 *using Lemma* 4.8. Suppose that the code defined by the field $\mathbb{F}_q$ ($q = 2^t$) and vectors $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(h)}$ is maximally recoverable, and without loss of generality we assume $m \ge n$. We will prove $t = \Omega\bigl((\log(n/h))^2\bigr)$.

We partition $[n]$ into h disjoint sets $P_1, P_2, \ldots, P_h$ such that $|P_k| \geq \lfloor n/h \rfloor$ for every $k \in [h]$. We consider the subgraphs of $K_{m,n}$ on $P_1 \times P_1, P_2 \times P_2, \ldots, P_h \times P_h$, and label their edges using the vector $\boldsymbol{\gamma}^{(1)}$, i.e., the label of the edge $e \in [m] \times [n]$ is $\gamma_e^{(1)}$.

If one of these subgraphs satisfies the condition of Lemma 4.8, i.e., the label sum over every simple cycle is nonzero, then Theorem 4.6 is proved immediately. Next, we assume the opposite and derive a contradiction. Suppose $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h \subseteq [m] \times [n]$ are simple cycles from each of these subgraphs such that

$$\sum_{e \in \mathcal{C}_k} \gamma_e^{(1)} = 0 \quad \forall k \in [h]. \tag{4.3}$$

We define $E = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_h$. Then the graph G_E contains h connected components. Since simple cycles contain the same number of vertices and edges, we have

$$|E| = |L_E| + |R_E| = |L_E| + |R_E| - h + h.$$

By Lemma 4.9, E is a recoverable pattern, and in order to recover E there must be

$$\operatorname{rank}\left\{\boldsymbol{\gamma}^{(1)}|_E, \boldsymbol{\gamma}^{(2)}|_E, \ldots, \boldsymbol{\gamma}^{(h)}|_E\right\} = h.$$

Recall that the matrix $H|_E$ has rank $|E|$ for a recoverable pattern E. We can extend $\left\{\boldsymbol{\gamma}^{(1)}|_E, \boldsymbol{\gamma}^{(2)}|_E, \ldots, \boldsymbol{\gamma}^{(h)}|_E\right\}$ to a basis of $\mathbb{F}_q^{|E|}$ by adding rows in the top part of $H|_E$. Say the basis is

$$\left\{\boldsymbol{\gamma}^{(1)}|_E, \boldsymbol{\gamma}^{(2)}|_E, \ldots, \boldsymbol{\gamma}^{(h)}|_E\right\} \cup \left\{\mathbf{1}_{\Gamma(v)\cap E}\right\}_{v \in S},$$

where $S \subseteq L_E \cup R_E$ has size $|E| - h$ and $\mathbf{1}_{\Gamma(v)\cap E} \in \{0,1\}^{|E|}$ denotes the indicator vector of the edges in E that are incident to v. Note that for the indicator vector $\mathbf{1}_{\mathcal{C}_k} \in \{0,1\}^{|E|}$ of every cycle $\mathcal{C}_k$ ($k \in [h]$), there is

$$\langle \mathbf{1}_{\Gamma(v)\cap E}, \mathbf{1}_{\mathcal{C}_k} \rangle = \sum_{\Gamma(v)\cap \mathcal{C}_k} 1 = 0 \quad \forall v \in L_E \cup R_E,$$

and by Equation (4.3) there is also

$$\langle \boldsymbol{\gamma}^{(1)}|_E, \mathbf{1}_{\mathcal{C}_k} \rangle = 0.$$

Thus we have found h linearly independent vectors $\mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}, \ldots, \mathbf{1}_{\mathcal{C}_h}$ orthogonal to

$$|S| + 1 = |E| - h + 1$$

different basis vectors of $\mathbb{F}_q^{|E|}$. We arrive at a contradiction. □

4.3 Proof of the graph labeling lemma

In this section, we prove Lemma 4.8. We first define some notations for a general graph $G = (V, E)$. For distinct vertices $v_1, v_2 \in V$, we use $P(v_1, v_2)$ to denote the set of simple paths from v_1 to v_2. For $k \in \mathbb{Z}^+$ and distinct $v_1, v_2 \in V$, we use $P_k(v_1, v_2)$ to denote the set of simple paths from v_1 to v_2 that have length exactly k. For a path in $P_k(v_1, v_2)$, we say that v_1 is first vertex, v_2 is the $(k+1)$th vertex, and the other $(k-1)$ vertices on the path are the second through the kth vertices according to their positions. For an edge labeling $\ell \colon E \to \Sigma$ over some alphabet Σ and a path $\mathcal{P} \subseteq E$, we use $\ell(\mathcal{P})$ to denote the label sum over $\mathcal{P}$, i.e.,

$$\ell(\mathcal{P}) = \sum_{e \in \mathcal{P}} \ell(e).$$

The key step of our proof is the following lemma:

Lemma 4.11. *Let $G = (V, E)$ be a graph with maximum degree d, and $\ell \colon E \to \Sigma$ be a labeling of the edges, where the alphabet Σ is some abelian additive group. Suppose that for any pair of vertices $v_1 \neq v_2 \in V$ and any two vertex-disjoint simple paths $\mathcal{P}_1, \mathcal{P}_2 \in P(v_1, v_2)$, there is*

$$\ell(\mathcal{P}_1) \neq \ell(\mathcal{P}_2).$$

Then for arbitrary $v_1 \neq v_2 \in V$, positive integer $k \leq \sqrt{d}$ and $\ell_0 \in \Sigma$, the set

$$S = \Big\{ \mathcal{P} \in P_k(v_1, v_2) : \ell(\mathcal{P}) = \ell_0 \Big\}$$

has cardinality at most $k^{\log_2 k+1} d^{k-\log_2 k-1}$.

Before starting the formal proof, we first give some high-level ideas. Let's only consider the simple case that k is a small constant. Then the goal of the lemma would be showing $|S| \lesssim d^{k-\log_2 k-1}$. The total number of paths in $|S|$ would be d^{k-1} if all the intermediate $k-1$ vertices of a path could be chosen "freely". The lemma is basically saying that there are $\log_2 k$ vertices that are not "free". In the proof, we will show that there is a large subset $R \subseteq S$ such that all paths in R share the same ith vertex for some $i \in [2, k]$. In other words, many paths in S are fixed at the ith vertex, and the choice of this ith vertex is not "free". Then we fix the prefix before (or the suffix after) the ith vertex, and the possible choices of remaining half of the path can be considered as elements of a new set S' which also has fixed endpoints, length and label sum. We recursively apply the argument to that half of the path and S'. Intuitively, we can do this for $\log_2 k$ rounds (since each time we halve the length of the paths) and find $\log_2 k$ vertices that are not "free".

We now prove Lemma 4.11 formally by an induction on the length k.

Proof. Define

$$f(k) = k^{\log_2 k+1} d^{k-\log_2 k-1}$$

and we will need to show $|S| \leq f(k)$.

For $k = 1$, we have

$$|S| \leq 1 = k^{\log_2 k+1} d^{k-\log_2 k-1} = f(k).$$

Assume that we have proved the lemma for lengths up to $k-1$, and we now consider the case of k, where $2 \le k \le \sqrt{d}$.

If $S = \emptyset$, the lemma is trivial. We only consider the case that $S \ne \emptyset$. Pick an arbitrary path $\mathcal{P}_0 \in S$. Then for any other path $\mathcal{P} \in S$, $\mathcal{P}$ must intersect $\mathcal{P}_0$ at some vertex other than v_1, v_2, because of $\ell(\mathcal{P}) = \ell(\mathcal{P}_0) = \ell_0$ and the condition in Lemma 4.11. That is, there exists $i, j \in [k-1]$ such that the $(i+1)$th vertex of $\mathcal{P}$ is the same as the $(j+1)$th vertex of $\mathcal{P}_0$. Let R_{ij} denote the set of all these paths, formally

$$R_{ij} = \Big\{ \mathcal{P} \in S : \text{the } (i+1)\text{th vertex of } \mathcal{P} \text{ is the } (j+1)\text{th vertex of } \mathcal{P}_0 \Big\}.$$

We note that $\mathcal{P}_0 \in R_{ii}$ for every $i \in [k-1]$ and $\bigcup_{i,j\in[k-1]} R_{ij} = S$. By the Pigeonhole principle, there must exist $i_0, j_0 \in [k-1]$ such that

$$|R_{i_0 j_0}| \ge \frac{|S|}{(k-1)^2}. \tag{4.4}$$

We consider the paths in $R_{i_0 j_0}$. These paths share the same (i_0+1)th vertex. We denote this vertex by v_3. See Figure 4.2.

