UNIVERSITY OF CALIFORNIA SAN DIEGO

What not to do when your data is lost ?

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Machine Learning and Data Science)

by

Sankeerth Rao Karingula

Committee in charge:

Professor Shachar Lovett, Chair
Professor Alex Vardy, Co-Chair
Professor Russell Impagliazzo
Professor Daniel Kane
Professor Alon Orlitsky

2020

The Dissertation of Sankeerth Rao Karingula is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

DEDICATION

I dedicate this dissertation to my father. I do not have words to describe all the wonderful things he has done for me.

EPIGRAPH

We have two lives, and the second begins when we realize we only have one.

Confucius

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor Shachar Lovett for his constant support. I am very fortunate to have someone like him to constantly learn from. He was always present to provide guidance, ask questions, suggest ways to handle various situations all through my PhD. I learnt a lot about how to think of problems that are of technical interest and also appeal to the community. Shachar thinks out aloud and taught me a great deal about problem solving techniques and changed how I think about mathematics. Every meeting with him was very inspiring and he is never afraid of picking up any problem, learning the tools needed, innovate and solve it or atleast improve the current knowledge on it significantly. I also learned a lot from Shachar in terms of his importance to family and how to balance various aspects of life and trying to be super efficient at whatever one does. These and many other skills I learnt from him are very valuable to me.

I would like to thank my co-advisor Professor Alexander Vardy for his support. Alex was very kind to me in times of need at various phases of my PhD and I am very grateful for that.

I would like to thank Professor Daniel Kane for wonderful courses and our regular weekly meetings. I admire his problem solving skill and there is always something to learn from every session.

I would like to thank Sergey Yekhanin for hosting me at MSR during the internship and for all the support and help. Sergey was very caring towards me and always ensured I was having a great time at MSR. It was a very valuable experience to work on various projects while at MSR with Sergey

I would like to thank Professor Russell Impagliazzo for his beautiful courses and for hosting me for various games at his home. Russell is a sweet person that is always warm to speak to. His big picture on research is very beautiful as opposed to just problem solving something I aspire to.

I would like to thank Professor Alon Orlitsky for being part of my committee.

I would like to thank all my friends, teachers and coaches for support and guidance all

through my research career.

I am extremely grateful for my parents for all the wonderful things they have done for me all through my life. There was never a single moment in their lives when I was not their first priority. No matter wherever, whenever or whatever obstacle I had faced, all I had to do was to mention it to my father and he would just put everything else aside and fix that issue for me asap no matter what. There was no extent to the sacrifices they have made for me and I am forever indebted to their unconditional love. Every step of my life has been deeply supported by my parents and I totally dedicate this thesis to my father.

Chapter 1, in full, is a reprint of the material as it appears in IEEE Foundations of Computer Science FOCS 2017 and SIAM Journal on Computing SICOMP. Kane, Daniel; Lovett, Shachar; Karingula, Sankeerth Rao. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is currently under review for publication of the material. Karingula, Sankeerth Rao; Lovett, Shachar. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a reprint of the material as it appears in SIAM Symposium on Discrete Algorithms SODA 2018 and Journal of Combinatorial Theory Series A JCTA. Karingula, Sankeerth Rao; Lovett, Shachar; Vardy, Alexander. The dissertation author was the primary investigator and author of this paper.

VITA

2014 BTech + MTech, Indian Institute of Technology Bombay

2020 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

1. “The independence number of the Birkhoff polytope graph, and applications to maximally recoverable codes” Foundations of Computer Science FOCS 2017 and SIAM Journal on Computing SICOMP.

2. ”Probabilistic existence of large sets of designs” Symposium on Discrete Algorithms SODA 2018 and Journal of Combinatorial Theory Series A JCTA.

3. ”An almost optimal PRG for Boolean quadratic PTFs” Computational complexity conference CCC 2018.

4. ”Codes over integers, and the singularity of random matrices with large entries” submitted to STOC 2020.

5. ”Torus polynomials: an algebraic approach to ACC lower bounds” Innovations in Theoretical Computer Science ITCS 19

6. ”Communication and Memory Efficient Testing of Discrete Distributions” Conference on Learning Theory COLT 19

7. ”Lower bounds on the redundancy of PIR Codes” Arxiv preprint.

FIELDS OF STUDY

Major Field: Machine Learning and Data Science.

ABSTRACT OF THE DISSERTATION

What not to do when your data is lost ?

by

Sankeerth Rao Karingula

Doctor of Philosophy in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2020

Professor Shachar Lovett, Chair
Professor Alex Vardy, Co-Chair

With ever increasing amount of digital data being generated everyday on various platforms the need for data storage techniques has increased tremendously. A central component of all data storage techniques are error correction codes. An ideal error correcting code is tolerant to noise, minimally redundant and computationally inexpensive. Figuring out the optimal trade offs between these properties forms the central theme of coding theory. In this thesis we will formulate a central question that underlies the computational performance of all error correction codes and answer this question in various contexts.

Introduction

There is a ginormous amount of digital data being generated on the internet every day. Here are the statistics from a Forbes survey [1] on this:

- **Social media**: 300 million new social media users each year
- **Email use**: 293 billion emails are sent daily
- **Mobile device data**: 21.9 billion text messages are sent daily
- **Internet of Things**: 2.5 quintillion bytes of data we create every day
- **Data Generating Services**: Amazon, Uber, Venmo,...

Every big tech company such as Amazon, Apple, Google, Facebook and Microsoft needs to deal with this ever growing data. They handle this by building many data centers all across the world. Every data center contains many servers that collectively store all the relevant data. It is a very active research area and new architectures and various ways of organizing the data are currently being explored. However the servers in such a data center could be unresponsive or there could be failures which could result in loss of data and lead to inability to respond to queries by the users.

To handle these issues error correcting codes are used and coding theory is the study of theory of error correcting codes. A code is a way to encode messages to codewords such that few errors introduced in the codeword could be corrected and the original message can be recovered. Codes are very useful in various contexts like data storage, data transmission, probabilistically

checkable proofs, etc. An ideal error correcting code is tolerant to noise, minimally redundant and computationally inexpensive. Figuring out the optimal trade offs between these properties forms the central theme of coding theory.

The computational cost of implementing an error correction code is directly related to the underlying alphabet the code uses. So a central question that comes up in various contexts in coding theory is whether a given code structure can be realized over a given alphabet size.

This is best understood in terms of the parity check matrix of the code. Every linear code is completely characterized by its parity check matrix H. The parity check matrix is a matrix whose rows span all the parity relations that the codewords in the linear code satisfy. In particular an erasure pattern is correctable by a linear error correcting code if and only if the corresponding columns of the parity check matrix are linearly independent.

So a typical application of coding theory in data storage would satisfy the following template: At first one looks at the statistics of data failures and makes a note of the kind of robustness we expect from the code in the needed data storage application. Then one comes up with a code topology that can provide the needed robustness against these failures. Once the code topology is fixed one tries to understand when does a code construction exist that meets the needed topology and application robustness specifications. Then the natural question that arises is to figure out the minimal field sizes needed for the existence of such a code. In order to do so one typically uses the algebraic structure of the problem to construct a lower bound on the alphabet size needed. The upper bounds are typically given with random and explicit constructions. However unless there is a magical construction most typically there is a huge gap between the upper and lower bounds. This has been a recurrent theme across many data storage applications where coding theory has been used.

The goal of this thesis is to build a general framework that combines all these different practical applications of coding theory and address the underlying issues in a systematic way and develop new tools and approaches to attack these problems. We now define this framework and see how various applications of coding theory fall into this framework and solve these problems

along the way.

So let's consider general a parity check matrix H and picking the entries of this matrix would correspond to construction the relevant codes. The topological constraints of the code would force some of the entries of the matrix H to be 0 while we are freely allowed to pick the other entries which correspond to the respective parity coefficients, let's denote these positions with *'s. Then the problem of seeking the minimum field size would reduce to finding the smallest alphabet A from which the entries of the matrix H in the positions of *'s can be filled in so that all the needed minors of the matrix H are non zero.

More concretely, given a $k \times n$ matrix with a given pattern of 0's and $*$'s, what is the optimal field size q so that there is a way to assign values to $*$'s from $\mathbb{F}_q$ such that a given set of minors of the matrix are non zero. Note that this needed set of minors comes from the specific coding application at hand. In the most optimistic framework like the maximally recoverable codes one expects every non trivial minor not trivially killed by the 0's to be non zero. However as we shall see depending on the coding application needs this requirement could be different and relaxed most of the times.

$$M = \begin{bmatrix} * & 0 & * & \cdots & 0 \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & 0 & \cdots & * \end{bmatrix}$$

This 0/* question is in coding theory and exactly captures the difficulties arising in effective data storage. For instance an all $*$ matrix would correspond to maximum distance separable codes (MDS), an upper triangular $*$ matrix would correspond to optimal codes for interactive communication (Tree codes), there would be a choice of $*$'s for maximally recoverable codes for every topology. The constructions in these specific instances are rather hit or miss, for instance a Reed Solomon code would give an MDS code with linear field size $q = O(n)$ where as a random construction would need $q = \Omega(n^k)$ and there is no understanding as to why this

particular construction works.

For the specific pattern that corresponds to maximally recoverable codes for the grid topology we characterized the field size needed exactly [24]. In particular we had an exponential improvement over the best known lower bound. This was done by reducing the problem to bounding the independence number of the Birkoff polytope and then using techniques from representation theory of S_n. Our result is presented in detail in Chapter 1.

It would be very interesting to build an unified theory that can explain the field size needed as we vary these patterns of 0's and $*$'s. In particular we are working on understanding the field size needed if the $*$'s and 0's were generated randomly. As a first step towards this we worked with codes over integers $\mathbb{Z}$ where in we fill the $*$'s with entries from $\{0,1,\ldots,m\} \subset \mathbb{Z}$ and see how large m needs to be to facilitate an union bound.

This leads to beautiful questions in random matrix theory. In particular if an $n \times n$ matrix is filled with random entries from $\{0,1,\ldots,m\}$ what is the probability that its singular ? We will present our results on this in chapter 2.

In chapter 3 we extend these probabilistic techniques to show the existence of rare combinatorial structures and in particular prove the existence of large sets of designs.

Chapter 1

Maximally recoverable codes

Maximally recoverable codes are codes designed for distributed storage which combine quick recovery from single node failure and optimal recovery from catastrophic failure. Gopalan et al [SODA 2017] studied the alphabet size needed for such codes in grid topologies and gave a combinatorial characterization for it.

Consider a labeling of the edges of the complete bipartite graph $K_{n,n}$ with labels coming from $\mathbb{F}_2^d$, that satisfies the following condition: for any simple cycle, the sum of the labels over its edges is nonzero. The minimal d where this is possible controls the alphabet size needed for maximally recoverable codes in $n \times n$ grid topologies.

Prior to the current work, it was known that d is between $\log(n)^2$ and $n \log n$. We improve both bounds and show that d is linear in n. The upper bound is a recursive construction which beats the random construction. The lower bound follows by first relating the problem to the independence number of the Birkhoff polytope graph, and then providing tight bounds for it using the representation theory of the symmetric group.

1.1 Introduction

The Birkhoff polytope is the convex hull of $n \times n$ doubly stochastic matrices. The Birkhoff polytope graph is the graph associated with its 1-skeleton. This graph is well studied as it plays an important role in combinatorics and optimization, see for example the book of Barvinok [5].

For us, this graph arose naturally in the study of certain maximally recoverable codes. Our main technical results are tight bounds on the independence number of the Birkhoff polytope graph, which translate to tight bounds on the alphabet size needed for maximally recoverable codes in grid topologies.

We start by describing the coding theory question that motivated the current work.