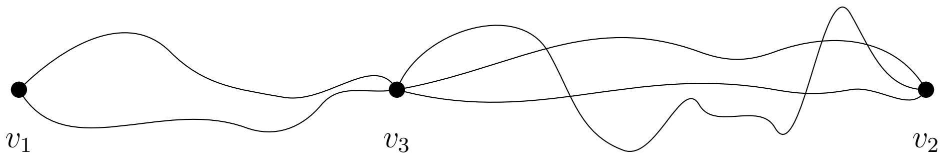

Figure 4.2: Paths in $R_{i_0 j_0}$ are fixed at three vertices v_1, v_3, v_2.

Every path $\mathcal{P} \in R_{i_0 j_0}$ can be considered as two parts: the *head* $\mathcal{P}_{\text{head}}$ from v_1 to v_3 (with length i_0) and the *tail* $\mathcal{P}_{\text{tail}}$ from v_3 to v_2 (with length $k - i_0$). We assume that the head is not shorter than the tail, i.e.,

$$i_0 \ge k/2.$$

If this condition does not hold, we can interchange the definitions of head and tail.

The number of possible tails is at most the number of simple paths from v_3 to v_2, which is bounded by d^{k-i_0-1}. We count the paths in $R_{i_0 j_0}$ according to their tails. For every choice of $\mathcal{P}_{\text{tail}}$, no matter what $\mathcal{P}_{\text{head}}$ is, the label sum of $\mathcal{P}_{\text{head}}$ is a fixed value:

$$\ell(\mathcal{P}_{\text{head}}) = \ell(\mathcal{P}) - \ell(\mathcal{P}_{\text{tail}}) = \ell_0 - \ell(\mathcal{P}_{\text{tail}}).$$

Hence by induction hypothesis, the number of possibilities of $\mathcal{P}_{\text{head}}$ for every fixed $\mathcal{P}_{\text{tail}}$ is bounded by $f(i_0) = {i_0}^{\log_2 i_0 + 1} d^{i_0 - \log_2 i_0 - 1}$. It follows that

$$\begin{aligned} |R_{i_0 j_0}| &\le d^{k-i_0-1} \cdot {i_0}^{\log_2 i_0 + 1} d^{i_0 - \log_2 i_0 - 1} \\ &= {i_0}^{\log_2 i_0 + 1} d^{k - \log_2 i_0 - 2}. \end{aligned}$$

Then by Inequality (4.4),

$$|S| \leq (k-1)^2 \cdot |R_{i_0 j_0}| \leq k^2 {i_0}^{\log_2 i_0 + 1} d^{k - \log_2 i_0 - 2}.$$

It remains to show that the right side is at most $f(k) = k^{\log_2 k + 1} d^{k - \log_2 k - 1}$. We finish the proof by considering the ratio:

$$\begin{aligned}
\frac{k^2 {i_0}^{\log_2 i_0 + 1} d^{k - \log_2 i_0 - 2}}{k^{\log_2 k + 1} d^{k - \log_2 k - 1}} &= \frac{{i_0}^{\log_2 i_0 + 1}}{k^{\log_2 k - 1} d^{\log_2 i_0 - \log_2 k + 1}} \\
&= \frac{{i_0}^{\log_2 i_0 + 1}}{k^{\log_2 k - 1} d^{\log_2 i_0 - \log_2 k + 1}} \cdot {i_0}^{-\log_2 k} \cdot k^{\log_2 i_0} \\
&= (k i_0 / d)^{\log_2 i_0 - \log_2 k + 1} \\
&= (k i_0 / d)^{\log_2 (2 i_0 / k)} \\
&\leq 1.
\end{aligned}$$

In the last step, we used $i_0 \leq k \leq \sqrt{d}$ and our assumption $i_0 \geq k/2$. □

We proceed to the proof of Lemma 4.8.

Proof of Lemma 4.8. We first claim that $K_{w,w}$ and labeling $\ell \colon [w] \times [w] \to \mathbb{F}_q$ ($q = 2^t$) satisfy the condition of Lemma 4.11. For two different vertices v_1, v_2 of $K_{w,w}$ and vertex-disjoint simple paths $\mathcal{P}_1, \mathcal{P}_2$ from v_1 to v_2, $\mathcal{P}_1$ and $\mathcal{P}_2$ form a simple cycle. Hence $\ell(\mathcal{P}_1) + \ell(\mathcal{P}_2) \neq 0$ by the condition in Lemma 4.8. Since the alphabet of the labeling has characteristic two, we have $\ell(\mathcal{P}_1) \neq \ell(\mathcal{P}_2)$.

Let $s = \lfloor (\sqrt{w} - 1)/2 \rfloor$ and $k = 2s + 1$. Clearly, $k \leq \sqrt{w}$. Pick an arbitrary pair of vertices v_1 and v_2 from the two sides of $K_{w,w}$. Then the number of simple paths from v_1 to v_2 with length k is

$$(w-1)(w-1)(w-2)(w-2)\cdots(w-s)(w-s) \geq (w-s)^{2s} = (w-s)^{k-1}.$$

Apply Lemma 4.11 with $d = w$. Then for every $\ell_0 \in \mathbb{F}_q$, the number of paths from v_1 to v_2 with length k and label sum ℓ_0 is at most

$$k^{\log_2 k + 1} w^{k - \log_2 k - 1}.$$

Note that there are $q = 2^t$ choices of ℓ_0. We have

$$2^t \geq \frac{(w-s)^{k-1}}{k^{\log_2 k + 1} w^{k - \log_2 k - 1}} = \frac{w^{\log_2 k}}{k^{\log_2 k + 1}} \cdot \left(\frac{w-s}{w} \right)^{k-1} = \frac{w^{\log_2 k}}{k^{\log_2 k + 1}} \cdot \Theta(1) = w^{\Omega(\log w)},$$

where we used the facts $s = \Theta(\sqrt{w})$ and $k = 2s + 1 = \Theta(\sqrt{w})$. It follows immediately that

$$t = \Omega\big((\log w)^2\big).$$ □

Appendix A

Constructions of Maximally Recoverable Codes

In this appendix, we present two new explicit constructions of MR codes for the topology $T_{m\times n}(1,0,h)$. We first explore some properties of this type of MR codes in Section A.1. Then we review known constructions and state our results in Section A.2. The details of our two constructions will be given in Section A.3 and Section A.4 respectively. The results in this appendix are also included in [GHK$^+$17] and [HY16].

A.1 The topology $T_{m\times n}(1,0,h)$

We have discussed rectangular topologies $T_{m\times n}(a,b,h)$ of storage systems in Section 4.1. Among them, the case $T_{m\times n}(1,0,h)$ is very widely used and has received a considerable amount of attention in the context of *locally recoverable codes* (see Section 1.2). According to Definition 4.1, to construct an MR code that instantiates $T_{m\times n}(1,0,h)$, we need to specify the field $\mathbb{F}$ and the vectors $\boldsymbol{\alpha}^{(1)} \in \mathbb{F}^m$, $\boldsymbol{\gamma}^{(k)} \in \mathbb{F}^{m\times n}$ ($k \in [h]$). In this appendix, we will give constructions with $\boldsymbol{\alpha}^{(1)}$ being the all-ones vector. In fact, as long as all entries of $\boldsymbol{\alpha}^{(1)}$ are nonzero (which must be true for all reasonable h by Lemma 4.4), one can always obtain an MR code with an arbitrary $\boldsymbol{\alpha}^{(1)}$ from any existing MR code, by multiplying $\left\{\alpha_i^{(1)}, \gamma_{i,j}^{(k)}\right\}_{j\in[n],k\in[h]}$ with a nonzero factor that depends on $i \in [m]$.