1.1.1 Maximally recoverable codes

Maximally recoverable codes, first introduced by Gopalan, Huang, Jenkins and Yekhanin [19], are codes designed for distributed storage which combine quick recovery from single node failure and optimal recovery from catastrophic failure. More precisely, they are systematic linear codes which combine two types of redundancy symbols: local parity symbols, which allow for fast recovery from single symbol erasure; and global parity symbols, which allow for recovery from the maximal information theoretic number of erasures. This was further studied in [3, 32, 46, 59, 60].

The present paper is motivated by a recent work of Gopalan, Hu, Kopparty, Saraf, Wang and Yekhanin [18], which studied the effect of the topology of the network on the code design. Concretely, they studied grid like topologies. In the simplest setting, a codeword is viewed as an $n \times n$ array, with entries in a finite field $\mathbb{F}_{2^d}$, where there is a single parity constraint for each row and each column, and an additional global parity constraint. More generally, a $T_{n\times m}(a,b,h)$ maximally recoverable code has codewords viewed as an $n \times m$ matrix over $\mathbb{F}_2^d$, with a parity constraints per row, b parity constraints per column, and h additional global parity constraints. An important problem in this context is, how small can we choose the alphabet size 2^d and still achieve information theoretical optimal resiliency against erausers.

Gopalan et al. [18] gave a combinatorial characterization for this problem, in the simplest setting of $m = n$ and $a = b = h = 1$. Their characterization is in terms of labeling the edges of the complete bipartite graph $K_{n,n}$ by elements of $\mathbb{F}_2^d$, which satisfy the property that in every simple cycle, the sum is nonzero.

Let $[n] = \{1, \ldots, n\}$. Let $\gamma : [n] \times [n] \to \mathbb{F}_2^d$ be a labeling of the edges of the complete

bipartite graph $K_{n,n}$ by bit vectors of length d.

Definition 1.1.1. *A labeling $\gamma : [n] \times [n] \to \mathbb{F}_2^d$ is* simple cycle free *if for any simple cycle C in $K_{n,n}$ it holds that*

$$\sum_{e \in C} \gamma(e) \neq 0.$$

Gopalan et al. [18] showed that the question on the minimal alphabet size needed for maximally recoverable codes, reduces to the question of how small can we take $d = d(n)$ so that a simple cycle free labeling exists. Concretely:

- The alphabet size needed for $T_{n \times n}(1,1,1)$ codes is $2^{d(n)}$.
- The alphabet size needed for $T_{n \times m}(a,b,h)$ codes is at least $2^{\min(d(n-a+1),d(m-b+1))/h}$.

Before the current work, there were large gaps between upper and lower bounds on $d(n)$. For upper bounds, as the number of simple cycles in $K_{n,n}$ is $n^{O(n)}$, a random construction with $d = O(n \log n)$ succeeds with high probability. There are also simple explicit constructions matching the same bounds, see e.g. [19]. In terms of lower bounds, it is simple to see that $d \geq \log n$ is necessary. The main technical lemma of Gopalan et al. [18] in this context is that in fact $d \geq \Omega(\log^2 n)$ is necessary. This implies a super-polynomial lower bound on the alphabet size 2^d in terms of n, which is one of their main results.

We improve on both upper and lower bounds and show that d is linear in n. We note that our construction improves upon the random construction, which for us was somewhat surprising. For convenience we describe it when n is a power of two, but note that it holds for any n with minimal modifications.

Theorem 1.1.2 (Explicit construction)**.** *Let n be a power of two. There exists $\gamma : [n] \times [n] \to \mathbb{F}_2^d$ for $d = 3n$ which is simple cycle free.*

Our main technical result is a nearly matching lower bound.

Theorem 1.1.3 (Lower bound)**.** *Let $\gamma : [n] \times [n] \to \mathbb{F}_2^d$ be simple cycle free. Then $d \geq n/2 - 2$.*

1.1.2 Labeling by general Abelian groups

The definition of simple cycles free labeling can be extended to labeling by general Abelian groups, not just $\mathbb{F}_2^d$. Let H be an Abelian group, and let $\gamma : [n] \times [n] \to H$. We say that γ is simple cycle free if for any simple cycle C,

$$\sum_{e \in C} \text{sign}(e)\gamma(e) \neq 0.$$

where $\text{sign}(e) \in \{-1, 1\}$ is an alternating sign assignment to the edges of C (these are sometimes called circulations). We note that the analysis of Gopalan et al. [18] can be extended to non-binary alphabets $\mathbb{F}_p$, in which case their combinatorial characterization extends to the one above with $H = \mathbb{F}_p$.

Theorem 1.1.4. *Let H be an Abelian group. Let $\gamma : [n] \times [n] \to H$ be simple cycle free. Then $|H| \geq 2^{n/2-2}$.*

As a side remark, we note that the study of graphs with nonzero circulations was instrumental in the recent construction of a deterministic quasi-polynomial algorithm for perfect matching in NC [17]. However, beyond some superficial similarities, the setup seems inherently different than ours. For starters, they study general bipartite graphs, while we study the complete graphs. Moreover, they need to handle certain families of cycles, not necessarily simple, while in this work we focus on simple cycles.

The proofs of Theorem 1.1.3 and Theorem 1.1.4 rely on the study of a certain Cayley graph of the permutation group, which encodes the property of simple cycle free labeling. Surprisingly, the corresponding graph is the Birkhoff polytope graph.

1.1.3 The Birkhoff polytope graph

Let S_n denote the symmetric group of permutations on $[n]$. A permutation $\tau \in S_n$ is said to be a *cycle* if, except for its fixed points, it contains a single non-trivial cycle (in

particular, the identity is not a cycle). We denote by $\mathscr{C}_n \subset S_n$ the set of cycles. The Cayley graph $\mathscr{B}_n = \mathrm{Cay}(S_n, \mathscr{C}_n)$ is a graph with vertex set S_n and edge set $\{(\pi, \tau\pi) : \pi \in S_n, \tau \in \mathscr{C}_n\}$. Note that this graph is undirected, as if $\tau \in \mathscr{C}_n$ then also $\tau^{-1} \in \mathscr{C}_n$.

The graph $\mathscr{B}_n$ turns out to be widely studied: it is the graph of the Birkhoff polytope, which is the convex hull of all $n \times n$ permutation matrices. See for example [6] for a proof. Our analysis does not use this connection; we use the description of the graph as a Cayley graph.

The following claim shows that Theorem 1.1.4 reduces to bounding the size of the largest independent set in the Birkhoff polytope graph.

Claim 1.1.5. *Let H be an Abelian group. Assume that $\gamma : [n] \times [n] \to H$ is simple cycle free. Then $\mathscr{B}_n$ contains an independent set of size $\geq n!/|H|$.*

Proof. Define

$$A = \left\{ \pi \in S_n : \sum_{i=1}^{n} \gamma(i, \pi(i)) = h \right\},$$

where $h \in H$ is chosen to maximize the size of A. Thus $|A| \geq n!/|H|$. We claim that A is an independent set in $\mathscr{B}_n$.

Assume not. Then there are two permutations $\pi, \pi' \in A$ such that $\tau = \pi(\pi')^{-1} \in \mathscr{C}_n$. Let $M_\pi = \{(i, \pi(i)) : i \in [n]\}$ denote the matching in $K_{n,n}$ associated with π, and define $M_{\pi'}$ analogously. Let $C = M_\pi \oplus M_{\pi'}$ denote their symmetric difference. The fact that $\tau \in \mathscr{C}_n$ has exactly one cycle, is equivalent to C being a simple cycle. Let $\mathrm{sign}(\cdot)$ be an alternating sign assignment to the edges of C. Then

$$\sum_{e \in C} \mathrm{sign}(e)\gamma(e) = \sum_{e \in M_\pi} \gamma(e) - \sum_{e \in M_{\pi'}} \gamma(e) = h - h = 0.$$

This violates the assumption that γ is simple cycle free. □

The construction of a simple cycle free labeling in Theorem 1.1.2, combined with Claim 1.1.5, implies that the Birkhoff polytope graph contains a large independent set.

Corollary 1.1.6. *Let n be a power of two. Then $\mathscr{B}_n$ contains an independent set of size $\geq n!/9^n$.*

We also give in the appendix a construction of a larger independent set in the Birkhoff polytope graph, not based on a simple cycle free labeling.

Theorem 1.1.7. *Let n be a power of two. Then $\mathscr{B}_n$ contains an independent set of size $\geq n!/4^n$.*

The best previous bounds we are aware of are by Onn [45] who proved that $\mathscr{B}_n$ contains an independent set of size $\geq n^{\Omega(\sqrt{n})}$.

Our main technical result is an upper bound on the largest size of an independent set in the Birkhoff polytope graph.

Theorem 1.1.8. *The largest independent set in $\mathscr{B}_n$ has size $\leq n!/2^{(n-4)/2}$.*

As a side remark, we note that general bounds on the independence number of graphs, such as the Hoffman bound, give much weaker bounds. A standard application of the Hoffman bound gives a much weaker bound for the independence number of $\mathscr{B}_n$ of $O(n!)$; and if we restrict all permutations to have the same sign, the bound improves to $O((n-1)!)$. The reason is that the Hoffman bounds (at least in its simplest form) directly relates to the minimal eigenvalues of the graph. However, in our case the eigenvalues are controlled by the irreducible representations of S_n, and the extreme eigenvalues are given by low dimensional representations. This prohibits obtaining strong bounds on the independence number directly.

In order to overcome this barrier, our analysis circumvents the effect of the low dimensional representations by appealing to a structure vs. randomness dichotomy specialized for our setting. It allows us to either reduce the dimension of the ambient group, or restrict to pseudo-random assumptions about the actions of the low dimensional representations.

Section outline

We prove Theorem 1.1.2 in Section 1.2 and Theorem 1.1.8 in Section 1.3. Theorem 1.1.7 is proved in Appendix 1.4.

1.2 A construction of a simple cycle free labeling

We prove Theorem 1.1.2 in this section. We first introduce some notation. For $x \in [n]$ denote by $e_x^n \in \mathbb{F}_2^n$ the unit vector with 1 in coordinate x and 0 in all other coordinates. We let $0^n \in \mathbb{F}_2^n$ denote the all zero vector.

Let n be a power of two. We define recursively a labeling $\gamma_n : [n] \times [n] \to \mathbb{F}_2^{3n}$. For $n = 2$ set (for example)

$$\gamma_2(0,0) = e_1^6, \gamma_2(0,1) = e_2^6, \gamma_2(1,0) = e_3^6, \gamma_2(1,1) = e_4^6.$$

Assume $n > 2$. Let $x' = x \mod (n/2)$ and $y' = y \mod (n/2)$, where $x', y' \in [n/2]$. Define $\gamma_n(x,y) \in \mathbb{F}_2^{3n}$ recursively as

(i) The first n bits of $\gamma_n(x,y)$ are e_x^n if $y \leq n/2$, and otherwise they are 0^n.

(ii) The next $n/2$ bits of $\gamma_n(x,y)$ are $e_{y'}^{n/2}$ if $x \leq n/2$, and otherwise they are $0^{n/2}$.

(iii) The last $3n/2$ bits of $\gamma_n(x,y)$ are defined recursively to be $\gamma_{n/2}(x',y')$.

We claim that γ_n is indeed simple cycle free. For $n = 2$ it is simple to verify this directly, so assume $n > 2$.

Let C be a simple cycle in $K_{n,n}$, and assume towards a contradiction that $\sum_{e \in C} \gamma_n(e) = 0$. Assume C has $2k$ nodes, for some $2 \leq k \leq n$, and let these be $C = (x_1, y_1, x_2, y_2, \ldots, x_k, y_k, x_1)$. We denote $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$. Define furthermore $L = \{1, \ldots, n/2\}$ and $U = \{n/2+1, \ldots, n\}$.