The parity check matrix of any concerned code is of the following form:

$$H = \left[\begin{array}{ccc|ccc|c|ccc}
1 & \cdots & 1 & & & & & & & \\
 & & & 1 & \cdots & 1 & & & & \\
 & & & & & & \ddots & & & \\
 & & & & & & & 1 & \cdots & 1 \\ \hline
\gamma^{(1)}_{1,1} & \cdots & \gamma^{(1)}_{m,1} & \gamma^{(1)}_{1,2} & \cdots & \gamma^{(1)}_{m,2} & \cdots\cdots\cdots & \gamma^{(1)}_{1,n} & \cdots & \gamma^{(1)}_{m,n} \\
\gamma^{(2)}_{1,1} & \cdots & \gamma^{(2)}_{m,1} & \gamma^{(2)}_{1,2} & \cdots & \gamma^{(2)}_{m,2} & \cdots\cdots\cdots & \gamma^{(2)}_{1,n} & \cdots & \gamma^{(2)}_{m,n} \\
 & \vdots & & & \vdots & & \vdots & & \vdots & \\
\gamma^{(h)}_{1,1} & \cdots & \gamma^{(h)}_{m,1} & \gamma^{(h)}_{1,2} & \cdots & \gamma^{(h)}_{m,2} & \cdots\cdots\cdots & \gamma^{(h)}_{1,n} & \cdots & \gamma^{(h)}_{m,n}
\end{array}\right]. \tag{A.1}$$

The matrix H contains $n+h$ rows and mn columns. The columns can be partitioned into n *column groups*, which correspond to the n columns of a codeword. A column group contains m columns, each corresponds to a symbol in a codeword. Precisely, The codeword symbol at location $(i,j) \in [m] \times [n]$ corresponds to the ith column of the jth column group. The *top part* of H contains n rows that are the column code linear constraints. Since we have assumed that $\boldsymbol{\alpha}^{(1)}$ is the all-ones vector, the top n rows of H are of the form shown in Equation (A.1). The *bottom part* contains h rows that are the global linear constraints. To construct an MR code, the major task is to construct the bottom part of H.

The recoverable patterns for $T_{m\times n}(1,0,h)$ are fully characterized in [BHH13, GHJY14], as shown in the following lemma:

Lemma A.1 ([GHJY14])**.** *$E \subseteq [m] \times [n]$ is a recoverable pattern for the topology $T_{m\times n}(1,0,h)$ if and only if E can be obtained by picking at most one entry in each column of $[m] \times [n]$ and up to h other entries at arbitrary locations.*

By the definitions of recoverable patterns (Definition 4.2) and MR codes (Definition 4.3), we immediately have the following standard lemma:

Lemma A.2 ([GHJY14])**.** *A code C that instantiates $T_{m\times n}(1,0,h)$ defined by a parity check matrix H as in Equation* (A.1) *is maximally recoverable, if and only if any set of $n+h$ columns of H that is obtained by picking one column from each column group and h additional columns has full rank.*

In Lemma A.2, let S be a submatrix of H that consists of $n+h$ columns obtained by picking one column from each column group and h additional columns. We consider the rows and columns of S. For a column group that has only one column included in S, without changing the rank of S we can eliminate this column from S, and also eliminate the row in the top part of S at which this column has a 1 (which is the only 1 in the entire row).

Let M be the remaining submatrix of S after the eliminations, $g \le \min\{h,n\}$ be the number of remaining column groups that have a column in M, and $r_1, r_2, \ldots, r_g \ge 2$ be the sizes of the parts of these column groups included in M. Since M is a square matrix and has $g+h$ rows (g rows in the top part and h rows in the bottom part), we can see that $r_1 + r_2 + \cdots + r_g$, which is the number of columns of M, is equal to $g+h$.

On the other hand, for any submatrix M of H with parameters $g \le \min\{h,n\}$ and $r_1, r_2, \ldots, r_g \ge 2$ such that $r_1 + r_2 + \cdots + r_g = g + h$, we can find a possible original S by adding one column from each of the other $n-g$ column groups in H and the row at which the column has a 1.

From the above procedure, we obtain our main tool to prove a construction is MR, which is formally stated below:

Lemma A.3 ([GHJY14])**.** *A code C that instantiates $T_{m\times n}(1,0,h)$ defined by a parity check matrix H as in Equation* (A.1) *is maximally recoverable, if and only if for any integers $g \le \min\{h,n\}$, $r_1, r_2, \ldots, r_g \ge 2$ such that $r_1 + r_2 + \cdots + r_g = g + h$, any* *distinct g indices $j_1, j_2, \ldots, j_g \in [n]$, and any distinct r_k indices $i_{k1}, i_{k2}, \ldots, i_{kr_k} \in [m]$ for each $k \in [g]$, the following matrix has full rank:*

$$M = \left[\begin{array}{ccc|ccc|c|ccc}
1 & \cdots & 1 & & & & & & & \\
 & & & 1 & \cdots & 1 & & & & \\
 & & & & & & \ddots & & & \\
 & & & & & & & 1 & \cdots & 1 \\
\gamma^{(1)}_{i_{11},j_1} & \cdots & \gamma^{(1)}_{i_{1r_1},j_1} & \gamma^{(1)}_{i_{21},j_2} & \cdots & \gamma^{(1)}_{i_{2r_2},j_2} & \cdots\cdots\cdots & \gamma^{(1)}_{i_{g1},j_g} & \cdots & \gamma^{(1)}_{i_{gr_g},j_g} \\
\gamma^{(2)}_{i_{11},j_1} & \cdots & \gamma^{(2)}_{i_{1r_1},j_1} & \gamma^{(2)}_{i_{21},j_2} & \cdots & \gamma^{(2)}_{i_{2r_2},j_2} & \cdots\cdots\cdots & \gamma^{(2)}_{i_{g1},j_g} & \cdots & \gamma^{(2)}_{i_{gr_g},j_g} \\
 & \vdots & & & \vdots & & \vdots & & \vdots & \\
\gamma^{(h)}_{i_{11},j_1} & \cdots & \gamma^{(h)}_{i_{1r_1},j_1} & \gamma^{(h)}_{i_{21},j_2} & \cdots & \gamma^{(h)}_{i_{2r_2},j_2} & \cdots\cdots\cdots & \gamma^{(h)}_{i_{g1},j_g} & \cdots & \gamma^{(h)}_{i_{gr_g},j_g}
\end{array}\right].$$

A.2 Known constructions and our contributions

We consider constructions of MR codes that instantiate the topology $T_{m\times n}(1,0,h)$. For $h = 0$, the problem is trivial since the code is fixed. For $h = 1$, explicit MR codes exist over a field of size $O(m)$ [BHH13], which is sub-linear of the code length mn. To see this, using Lemma A.3 it suffices to ensure that $\gamma^{(1)}_{1,j}, \gamma^{(1)}_{2,j}, \ldots, \gamma^{(1)}_{m,j}$ are distinct for each $j \in [n]$.

For $h \geq 2$ there is a linear lower bound $\Omega(mn)$ on the field size [Bal12, GHJY14], and this bound was known to be tight for the case $h = 2$ (see constructions in [Bla13, BPSY16]). We will give another construction in Section A.3 that also matches this lower bound, which is formally stated as the following theorem:

Theorem A.4. *Suppose that positive integers m, n satisfy $(m-1)n > 2$ (so that the topology $T_{m\times n}(1,0,2)$ is well defined). Then there is an explicit construction of MR codes instantiating $T_{m\times n}(1,0,2)$ over a characteristic-two field of size $O(mn)$.*

We note that when m and n are powers of 2, our construction has field size exactly mn, which is about half of the field size of the previous constructions.

For general h, the first explicit construction was given in [GHJY14] over a field of size roughly $O\big(2^m n^{(h-1)(1-2^{-m})}\big)$. For small n and growing m, a better estimate on the field size of this construction is $O\big((mn)^{\lfloor (n+h)/2 \rfloor}\big)$, which is formally stated below:

Theorem A.5 ([GHJY14])**.** *Suppose that m, n are powers of 2 and $(m-1)n > h$. There is an explicit construction of MR codes instantiating the topology $T_{m\times n}(1,0,h)$ over a field $\mathbb{F}_q$ of size $q = (mn)^{(n+h)/2}$ when $n+h$ is even, or $q = 2(mn)^{(n+h-1)/2}$ when $n+h$ is odd.*

Proof sketch. Since m, n are powers of 2, the field $\mathbb{F}_q$ has characteristic two. Set $\gamma^{(k)}_{i,j} = x^{2^{k-1}}_{i,j}$ in the parity check matrix H as shown in Equation (A.1) for all $i \in [m]$, $j \in [n]$, $k \in [h]$, where $\{x_{i,j}\}$ are $\mathbb{F}_q$ elements to be determined. By a standard linear algebra argument (e.g., [LN97, Lemma 3.51]), the matrix M in Lemma A.3 has full rank if the elements $\{x_{i,j}\}$ are $(n+h)$*-wise independent*, i.e., any subset of $\{x_{i,j}\}$ with $n+h$ or fewer elements do not sum to zero. We construct $\{x_{i,j}\}$ as follows: If $n+h$ is even, set $x_{i,j}$ to have the form $z \circ z^3 \cdots \circ z^{n+h-1}$, and if $n+h$ is odd, set $x_{i,j}$ to have the form $1 \circ z \circ z^3 \cdots \circ z^{n+h-2}$, where z runs through the subfield $\mathbb{F}_{mn} \subseteq \mathbb{F}_q$ and $\circ$ denotes concatenation of binary strings. One can verify that $\{x_{i,j}\}$ are $(n+h)$-wise independent. □

In Section A.4, we will give another construction for general h. Our construction can beat the above Theorem A.5 in a narrow case. The more interesting part of the construction is probably that it uses a completely new technique, which will be discussed later. The field size of our construction is stated as the following theorem:

Theorem A.6. *Let p be a prime. Suppose that $m, n \geq 2$ are powers of p and $2 \leq h < (m-1)n$. Then there is an explicit construction of MR codes instantiating the topology $T_{m\times n}(1,0,h)$ over a field of size $(mn)^{n+h-\lceil h/n\rceil-1}$.*

Under some additional conditions, the above result can be improved slightly:

Theorem A.7. *Let p be a prime. Suppose* (1) *m, n are powers of p and $3 \leq h \leq (m-1)n$;* (2) *$h \not\equiv 1 \pmod{n}$; and* (3) *$\lceil h/n \rceil \not\equiv p-1 \pmod{p}$. Then there is an explicit construction of MR codes instantiating the topology $T_{m\times n}(1,0,h)$ over a field of size $(mn)^{n+h-\lceil h/n\rceil-2}$.*

Setting $n=p=2$ in Theorem A.7, we have the following corollary:

Corollary A.8. *Suppose that m is a power of 2, $n=2$, $h \equiv 0 \pmod{4}$ and $3 \leq h < (m-1)n$. Then there is an explicit construction of MR codes instantiating the topology $T_{m\times n}(1,0,h)$ over a field of size $(mn)^{h/2}$.*

This corollary improves the result $(mn)^{h/2+1}$ in Theorem A.5 to a clean $(mn)^{h/2}$ for the case $n=2$ and $h \equiv 0 \pmod{4}$. Perhaps more importantly, unlike most previously known constructions that work for all h, our code family is "Vandermonde type" rather than "linearized". (An exception is [TPD13], which implicitly yields a family of MR codes over field size $(mn)^{O(mn)}$.) Our constructions for $\gamma_{i,j}^{(1)}, \gamma_{i,j}^{(2)}, \gamma_{i,j}^{(3)}, \gamma_{i,j}^{(4)}, \ldots$ in the parity check matrix are based on powers with consecutive exponents, i.e., $x_{i,j}, x_{i,j}^2, x_{i,j}^3, x_{i,j}^4 \ldots$, rather than linear functions $x_{i,j}, x_{i,j}^2, x_{i,j}^4, x_{i,j}^8 \ldots$ (over characteristic-two). Because of the technique used, our construction does not depend on characteristic-two fields and works for arbitrary finite fields. One can show that no "linearized" construction can beat $O\big((mn)^{h/2}\big)$ for the field size. This can be seen by considering the case $n=2$. Roughly speaking, if the field size was much smaller than $(mn)^{h/2} \approx m^{h/2}$, we would be able to find $\sim h/2$ columns from each column group such that these two column sets have the same sum of bottom parts. And this will leads to a singular matrix M in Lemma A.3. Therefore new techniques are of vital interest.

A.3 New construction for $T_{m\times n}(1,0,2)$

In this section, we prove Theorem A.4 by giving a construction of MR codes that instantiate the topology $T_{m\times n}(1,0,2)$ over a field of size $O(mn)$.

Construction A.1. Given integers m, n, we construct $\gamma_{i,j}^{(k)}$ for all $i \in [m]$, $j \in [n]$ and $k \in \{1,2\}$. Let $m' \in [m, 2m)$ and $n' \in [n, 2n)$ be powers of 2. Pick an additive subgroup $G = \{s_1, s_2, \dots, s_{m'}\} \subseteq \mathbb{F}_{m'n'}$ and elements $d_1, d_2, \dots, d_{n'} \in \mathbb{F}_{m'n'}$ such that $d_{j_1} - d_{j_2} \notin G$ for all $j_1 \neq j_2$, i.e., $d_1, d_2, \dots, d_{n'}$ belong to different cosets of $\mathbb{F}_{m'n'}$ modulo the subgroup G. For $i \in [m]$ and $j \in [n]$, we set

$$\begin{aligned}\gamma_{ij}^{(1)} &= s_i,\\ \gamma_{ij}^{(2)} &= s_i^2 + d_j \cdot s_i.\end{aligned}$$

Proof of Theorem A.4. We show that Construction A.1 satisfies the requirement. We prove the code is MR using Lemma A.3 with $h = 2$. For integers $g, r_1, r_2, \dots, r_g$ satisfying the conditions $g \le \min\{h, n\} \le 2$, $r_1, r_2, \dots, r_g \ge 2$ and $r_1 + r_2 + \dots + r_g = g + h$, there are only two possibilities:

Case 1: $g = 1$ and $r_1 = 3$. In this case, the matrix M in Lemma A.3 is

$$M = \begin{bmatrix} 1 & 1 & 1 \\ s_{i_{11}} & s_{i_{12}} & s_{i_{13}} \\ s_{i_{11}}^2 + d_{j_1}\cdot s_{i_{11}} & s_{i_{12}}^2 + d_{j_1}\cdot s_{i_{12}} & s_{i_{13}}^2 + d_{j_1}\cdot s_{i_{13}} \end{bmatrix},$$

where $j_1 \in [n]$ and distinct $i_{11}, i_{12}, i_{13} \in [m]$. If we multiple the second row by $-d_{j_1}$ and add it to the third row, M will become a Vandermonde matrix, which has full rank:

$$\det(M) = \det \begin{bmatrix} 1 & 1 & 1 \\ s_{i_{11}} & s_{i_{12}} & s_{i_{13}} \\ s_{i_{11}}^2 & s_{i_{12}}^2 & s_{i_{13}}^2 \end{bmatrix} \neq 0.$$

Case 2: $g = 2$ and $r_1 = r_2 = 2$. In this case, the matrix M in Lemma A.3 is

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ s_{i_{11}} & s_{i_{12}} & s_{i_{21}} & s_{i_{22}} \\ s_{i_{11}}^2 + d_{j_1}\cdot s_{i_{11}} & s_{i_{12}}^2 + d_{j_1}\cdot s_{i_{12}} & s_{i_{21}}^2 + d_{j_2}\cdot s_{i_{21}} & s_{i_{22}}^2 + d_{j_2}\cdot s_{i_{22}} \end{bmatrix},$$

where $j_1 \neq j_2$ are from $[n]$, $i_{11} \neq i_{12}$ and $i_{21} \neq i_{22}$ are from $[m]$. Note that we are working over a characteristic-two field. If we add the first column to the second column, and the third column to the fourth column, the first two entries of the second and fourth columns will become zeros. It is easy to see

$$\begin{aligned}\det(M) &= \det\begin{bmatrix} s_{i_{11}}+s_{i_{12}} & s_{i_{21}}+s_{i_{22}} \\ s_{i_{11}}^2+s_{i_{12}}^2+d_{j_1}(s_{i_{11}}+s_{i_{12}}) & s_{i_{21}}^2+s_{i_{22}}^2+d_{j_2}(s_{i_{21}}+s_{i_{22}})\end{bmatrix} \\ &= \det\begin{bmatrix} s_{i_{11}}+s_{i_{12}} & s_{i_{21}}+s_{i_{22}} \\ (s_{i_{11}}+s_{i_{12}})^2+d_{j_1}(s_{i_{11}}+s_{i_{12}}) & (s_{i_{21}}+s_{i_{22}})^2+d_{j_2}(s_{i_{21}}+s_{i_{22}})\end{bmatrix} \\ &= (s_{i_{11}}+s_{i_{12}})(s_{i_{21}}+s_{i_{22}})\det\begin{bmatrix} 1 & 1 \\ s_{i_{11}}+s_{i_{12}}+d_{j_1} & s_{i_{21}}+s_{i_{22}}+d_{j_2}\end{bmatrix} \\ &\neq 0.\end{aligned}$$

In the last step, we used the fact $s_{i_{11}}+s_{i_{12}}+d_{j_1} \neq s_{i_{21}}+s_{i_{22}}+d_{j_2}$, which follows from $s_{i_{11}}, s_{i_{12}}, s_{i_{21}}, s_{i_{22}} \in G$ and $d_{j_1}-d_{j_2} \notin G$. □

A.4 New construction for general $T_{m\times n}(1,0,h)$

We now present our Vandermonde-type construction of MR codes that instantiate the topology $T_{m\times n}(1,0,h)$ for general h.

Construction A.2. Let p be a prime, m and n be powers of p, $h \geq 2$ be an integer, and $t \in [2,h]$ be an arbitrary parameter. This construction will be over the finite field $\mathbb{F}_q$ of size $q=(mn)^{n+h-t}$.