Claim 1.2.1. *Either $Y \subset L$ or $Y \subset U$.*

Proof. Assume that both $Y \cap L$ and $Y \cap U$ are nonempty. Then there must exist $i \in [k]$ with $y_i \in L$ and $y_{i+1} \in U$, where if $i = k$ then we take the subscript modulo k. Recall that x_{i+1} is the neighbour of y_i, y_{i+1} in C. Its contribution to the first n bits of the sum is $e_{x_{i+1}}^n$, since $y_i \leq n/2$

and $y_{i+1} > n/2$. Note that no other edge in C has a nonzero value in coordinate x_{i+1}. Thus the x_{i+1} coordinate in the sum over C is 1, which contradicts the assumption that the sum over C is zero. □

Thus we can assume from now on that either $Y \subset L$ or $Y \subset U$.

Claim 1.2.2. *Either $X \subset L$ or $X \subset U$.*

Proof. Assume that $Y \subset L$, and the case of $Y \subset U$ is identical. Assume that both $X \cap L$ and $X \cap U$ are both nonempty. Then there must exist $i \in [k]$ with $x_i \in L$ and $x_{i+1} \in U$. Recall that y_i is the neighbour of x_i, x_{i+1} in C. Its contribution to the 2nd batch (of $n/2$ bits) of the sum is $e^{n/2}_{y'_i}$, since $x_i \leq n/2$ and $x_{i+1} > n/2$. Note that no other edge in C has a nonzero value in coordinate $n + y'_i$, where we here we need the assumption that $Y \subset L$ or $Y \subset U$. Thus the $n + y'_i$ coordinate in the sum over C is 1, which contradicts the assumption that the sum over C is zero. □

Thus we have that $X \subset U$ or $X \subset L$, and similarly $Y \subset U$ or $Y \subset L$. Thus, C is a simple cycle in $K_{n/2,n/2}$ embedded in $K_{n,n}$ in one of four disjoint ways: $L \times L$, $L \times U$, $U \times L$ or $U \times U$. Observe that in each of these copies, the last $3n/2$ coordinates of the sum are precisely $\gamma_{n/2}$, so by induction C cannot have zero sum.

1.3 The independence number of the Birkhoff polytope graph

We prove Theorem 1.1.8 in this section. Let A be an independent set in $\mathscr{B}_n$. We prove an upper bound on the size of A. Concretely, we will show that $|A| \leq \frac{a}{c^n} n!$ for some absolute constants $a, c > 1$. As we will see at the end, the choice of $a = 4, c = \sqrt{2}$ works.

The proof relies on representation theory, in particular representation theory of the symmetric group. We refer readers to the excellent book of Sagan [51], which provides a thorough introduction to the topic. We will try to adhere to the notations in that book whenever possible.

Overall Strategy.

Our basic plan will be to break our analysis into two cases based on whether or not the action of A on m-tuples is nearly uniform for all m. This will be in analogy with standard structure vs. randomness arguments. If the action on m-tuples is highly non-uniform, this will allow us to take advantage of this non-uniformity to reduce to a lower-dimensional case. On the other hand, if A acts nearly uniformly on m-tuples, this suggests that it behaves somewhat randomly. This intuition can be cashed out usefully by considering the Fourier-analytic considerations of this condition, which will allow us to prove that some pair of elements of A differ by a simple cycle using Fourier analysis on S_n.

Non-Uniform Action on Tuples.

Let $[n]_m = \{(i_1,\dots,i_m) : i_1,\dots,i_m \in [n] \text{ distinct}\}$ denote the family of ordered m-tuples of distinct elements of $[n]$. Its size is $(n)_m = n(n-1)\cdots(n-m+1)$. A permutation $\pi \in S_n$ acts on $[n]_m$ by sending $I = (i_1,\dots,i_m)$ to $\pi(I) = (\pi(i_1),\dots,\pi(i_m))$. Below when we write $\Pr_{\pi\in A}[\cdot]$ we always mean the probability of an event under a uniform choice of $\pi \in A$.

Notice that if $\Pr_{\pi\in A}[\pi(I) = J] \geq c^m/(n)_m$ for some pair $I,J \in [n]_m$, this will allow us to reduce to a lower dimensional version of the problem. In particular, if we let $A' = \{\pi \in A : \pi(I) = J\}$, we note that $|A| \leq |A'|(n)_m/c^m$. On the other hand, after multiplying on the left and right by appropriate permutations (an operation which doesn't impact our final problem), we can assume that $I = J = \{n-m+1,\dots,n\}$. Then, if A were an independent set for $\mathscr{B}_n$, A' would correspond to an independent set for $\mathrm{Cay}(S_{n-m},\mathscr{C}_{n-m})$. Then, if we could prove the bound that $|A'| \leq \frac{a}{c^{n-m}}(n-m)!$, we could inductively prove that $|A| \leq \frac{a}{c^n}n!$.

Uniform Action on Tuples.

When the action of A on m-tuples is near uniform for all m, we will attempt to show that two elements of A differ by a simple cycle using techniques from the Fourier analysis of S_n. In fact, we will show the stronger statement that some pair of elements of A differ by a single cycle of length n.

Some slight complications arise here when parity of the permutations here is considered. In particular, all n-cycles have the same parity. This is actually a problem for n even, as all such cycles will be odd, and thus our statement will fail if A consists only of permutations with the same parity. Thus, we will have to consider our statement only in the case of n odd. Even in this case though, parity will still be relevant. In particular, note that the difference between two permutations in A can be a cycle of length n only if the initial permutations had the same parity. Thus, we lose very little by restricting our attention to only elements of A with the more common parity. This will lose us a factor of 2 in the size of A, but will make our analysis somewhat easier. We are now prepared to state our main technical proposition:

Proposition 1.3.1. *Let n be an odd integer and let $c > 1$ be a sufficiently small constant. Let $A \subset S_n$ be a set of permutations satisfying:*

(i) All elements of A are of the same sign.