Note that $\mathbb{F}_{mn}$ is a subfield of $\mathbb{F}_q$. There exists some $\varphi \in \mathbb{F}_q$ such that every element of $\mathbb{F}_q$ can be uniquely represented as

$$\lambda_0+\lambda_1\varphi+\cdots+\lambda_{n+h-t-1}\varphi^{n+h-t-1} \quad (\lambda_0,\lambda_1,\ldots,\lambda_{n+h-t-1}\in\mathbb{F}_{mn}).$$

Let $\Phi \in \mathbb{F}_q^{(n-1)\times(n+h-t)}$ denote the matrix $\left\{\varphi^{(j-1)(mn)^{i-1}}\right\}_{i\in[n-1],j\in[n+h-t]}$, i.e.,

$$\Phi = \begin{bmatrix} 1 & \varphi & \cdots & \varphi^{n+h-t-1} \\ 1 & \varphi^{mn} & \cdots & \varphi^{(n+h-t-1)(mn)} \\ & & \vdots & \\ 1 & \varphi^{(mn)^{n-2}} & \cdots & \varphi^{(n+h-t-1)(mn)^{n-2}}\end{bmatrix}.$$

We find $(n+h-t)-(n-1)=h-t+1$ linearly independent vectors $\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_{h-t+1}\in\mathbb{F}_q^{n+h-t}$ that are orthogonal to all rows of Φ, i.e.,

$$\Phi\cdot\boldsymbol{u}_k=\mathbf{0} \quad \forall k\in[h-t+1]. \tag{A.2}$$

Pick an additive subgroup $G=\{s_1,s_2,\ldots,s_m\}\subseteq\mathbb{F}_{mn}$ and elements $d_1,d_2,\ldots,d_n\in\mathbb{F}_{mn}$ such that $d_{j_1}-d_{j_2}\notin G$ for $j_1\neq j_2$, i.e., $d_1,d_2,\ldots,d_n$ belong to different cosets of $\mathbb{F}_{mn}$ modulo the subgroup G. We use $G_j=\{x_{1,j},x_{2,j},\ldots,x_{m,j}\}$ ($j\in[n]$) to denote the coset $G+d_j=\left\{s_i+d_j:\forall i\in[m]\right\}$, where the order of the elements are arbitrary.

For $i \in [m]$, $j \in [n]$ and $k \in [h]$, we set $\gamma_{i,j}^{(k)}$ in Equation (A.1) as follows:

$$\gamma_{i,j}^{(k)} = \begin{cases} x_{i,j}^k & k \leq t-1, \\ \langle \boldsymbol{u}_{k-t+1}, \boldsymbol{w}(x_{i,j}) \rangle & k \geq t, \end{cases} \tag{A.3}$$

where $\boldsymbol{w} \colon \mathbb{F}_q \to \mathbb{F}_q^{n+h-t}$ is defined as $\boldsymbol{w}(x) = (x^t, x^{t+1}, \ldots, x^{n+h-1})^\mathsf{T}$.

We will prove that the above Construction A.2 satisfies the requirements in Theorem A.6 and Theorem A.7. Our proofs will be based on Lemma A.3. Note that in the construction, the order of column groups and the order of columns in each column group are arbitrary. Without loss of generality, we assume the indices $j_1 = 1, j_2 = 2, \ldots, j_g = g$ and $i_{k1} = 1, i_{k2} = 2, \ldots, i_{kr_k} = r_k$ for every $k \in [g]$ in the matrix M defined in Lemma A.3. We add the first $g-1$ rows of M to the gth row, and obtain a matrix as following:

$$M' = \left[\begin{array}{ccc|ccc|c|ccc} 1 & \cdots & 1 & & & & & & & \\ & & & 1 & \cdots & 1 & & & & \\ & & & & & & \ddots & & & \\ 1 & \cdots & 1 & 1 & \cdots & 1 & \cdots\cdots\cdots & 1 & \cdots & 1 \\ \gamma_{1,1}^{(1)} & \cdots & \gamma_{r_1,1}^{(1)} & \gamma_{1,2}^{(1)} & \cdots & \gamma_{r_2,2}^{(1)} & \cdots\cdots\cdots & \gamma_{1,g}^{(1)} & \cdots & \gamma_{r_g,g}^{(1)} \\ \gamma_{1,1}^{(2)} & \cdots & \gamma_{r_1,1}^{(2)} & \gamma_{1,2}^{(2)} & \cdots & \gamma_{r_2,2}^{(2)} & \cdots\cdots\cdots & \gamma_{1,g}^{(2)} & \cdots & \gamma_{r_g,g}^{(2)} \\ & \vdots & & & \vdots & & \vdots & & \vdots & \\ \gamma_{1,1}^{(h)} & \cdots & \gamma_{r_1,1}^{(h)} & \gamma_{1,2}^{(h)} & \cdots & \gamma_{r_2,2}^{(h)} & \cdots\cdots\cdots & \gamma_{1,g}^{(h)} & \cdots & \gamma_{r_g,g}^{(h)} \end{array} \right]. \tag{A.4}$$

We need to set the parameter t in Construction A.2 and prove that the matrix M' has full rank. A key step is the following lemma:

Lemma A.9. *We use the variables defined in Construction* A.2. *Suppose the matrix M' shown in Equation* (A.4) *does not have full rank. Then there exists a polynomial $f(x)$ with $1 \leq \deg(f) \leq t-1$ and $\mu_1, \ldots, \mu_g \in \mathbb{F}_q$ such that $f(x_{i,j}) = \mu_j$ for all $i \in [r_j]$, $j \in [g]$.*

Proof. Suppose M' does not have full rank. We pick $\mu_1, \mu_2, \ldots, \mu_{g-1}, \nu_0, \nu_1, \ldots, \nu_h \in \mathbb{F}_q$ such that they are not all zeros and

$$(-\mu_1, -\mu_2, \ldots, -\mu_{g-1}, \nu_0, \nu_1, \ldots, \nu_h) \cdot M' = \mathbf{0}^\mathsf{T}.$$

We define $\mu_g = 0$. From the construction of $\gamma_{i,j}^{(k)}$ (Equation (A.3)), we can see that the following polynomial $f(x)$ has degree at most $n+h-1$ and satisfies $f(x_{i,j}) = \mu_j$ for all $i \in [r_j]$, $j \in [g]$:

$$f(x) = \sum_{k=0}^{t-1} \nu_k x^k + \sum_{k=t}^{h} \nu_k \langle \boldsymbol{u}_{k-t+1}, \boldsymbol{w}(x) \rangle.$$

It remains to show $1 \leq \deg(f) \leq t-1$. Let $c_t, c_{t+1}, \ldots, c_{n+h-1}$ be the coefficients of $x^t, x^{t+1}, \ldots, x^{n+h-1}$ in $f(x)$, i.e.,

$$(c_t, c_{t+1}, \ldots, c_{n+h-1})^\mathsf{T} = \nu_t \boldsymbol{u}_1 + \nu_{t+1}\boldsymbol{u}_2 + \cdots + \nu_h \boldsymbol{u}_{h-t+1},$$

and

$$f(x) = \sum_{k=0}^{t-1} \nu_k x^k + \sum_{k=t}^{n+h-1} c_k x^k. \tag{A.5}$$

By the definition of $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{h-t+1}$ (Equation (A.2)), we have

$$\Phi \cdot (c_t, c_{t+1}, \ldots, c_{n+h-1})^\mathsf{T} = \mathbf{0}. \tag{A.6}$$

Claim A.10. *$f(x)$ is not a constant, i.e.,* $\deg(f) \geq 1$.

Proof of Claim A.10. Assume $f(x)$ is a constant. By Equation (A.5), there must be $\nu_1 = \nu_2 = \cdots = \nu_{t-1} = 0$ and $c_t = c_{t+1} = \cdots = c_{n+h-1} = 0$. Since the vector $(c_t, c_{t+1}, \ldots, c_{n+h-1})^\mathsf{T}$ is a linear combination of linearly independent vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{h-t+1}$, the coefficients of the linear combination, which are $\nu_t, \nu_{t+1}, \ldots, \nu_h$, must also be zeros. Now we have $\nu_1 = \nu_2 = \cdots = \nu_h = 0$. It follows that the first g rows of M' are linearly dependent (with coefficients $-\mu_1, -\mu_2, \ldots, -\mu_{g-1}, \nu_0$), which is clearly false. ■

In order to show $\deg(f) \leq t-1$, we need the following claim first:

Claim A.11. *The rank of $\{\nu_0, \nu_1, \ldots, \nu_{t-1}, c_t, c_{t+1}, \ldots, c_{n+h-1}\}$ (i.e., the coefficients of $f(x)$) over the subfield $\mathbb{F}_{mn}$ is at most $n-1$.*

Proof of Claim A.11. Using Lagrange interpolating polynomials, we can find a polynomial $\widetilde{f}(x)$ that agrees with $f(x)$ on $g+h$ different values $x_{i,j}$ ($i \in [r_j]$, $j \in [g]$):

$$\widetilde{f}(x) = \sum_{j=1}^{g} \sum_{i=1}^{r_j} \mu_j \frac{\prod_{(i',j') \neq (i,j)} (x - x_{i',j'})}{\prod_{(i',j') \neq (i,j)} (x_{i,j} - x_{i',j'})}. \tag{A.7}$$

Recall that in Construction A.2, we defined $\{x_{i,j}\}_{i \in [m], j \in [n]}$ as elements of $\mathbb{F}_{mn}$. By expanding Equation (A.7), we can see that the coefficients of $\widetilde{f}(x)$ are $\mathbb{F}_{mn}$ linear combinations of $\mu_1, \mu_2, \ldots, \mu_{g-1}$. (Note that we defined $\mu_g = 0$.) Therefore the rank of the coefficients of $\widetilde{f}(x)$ over $\mathbb{F}_{mn}$ is at most $g-1$. (For $g = 1$, $\widetilde{f}(x) \equiv 0$.)

For the case $g < n$, since $f(x)$ agrees with $\widetilde{f}(x)$ on $g+h$ values $x_{i,j}$ ($i \in [r_j]$, $j \in [g]$), there must exist a polynomial $\psi(x)$ with degree at most

$$\deg(f) - (g+h) \leq (n+h-1) - (g+h) = n-g-1$$

such that

$$f(x) \equiv \widetilde{f}(x) + \psi(x) \cdot \prod_{i \in [r_j], j \in [g]} (x - x_{i,j}). \tag{A.8}$$

The rank of the coefficients of $\psi(x)$ over $\mathbb{F}_{mn}$ is at most $n-g$ since there are at most $n-g$ terms in $\psi(x)$.

For the case $g=n$, there must be $f(x) \equiv \widetilde{f}(x)$ since $\deg(f), \deg(\widetilde{f}) < g+h = n+h$. In this case, Equation (A.8) also holds if we set $\psi(x) \equiv 0$, and the rank of the coefficients of $\psi(x) \equiv 0$ is also at most $n-g=0$.

By expanding Equation (A.8), the coefficients of $f(x)$ are $\mathbb{F}_{mn}$ linear combinations of the coefficients of $\widetilde{f}(x)$ and the coefficients of $\psi(x)$. Therefore the rank of the coefficients of $f(x)$ over F_{mn} is at most $g-1+n-g=n-1$. ■

Claim A.12. *$c_t = c_{t+1} = \cdots = c_{n+h-1} = 0$, i.e.,* $\deg(f) \leq t-1$.

Proof of Claim A.12. Let r be the rank of $\{c_t, c_{t+1}, \ldots, c_{n+h-1}\}$ over the subfield $\mathbb{F}_{mn}$. It suffices to show $r=0$. For the sake of contradiction, assume $r>0$. Let $\{\beta_1, \beta_2, \ldots, \beta_r\} \in \mathbb{F}_q^r$ be a basis of $\{c_t, c_{t+1}, \ldots, c_{n+h-1}\}$, and $\Xi = \{\xi_{i,j}\}_{i \in [n+h-t], j \in [r]}$ be the matrix over $\mathbb{F}_{mn}$ with $\operatorname{rank}(\Xi) = r$ such that

$$(c_t, c_{t+1}, \ldots, c_{n+h-1})^{\mathsf{T}} = \Xi \cdot (\beta_1, \beta_2, \ldots, \beta_r)^{\mathsf{T}}.$$

By Equation (A.6), we have

$$\mathbf{\Phi} \cdot \Xi \cdot (\beta_1, \beta_2, \ldots, \beta_r)^{\mathsf{T}} = 0. \tag{A.9}$$

We will derive a contradiction by showing that the matrix $\mathbf{\Phi} \cdot \Xi$ has full column rank. For every $j \in [r]$, let

$$\tau_j = \xi_{1,j} + \xi_{2,j}\varphi + \cdots + \xi_{n+h-t,j}\varphi^{n+h-t-1}$$

be the jth entry of the first row of the matrix $\mathbf{\Phi} \cdot \Xi$. Since $\xi_{i,j} \in \mathbb{F}_{mn}$, we have $\xi_{i,j}^{(mn)^k} = \xi_{i,j}$ for all $k \in \mathbb{N}$. Therefore the jth entry of the $(k+1)$th row ($j \in [r], k \in [n-2]$) of the matrix $\mathbf{\Phi} \cdot \Xi$ is

$$\begin{aligned} & \xi_{1,j} + \xi_{2,j}\varphi^{(mn)^k} + \cdots + \xi_{n+h-t,j}\varphi^{(n+h-t-1)(mn)^k} \\ =& \xi_{1,j}^{(mn)^k} + (\xi_{2,j}\varphi)^{(mn)^k} + \cdots + (\xi_{n+h-t,j}\varphi^{n+h-t-1})^{(mn)^k} \\ =& \tau_j^{(mn)^k}, \end{aligned}$$

where we used the fact that m, n are powers of p (the field characteristic). Thus

$$\mathbf{\Phi} \cdot \Xi = \begin{bmatrix} \tau_1 & \tau_2 & \cdots & \tau_r \\ \tau_1^{mn} & \tau_2^{mn} & \cdots & \tau_r^{mn} \\ & \vdots & & \\ \tau_1^{(mn)^{n-2}} & \tau_2^{(mn)^{n-2}} & \cdots & \tau_r^{(mn)^{n-2}} \end{bmatrix}.$$

The $r \times r$ submatrix (note that $r \leq n-1$ by Claim A.11) at the top part of $\mathbf{\Phi} \cdot \Xi$ has full rank if and only if $\tau_1, \tau_2, \ldots, \tau_r \in \mathbb{F}_q$ are linearly independent over the subfield $\mathbb{F}_{mn}$ (see [LN97, Lemma 3.51]). Recall that we picked $\varphi \in \mathbb{F}_q$ in Construction A.2 so that every element of $\mathbb{F}_q$ can be uniquely represented as

$$\lambda_0 + \lambda_1\varphi + \cdots + \lambda_{n+h-t-1}\varphi^{n+h-t-1} \quad (\lambda_0, \lambda_1, \ldots, \lambda_{n+h-t-1} \in \mathbb{F}_{mn}).$$

Combining this with $\operatorname{rank}(\Xi) = r$, we can see that $\tau_1, \tau_2, \ldots, \tau_r$ are linearly independent. Hence the matrix $\Phi \cdot \Xi$ has rank r, contradicting Equation (A.9). ■

By Claim A.10 and Claim A.12, we have $1 \le \deg(f) \le t-1$. This concludes the proof of Lemma A.9. □

Using Lemma A.9, we are able to prove Theorem A.6 and Theorem A.7.

Proof of Theorem A.6. The average value of $r_1, r_2, \ldots, r_g$ is

$$\frac{g+h}{g} = \frac{h}{g} + 1 \ge \frac{h}{n} + 1.$$

We set $t = \lceil h/n \rceil + 1$ in Construction A.2 (one can verify $t \in [2, h]$) and pick some $j_0 \in [g]$ with $r_{j_0} \ge t$. Assume that M' does not have full rank. By Lemma A.9, there exists a polynomial $f(x)$ with $1 \le \deg(f) \le t-1$ satisfying $f(x) = \mu_{j_0}$ for some $\mu_{j_0} \in \mathbb{F}_q$ at $r_{j_0} \ge t > \deg(f)$ different values $x = x_{i,j_0}$ $(i \in [r_{j_0}])$. We arrive at a contradiction. Therefore M' has full rank and Construction A.2 gives an MR code construction over field size $(mn)^{n+h-t} = (mn)^{n+h-\lceil h/n \rceil - 1}$. □

Proof of Theorem A.7. We set $t = \lceil h/n \rceil + 2$ in Construction A.2. One can verify that $t \in [2, h]$ holds unless $h = 3$ and $n = 2$, in which case the condition $h \not\equiv 1 \pmod{n}$ is violated. For the sake of contradiction, we assume that M' does not have full rank. If there exists $j_0 \in [g]$ with $r_{j_0} \ge \lceil h/n \rceil + 2 = t$, we can derive a contradiction with Lemma A.9 along the lines of the proof of Theorem A.6. We only consider the case that $r_j \le \lceil h/n \rceil + 1$ for all $j \in [g]$.