(ii) For any even $m < n$ and any $I, J \in [n]_m$, $\Pr_{\pi \in A}[\pi(I) = J] < \frac{c^m}{(n)_m}$.

Then there exist two elements of A that differ by a cycle of length n. In particular, we can take $c = \sqrt{2}$.

Remark.

In the second condition above, we consider only even m. This is because if this condition fails, we are going to use our other analysis to recursively consider permutations of $[n-m]$, and would like $n-m$ to also be odd.

We prove Proposition 1.3.1 below, and then show that it implies Theorem 1.1.8.

Proof. First, note that by replacing all $\pi \in A$ by $\pi\sigma$ for some odd permutation σ if necessary, it suffices to assume that all $\pi \in A$ are even. We will assume this henceforth.

Rephrasing the problem using class functions.

Let $\mathscr{C}'_n$ denote the set of n-cycles in S_n. Define two class functions $\varphi, \psi \in \mathbb{R}[S_n]$ as

$$\varphi = \frac{1}{|S_n||A|^2} \sum_{\sigma \in S_n, \pi, \pi' \in A} \sigma\pi(\pi')^{-1}\sigma^{-1}, \qquad \psi = \frac{1}{|\mathscr{C}'_n|} \sum_{\tau \in \mathscr{C}'_n} \tau.$$

It is easy to see that our conclusion is equivalent to showing that $\langle \varphi, \psi \rangle > 0$.

Let $\lambda \vdash n$ denote a partition of n, namely $\lambda = (\lambda_1, \ldots, \lambda_k)$ where $\lambda_1 \geq \ldots \geq \lambda_k \geq 1$ and $\sum \lambda_i = n$. The irreducible representations of S_n are the Specht modules, which are indexed by partitions $\{S^\lambda : \lambda \vdash n\}$. Let $\chi^\lambda : S_n \to \mathbb{R}$ denote their corresponding characters. Their action extends linearly to $\mathbb{R}[S_n]$. Namely, if $\zeta \in \mathbb{R}[S_n]$ is given by $\zeta = \sum_{\pi \in S_n} \zeta_\pi \pi \in \mathbb{R}[S_n]$ where $\zeta_\pi \in \mathbb{R}$ then $\chi^\lambda(\zeta) = \sum_{\pi \in S_n} \zeta_\pi \chi^\lambda(\pi)$.

As $\varphi, \psi \in \mathbb{R}[S_n]$ are class functions, their inner product equals

$$\langle \varphi, \psi \rangle = \sum_{\lambda \vdash n} \chi^\lambda(\varphi)\chi^\lambda(\psi). \tag{1.1}$$

Let $(n) \in \mathscr{C}'_n$ be a fixed cycle of length n. As all elements in ψ are conjugate to (n), we have $\chi^\lambda(\psi) = \chi^\lambda((n))$ and we can simplify Equation (1.1) to

$$\langle \varphi, \psi \rangle = \sum_{\lambda \vdash n} \chi^\lambda(\varphi)\chi^\lambda((n)). \tag{1.2}$$

Thus, we are lead to explore the action of the irreducible characters on the full cycle (n).

Characters action on the full cycle.

The Murnaghan-Nakayama rule is a combinatorial method to compute the value of a character χ^λ on a conjugacy class, which in our case is (n). In this special case it is very simple. It equals zero unless λ is a hook, e.g. its corresponding tableaux has only one row and one column, and otherwise its either -1 or 1. Concretely, let $h_m = (n-m, 1, 1, \ldots, 1)$ for

$0 \leq m \leq n-1$ denote the partition corresponding to a hook. Then

$$\chi^{\lambda}((n)) = \begin{cases} (-1)^m & \text{if } \lambda = h_m \\ 0 & \text{otherwise} \end{cases}. \tag{1.3}$$

Thus we can simplify Equation (1.2) to

$$\langle \varphi, \psi \rangle = \sum_{m=0}^{n-1} (-1)^m \chi^{h_m}(\varphi). \tag{1.4}$$

Bounding the characters on φ.

The character h_0 corresponds to the trivial representation, and by our definition of φ it equals $\chi^{h_0}(\varphi) = 1$. Observe that we can simplify $\chi^{\lambda}(\varphi)$ as

$$\chi^{\lambda}(\varphi) = \frac{1}{|A|^2|S_n|} \sum_{\pi,\pi' \in A, \sigma \in S_n} \chi^{\lambda}(\sigma\pi(\pi')^{-1}\sigma^{-1}) = \frac{1}{|A|^2} \sum_{\pi,\pi' \in A} \chi^{\lambda}(\pi(\pi')^{-1}). \tag{1.5}$$

First, we argue that the evaluation of characters on φ is always nonnegative.

Claim 1.3.2. *$\chi^{\lambda}(\varphi) \geq 0$ for all $\lambda \vdash n$.*

Proof. Let $\zeta \in \mathbb{R}[S_n]$ be given by $\zeta = \frac{1}{|A|}\sum_{\pi \in A} \pi$. Then

$$\chi^{\lambda}(\varphi) = \frac{1}{|A|^2} \sum_{\pi,\pi' \in A} \mathrm{Tr}\left(S^{\lambda}(\pi) S^{\lambda}((\pi')^{-1})\right) = \mathrm{Tr}\left(S^{\lambda}(\zeta) S^{\lambda}(\zeta)^T\right) = \|S^{\lambda}(\zeta)\|_F^2,$$

where for a matrix M its Frobenius norm is given by $\|M\|_F^2 = \sum |M_{i,j}|^2$. In particular it is always nonnegative. □

The following lemma bounds $\chi^{h_m}(\varphi)$. Observe that in particular for $c = 1$ it gives $\chi^{h_m}(\varphi) = 0$. However, we would use it to obtain effective bounds when $c > 1$.