The average value of $r_1, r_2, \ldots, r_g$ is at least $h/n + 1$ as shown in the proof of Theorem A.6. It follows that there exists some $j_0 \in [g]$ with $r_{j_0} = \lceil h/n \rceil + 1$. We claim that there must be a different $j_1 \in [g] \setminus \{j_0\}$ with $r_{j_1} = \lceil h/n \rceil + 1$. If there was not such a j_1 we would have

$$g + h = \sum_{j \in [g]} r_j \le \Big(\Big\lceil \frac{h}{n} \Big\rceil + 1\Big) + (g-1) \cdot \Big\lceil \frac{h}{n} \Big\rceil = g \cdot \Big\lceil \frac{h}{n} \Big\rceil + 1.$$

By $\lceil h/n \rceil \ge 1$ and $g \le n$,

$$n + h \le n \cdot \Big\lceil \frac{h}{n} \Big\rceil + 1.$$

It follows that

$$n - 1 \le n \cdot \Big(\Big\lceil \frac{h}{n} \Big\rceil - \frac{h}{n}\Big).$$

This inequality holds only if $\lceil h/n \rceil - h/n$ attains its maximum value $(n-1)/n$, which happens only when $h \equiv 1 \pmod{n}$. Hence under the condition $h \not\equiv 1 \pmod{n}$, there must exist distinct $j_0, j_1 \in [g]$ with $r_{j_0} = r_{j_1} = \lceil h/n \rceil + 1 = t - 1$.

By Lemma A.9, there exists a polynomial $f(x)$ with $1 \leq \deg(f) \leq t-1$ such that $f(x) = \mu_{j_0}$ for some $\mu_{j_0} \in \mathbb{F}_q$ at $r_{j_0} = t-1$ different values $x = x_{i,j_0}$ $(i \in [r_{j_0}])$, and $f(x) = \mu_{j_1}$ for some $\mu_{j_1} \in \mathbb{F}_q$ at $r_{j_1} = t-1$ different values $x = x_{i,j_1}$ $(i \in [r_{j_1}])$. We can see that $f(x)$ can be written in two ways

$$\begin{aligned} f(x) &\equiv \lambda_0(x - x_{1,j_0})(x - x_{2,j_0}) \cdots (x - x_{t-1,j_0}) + \mu_{j_0} \\ &\equiv \lambda_1(x - x_{1,j_1})(x - x_{2,j_1}) \cdots (x - x_{t-1,j_1}) + \mu_{j_1}, \end{aligned}$$

where $\lambda_0, \lambda_1 \in \mathbb{F}_q$ are nonzero. Consider the x^{t-1} term in the expansions of these two representations, and there must be $\lambda_0 = \lambda_1$. Then we consider the x^{t-2} term. The coefficients of x^{t-2} in the two representations should be equal. Hence

$$\sum_{i=1}^{t-1}(x_{i,j_0} - x_{i,j_1}) = \left(\sum_{i=1}^{t-1} x_{i,j_0}\right) - \left(\sum_{i=1}^{t-1} x_{i,j_1}\right) = 0.$$

For every $i \in [t-1]$, $x_{i,j_0} - x_{i,j_1}$ can be written as $s + d_{j_0} - d_{j_1}$ for some $s \in G$, i.e., $x_{i,j_0} - x_{i,j_1}$ is in the coset $G + d_{j_0} - d_{j_1}$. The left side of the above equality must be in the coset $G + (t-1) \cdot (d_{j_0} - d_{j_1})$, where $(t-1) \cdot (d_{j_0} - d_{j_1})$ denotes summing $d_{j_0} - d_{j_1}$ for $t-1$ times. Since 0 is contained only in the coset G, there must be $(t-1) \cdot (d_{j_0} - d_{j_1}) = 0$. It follows that $t - 1 \equiv 0 \pmod{p}$. However, this contradicts the condition $\lceil h/n \rceil = t - 2 \not\equiv p - 1 \pmod{p}$. This concludes the proof of Theorem A.7. □

Bibliography

[ADSW14] Albert Ai, Zeev Dvir, Shubhangi Saraf, and Avi Wigderson. Sylvester-Gallai type theorems for approximate collinearity. *Forum of Mathematics, Sigma*, 2(e3), 2014.

[Alo09] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability and Computing*, 18(1–2):3–15, 2009.

[Bal12] Simeon Ball. On sets of vectors of a finite vector space in which every subset of basis size is a basis. *Journal of the European Mathematical Society*, 14(3):733–748, 2012.

[Bar98] Franck Barthe. On a reverse form of the Brascamp-Lieb inequality. *Inventiones Mathematicae*, 134(2):335–361, 1998.

[BCCT08] Jonathan Bennett, Anthony Carbery, Michael Christ, and Terence Tao. The Brascamp-Lieb inequalities: Finiteness, structure and extremals. *Geometric and Functional Analysis*, 17(5):1343–1415, 2008.

[BDHS14] Jop Briët, Zeev Dvir, Guangda Hu, and Shubhangi Saraf. Lower bounds for approximate LDCs. In *Proceedings of the 41st Annual International Colloquium on Automata, Languages, and Programming (ICALP '14)*, pages 259–270, 2014.

[BDWY11] Boaz Barak, Zeev Dvir, Avi Wigderson, and Amir Yehudayoff. Rank bounds for design matrices with applications to combinatorial geometry and locally correctable codes. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC '11)*, pages 519–528, 2011.

[BDWY13] Boaz Barak, Zeev Dvir, Avi Wigderson, and Amir Yehudayoff. Fractional Sylvester-Gallai theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19213–19219, 2013.

[BET10] Avraham Ben-Aroya, Klim Efremenko, and Amnon Ta-Shma. Local list decoding with a constant number of queries. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '10)*, pages 715–722, 2010.

[BF90] Donald Beaver and Joan Feigenbaum. Hiding instances in multioracle queries. In *Proceedings of the 7th Annual Symposium on Theoretical Aspects of Computer Science (STACS '90)*, pages 37–48, 1990.

[BFLS91] László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing (STOC '91)*, pages 21–32, 1991.

[BHH13] Mario Blaum, James L. Hafner, and Steven Hetzler. Partial-MDS codes and their application to RAID type of architectures. *IEEE Transactions on Information Theory*, 59(7):4510–4519, 2013.

[BK95] Manuel Blum and Sampath Kannan. Designing programs that check their work. *Journal of the ACM*, 42(1):269–291, 1995.

[BK15] S. B. Balaji and P. Vijay Kumar. On partial maximally-recoverable and maximally-recoverable codes. In *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT '15)*, pages 1881–1885, 2015.

[Bla13] Mario Blaum. Construction of PMDS and SD codes extending RAID 5. arXiv 1305.0032, 2013. `https://arxiv.org/abs/1305.0032`.

[BM90] Peter Borwein and William O. J. Moser. A survey of Sylvester's problem and its generalizations. *Aequationes Mathematicae*, 40(1):111–135, 1990.

[Bol86] Béla Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability.* Cambridge University Press, 1986.

[BPSY16] Mario Blaum, James S. Plank, Moshe Schwartz, and Eitan Yaakobi. Construction of partial MDS and sector-disk codes with two global parity symbols. *IEEE Transactions on Information Theory*, 62(5):2673–2681, 2016.

[BW89] Joel G. Broida and Stanley Gill Williamson. *A Comprehensive Introduction to Linear Algebra.* Addison-Wesley, 1989.

[CFL$^+$13] Yeow Meng Chee, Tao Feng, San Ling, Huaxiong Wang, and Liang Feng Zhang. Query-efficient locally decodable codes of subexponential length. *Computational Complexity*, 22(1):159–189, 2013.

[CGKS98] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–981, 1998.

[CHL07] Minghua Chen, Cheng Huang, and Jin Li. On the maximally recoverable property for multi-protection group codes. In *Proceedings of the 2007 IEEE International Symposium on Information Theory (ISIT '07)*, pages 486–490, 2007.

[CM15] Viveck R. Cadambe and Arya Mazumdar. Bounds on the size of locally recoverable codes. *IEEE Transactions on Information Theory*, 61(11):5787–5794, 2015.

[CR70] Henry H. Crapo and Gian C. Rota. *On the Foundations of Combinatorial Theory: Combinatorial Geometries*. MIT Press, 1970.

[CRT06] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[CSYS15] Junyu Chen, Kenneth W. Shum, Quan Yu, and Chi Wan Sung. Sector-disk codes and partial MDS codes with up to three global parities. In *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT '15)*, pages 1876–1880, 2015.

[DGOS16] Zeev Dvir, Ankit Garg, Rafael Oliveira, and József Solymosi. Rank bounds for design matrices with block entries and geometric applications. arXiv 1610.08923, 2016. `https://arxiv.org/abs/1610.08923`.