Lemma 1.3.3. *Let $m \in \{1, \ldots, n-1\}$. For any even $k \in \{m, \ldots, n\}$ it holds that $\chi^{h_m}(\varphi) \leq \frac{c^k - 1}{\binom{k}{m}}$.*

Proof. Let M^{μ} denote the (not irreducible) Young module associated with a partition $\mu \vdash n$. In the case of $\mu = h_k$ it corresponds to the action of S_n on $[n]_k$. That is, for any $\pi \in S_n$ we have that $M^{h_k}(\pi)$ is a matrix whose rows and columns are indexed by $I, J \in [n]_k$ respectively, where $M^{h_k}(\pi)_{I,J} = 1_{\pi(I)=J}$. Observe that $M^{h_k}(\pi^{-1}) = \left(M^{h_k}(\pi)\right)^T$. We extend this action to $\mathbb{R}[S_n]$ linearly.

Recall that $\zeta = \frac{1}{|A|}\sum_{\pi \in A} \pi \in \mathbb{R}[S_n]$. By assumption (ii) in Proposition 1.3.1 we have

$$\left(M^{h_k}(\zeta)\right)_{I,J} = \Pr_{\pi \in A}[\pi(I) = J] \leq \frac{c^k}{(n)_k}.$$

Thus, we can bound the Frobenius norm of $M^{h_k}(\zeta)$ by

$$\|M^{h_k}(\zeta)\|_F^2 = \sum_{I,J} |\left(M^{h_k}(\zeta)\right)_{I,J}|^2 \leq \left(\frac{c^k}{(n)_k}\right) \sum_{I,J} |\left(M^{h_k}(\zeta)\right)_{I,J}| = c^k.$$

This is useful as

$$\mathrm{Tr}(M^{h_k}(\varphi)) = \mathrm{Tr}\left(M^{h_k}(\zeta)\left(M^{h_k}(\zeta)\right)^T\right) = \|M^{h_k}(\zeta)\|_F^2 \leq c^k.$$

The Kostka numbers $K_{\lambda,\mu}$ denote the multiplicity of the Specht module S^{λ} in the Young module M^{μ}. We can thus decompose

$$\mathrm{Tr}(M^{\mu}(\varphi)) = \sum_{\lambda} K_{\lambda,\mu} \chi^{\lambda}(\varphi).$$

We saw that $\chi^{\lambda}(\varphi) \geq 0$ for all λ. By Young's rule, $K_{\lambda,\mu}$ equals the number of semistandard tableaux of shape λ and content μ. In particular, it is always a nonnegative integer. In the special case of $\lambda = h_m$ and $\mu = h_k$ for $k \geq m$, Young's rule is simple to compute and gives

$$K_{h_m,h_k} = \binom{k}{m}.$$

Recall that χ^{h_0} is the trivial representation, for which $K_{h_0,h_k} = 1$ and $\chi^{h_0}(\varphi) = 1$. Thus

$$1 + \binom{k}{m}\chi^{h_m}(\varphi) \leq \sum_{\lambda} K_{\lambda,h_k}\chi^{\lambda}(\varphi) = \mathrm{Tr}(M^{h_k}(\varphi)) \leq c^k.$$

□

We next apply Lemma 1.3.3 to bound $\chi^{h_m}(\varphi)$ for all $1 \leq m \leq n-1$. If $m \leq n/2$ then we can apply Lemma 1.3.3 for $k = 2m$ and obtain the bound

$$\chi^{h_m}(\varphi) \leq \frac{c^{2m}-1}{\binom{2m}{m}}.$$

For $m > n/2$ we need the following claim, relating χ^{h_m} to $\chi^{h_{n-1-m}}$.

Claim 1.3.4. *For any $1 \leq m \leq n-1$ it holds that $\chi^{h_m}(\varphi) = \chi^{h_{n-1-m}}(\varphi)$.*

Proof. For any partition λ let λ' denote the transpose (also known as conjugate) partition. It satisfies $\chi^{\lambda'}(\pi) = \chi^{\lambda}(\pi)\mathrm{sign}(\pi)$ for all $\pi \in S_n$, where $\mathrm{sign} : S_n \to \{-1,1\}$ is the sign representation. As all elements in A are even permutations, it holds by the definition of φ that

$$\chi^{\lambda'}(\varphi) = \frac{1}{|A|^2}\sum_{\pi,\pi' \in A}\chi^{\lambda'}(\pi(\pi')^{-1}) = \frac{1}{|A|^2}\sum_{\pi,\pi' \in A}\chi^{\lambda}(\pi(\pi')^{-1}) = \chi^{\lambda}(\varphi).$$

In particular if $\lambda = h_m$ then $\lambda' = h_{n-1-m}$. □

Next, we lower bound $\langle \varphi, \psi' \rangle$ as follows. The dominant terms are $\chi^{h_0}(\varphi) = \chi^{h_{n-1}}(\varphi) = 1$. For any $1 \leq m \leq (n-1)/2 - 1$, the corresponding term in Equation (1.4) appears twice, once as $(-1)^m\chi^{h_m}(\varphi)$ and once as $(-1)^{n-1-m}\chi^{h_{n-1-m}}(\varphi) = (-1)^m\chi^{h_m}(\varphi)$. The term for $m = (n-1)/2$ appears once.

Furthermore, as $\chi^{h_m}(\varphi) \geq 0$ for all m by Claim 1.3.2, the only negative terms correspond

to odd $1 \leq m \leq (n-1)/2$. Thus we can lower bound

$$\frac{1}{2}\langle \varphi, \psi' \rangle \geq 1 - \sum_{m \geq 1,\, m \text{ odd}} \frac{c^{2m}-1}{\binom{2m}{m}}. \tag{1.6}$$

It is not hard to show that this is positive if $c > 1$ is small enough. If we take $c = \sqrt{2}$, the right hand side of Equation (1.6) is slightly negative for large enough m (the limit as $m \to \infty$ is $-0.02451...$). However, when $n \geq 8$, the second term can be replaced by $\frac{c^8-1}{\binom{8}{3}}$ rather than $\frac{c^6-1}{\binom{6}{3}}$, making our lower bound on $\frac{1}{2}\langle \varphi, \psi' \rangle$ at least 0.057. This completes our proof.