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

[DH16] Zeev Dvir and Guangda Hu. Sylvester-Gallai for arrangements of subspaces. *Discrete & Computational Geometry*, 56(4):940–965, 2016.

[Don06] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[DS07] Zeev Dvir and Amir Shpilka. Locally decodable codes with two queries and polynomial identity testing for depth 3 circuits. *SIAM Journal on Computing*, 36(5):1404–1434, 2007.

[DSW14a] Zeev Dvir, Shubhangi Saraf, and Avi Wigderson. Breaking the quadratic barrier for 3-LCC's over the reals. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC '14)*, pages 784–793, 2014.

[DSW14b] Zeev Dvir, Shubhangi Saraf, and Avi Wigderson. Improved rank bounds for design matrices and a new proof of Kelly's theorem. *Forum of Mathematics, Sigma*, 2(e4), 2014.

[Dvi11] Zeev Dvir. On matrix rigidity and locally self-correctable codes. *Computational Complexity*, 20(2):367–388, 2011.

[Efr12] Klim Efremenko. 3-query locally decodable codes of subexponential length. *SIAM Journal on Computing*, 41(6):1694–1703, 2012.

[EPS06] Noam Elkies, Lou M. Pretorius, and Konrad J. Swanepoel. Sylvester-Gallai theorems for complex numbers and quaternions. *Discrete & Computational Geometry*, 35(3):361–373, 2006.

[Erd43] Paul Erdös. Problem 4065, in Problems for solutions. *The American Mathematical Monthly*, 50(1):65, 1943.

[FF93] Joan Feigenbaum and Lance Fortnow. Random-self-reducibility of complete sets. *SIAM Journal on Computing*, 22(5):994–1005, 1993.

[FGT16] Stephen Fenner, Rohit Gurjar, and Thomas Thierauf. Bipartite perfect matching is in quasi-NC. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC '16)*, pages 754–763, 2016.

[GHJY14] Parikshit Gopalan, Cheng Huang, Bob Jenkins, and Sergey Yekhanin. Explicit maximally recoverable codes with locality. *IEEE Transactions on Information Theory*, 60(9):5245–5256, 2014.

[GHK+17] Parikshit Gopalan, Guangda Hu, Swastik Kopparty, Shubhangi Saraf, Carol Wang, and Sergey Yekhanin. Maximally recoverable codes for grid-like topologies. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '17)*, 2017.

[GHSY12] Parikshit Gopalan, Cheng Huang, Huseyin Simitci, and Sergey Yekhanin. On the locality of codeword symbols. *IEEE Transactions on Information Theory*, 58(11):6925–6934, 2012.

[GKST06] Oded Goldreich, Howard Karloff, Leonard J. Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. *Computational Complexity*, 15(3):263–296, 2006.

[GLR+91] Peter Gemmell, Richard J. Lipton, Ronitt Rubinfeld, Madhu Sudan, and Avi Wigderson. Self-testing/correcting for polynomials and for approximate functions. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing (STOC '91)*, pages 33–42, 1991.

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

[HCL07] Cheng Huang, Minghua Chen, and Jin Li. Pyramid codes: Flexible cchemes to trade space for access efficiency in reliable data storage systems. In *Proceedings of the 6th IEEE International Symposium on Network Computing and Applications (NCA '07)*, pages 79–86, 2007.

[Hil73] Anthony J. W. Hilton. On double diagonal and cross latin squares. *Journal of the London Mathematical Society*, s2-6(4):679–689, 1973.

[HSX$^+$12] Cheng Huang, Huseyin Simitci, Yikang Xu, Aaron Ogus, Brad Calder, Parikshit Gopalan, Jin Li, and Sergey Yekhanin. Erasure coding in windows azure storage. In *Proceedings of the 2012 USENIX Annual Technical Conference (USENIX ATC '12)*, pages 15–26, 2012.

[HY16] Guangda Hu and Sergey Yekhanin. New constructions of SD and MR codes over small finite fields. In *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT '16)*, pages 1591–1595, 2016.

[IS10] Toshiya Itoh and Yasuhiro Suzuki. Improved constructions for query-efficient locally decodable codes of subexponential length. *IEICE Transactions on Information and Systems*, E93-D(2):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(3):395–420, 2004.

[Kel86] Leroy M. Kelly. A resolution of the Sylvester-Gallai problem of J.-P. Serre. *Discrete & Computational Geometry*, 1(2):101–104, 1986.

[KORW12] Guy Kindler, Ryan O'Donnell, Anup Rao, and Avi Wigderson. Spherical cubes: Optimal foams from computational hardness amplification. *Communications of the ACM*, 55(10):90–97, 2012.

[KSY14] Swastik Kopparty, Shubhangi Saraf, and Sergey Yekhanin. High-rate codes with sublinear-time decoding. *Journal of the ACM*, 61(5):28:1–28:20, 2014.

[KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC '00)*, 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(5):1952–1969, 2009.

[Lax07] Peter D. Lax. *Linear Algebra and Its Applications.* Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, 2007.

[Lev87] Leonid A. Levin. One way functions and pseudorandom generators. *Combinatorica*, 7(4):357–363, 1987.

[Lip90] Richard J. Lipton. Efficient checking of computations. In *Proceedings of the 7th Annual Symposium on Theoretical Aspects of Computer Science (STACS '90)*, pages 207–215, 1990.

[LL15] V. Lalitha and Satyanarayana V. Lokam. Weight enumerators and higher support weights of maximally recoverable codes. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton '15)*, pages 835–842, 2015.

[LN97] Rudolf Lidl and Harald Niederreiter. *Finite Fields*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1997.

[Mel40] Eberhard Melchior. Über vielseite der projektiven ebene. *Deutsche Mathematik*, 5:461–475, 1940.

[MLR$^+$14] Subramanian Muralidhar, Wyatt Lloyd, Sabyasachi Roy, Cory Hill, Ernest Lin, Weiwen Liu, Satadru Pan, Shiva Shankar, Viswanath Sivakumar, Linpeng Tang, and Sanjeev Kumar. f4: Facebook's warm BLOB storage system. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI '14)*, pages 383–398, 2014.

[MS77] Florence J. MacWilliams and Neil J. A. Sloane. *The Theory of Error Correcting Codes*. North-Holland Mathematical Library. North-Holland Publishing Company, 1977.

[PD14] Dimitris S. Papailiopoulos and Alexandros G. Dimakis. Locally repairable codes. *IEEE Transactions on Information Theory*, 60(10):5843–5855, 2014.

[PGM13] James S. Plank, Kevin M. Greenan, and Ethan L. Miller. Screaming fast Galois field arithmetic using Intel SIMD instructions. In *Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST '13)*, pages 299–306, 2013.

[PKLK12] N. Prakash, Govinda M. Kamath, V. Lalitha, and P. Vijay Kumar. Optimal linear codes with a local-error-correction property. In *Proceedings of the 2012 IEEE International Symposium on Information Theory (ISIT '12)*, pages 2776–2780, 2012.

[Rag07] Prasad Raghavendra. A note on Yekhanin's locally decodable codes. Electronic Colloquium on Computational Complexity (ECCC) TR07-016, 2007. `http://eccc.hpi-web.de/report/2007/016`.

[RR72] Sudhakar M. Reddy and John P. Robinson. Random error and burst correction by iterated codes. *IEEE Transactions on Information Theory*, 18(1):182–185, 1972.

[Syl93] James J. Sylvester. Mathematical question 11851. *Mathematical Questions and Solutions from the "Educational Times"*, 59:98, 1893.

[TB14] Itzhak Tamo and Alexander Barg. A family of optimal locally recoverable codes. *IEEE Transactions on Information Theory*, 60(8):4661–4676, 2014.

[TPD13] Itzhak Tamo, Dimitris S. Papailiopoulos, and Alexandros G. Dimakis. Optimal locally repairable codes and connections to matroid theory. In *Proceedings of the 2013 IEEE International Symposium on Information Theory (ISIT '13)*, pages 1814–1818, 2013.

[Woo07] David P. Woodruff. New lower bounds for general locally decodable codes. Electronic Colloquium on Computational Complexity (ECCC) TR07-006, 2007. `http://eccc.hpi-web.de/report/2007/006`.

[Woo12] David P. Woodruff. A quadratic lower bound for three-query linear locally decodable codes over any field. *Journal of Computer Science and Technology*, 27(4):678–686, 2012.

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

[Yek12] Sergey Yekhanin. Locally decodable codes. *Foundations and Trends in Theoretical Computer Science*, 6(3):139–255, 2012.