□

We are now prepared to prove Theorem 1.1.8.

Proof. We first prove that if n is odd and if all permutations in A have the same sign, then $|A| \leq \frac{n!}{2^{(n-1)/2}}$.

We proceed by induction on n. Firstly, we note that if $n = 1$, the bound follows trivially.

For odd $n > 1$, we note that unless there is some even $m < n$ and some $I, J \in [n]_m$ with $\Pr_{\pi \in A}[\pi(I) = J] \geq 2^{m/2}/(n)_m$, then our result follows immediately from Proposition 1.3.1. Otherwise, we may assume without loss of generality that $I = J = (n-m+1, \ldots, n)$. It then follows that letting $A' = \{\pi \in A : \pi(I) = J\}$, we can think of A' as a set of permutations on $[n-m]$. Also, note that A being an independent set for $\mathscr{B}_n$, implies that A' is an independent set for $\mathrm{Cay}(S_{n-m}, \mathscr{C}_{n-m})$. Therefore, by the inductive hypothesis:

$$|A| \leq (n)_m 2^{-m/2} |A'| \leq (n)_m 2^{-m/2} (n-m)!/2^{(n-m-1)/2} = n!/2^{(n-1)/2}.$$

We now need to reduce to the case of n odd and A consisting only of permutations of the same sign. First, restricting A to only permutations of the most common sign, we can assume that all permutations in A have the same sign, losing only a factor of 2 in $|A|$. Now, if n is odd, we are done. otherwise, let j be the most likely value of $\pi(n)$ for π taken from A. We have that

$\Pr_{\pi\in A}[\pi(n)=j]\geq 1/n$. Without loss of generality, $j=n$ and we can let $A'=\{\pi\in A:\pi(n)=n\}$. Since A' is an independent set in $\mathrm{Cay}(S_{n-1},\mathscr{C}_{n-1})$, and since $n-1$ is odd, we have

$$|A|\leq n|A'|\leq n(n-1)!/2^{(n-2)/2}=n!/2^{n/2-1}.$$

□

1.4 A construction of a larger independent set

We prove Theorem 1.1.7 in this section. Assume that $n=2^m$. We construct $A\subset S_n$ of size $|A|\geq n!/4^n$, such that A is an independent set in $\mathscr{B}_n$.

Let $T_{i,j}=\{2^{m-i}(j-1)+1,\ldots,2^{m-i}j\}$ for $0\leq i\leq m, 1\leq j\leq 2^i$. Note that $\{T_{i,j}:j\in[2^i]\}$ is a partition of $[n]$ for every i, that $|T_{i,j}|=2^{m-i}$ and that $T_{i,2j-1}\cup T_{i,2j}$ is a partition of $T_{i-1,j}$.

We define a sequence of subsets of S_n. For $1\leq i\leq m$ let $M_i=\binom{2^{m-i+1}}{2^{m-i}}$. For any set R of size $|R|=2^{m-i+1}$ let $\mathrm{ind}_i(R,\cdot)$ be a bijection between subsets of R of size 2^{m-i} and $\mathbb{Z}_{M_i}$. Define $A_0=S_n$ and

$$A_i=\left\{\pi\in A_{i-1}:\sum_{j=1}^{2^{i-1}}\mathrm{ind}_i(\pi(T_{i-1,j}),\pi(T_{i,2j-1}))\equiv 0 \mod M_i\right\}.$$

Since each value mod M_i occurs equally often as a $\mathrm{ind}_i(\pi(T_{i-1,j}),\pi(T_{i,2j-1}))$ for each j, and since these values are independent of one another, $|A_i|=|A_{i-1}|/M_i$. Finally set $A=A_m$. The following claim (applied for $i=m$) shows that A is an independent set in $\mathscr{B}_n$.

Claim 1.4.1. *Let $1\leq i\leq m$. Let $\pi,\pi'\in A_i$ be such that $\tau=\pi(\pi')^{-1}\in\mathscr{C}_n$. Then there exists $j_i\in[2^i]$ such that*

1. *$\tau(T_{i,j_i})=T_{i,j_i}$.*
2. *$\tau(x)=x$ for all $x\in T_{i,j}, j\neq j_i$.*

Proof. We prove the claim by induction on i. The case of $i = 1$ follows from the definition of A_1. By assumption π, π' fix both $T_{1,1}$ and $T_{1,2}$. However, as $\tau = \pi(\pi')^{-1}$ is a cycle, it must be contained in either $T_{1,1}$ or $T_{1,2}$. This implies that $\tau(x) = x$ for all $x \in T_{1,1}$ or all $x \in T_{1,2}$.

Consider next the case of $i > 1$. By induction $\pi(T_{i-1,j}) = \pi'(T_{i-1,j})$ for all $j \in [2^{i-1}]$. Moreover, there exists $j' = j_{i-1}$ such that $\pi(x) = \pi'(x)$ for all $x \in T_{i-1,j}, j \neq j'$. This implies that $\pi(T_{i,j}) = \pi'(T_{i,j})$ for all $j \notin \{2j'-1, 2j'\}$.

Next, the assumption that $\pi, \pi' \in A_i$ guarantees that

$$\sum_{j=1}^{2^{i-1}} \mathrm{ind}_i(\pi(T_{i-1,j}), \pi(T_{i,2j-1})) \equiv \sum_{j=1}^{2^{i-1}} \mathrm{ind}_i(\pi'(T_{i-1,j}), \pi'(T_{i,2j-1})) \equiv 0 \mod M_i.$$

For any $j \neq j'$ we know that $\pi(T_{i-1,j}) = \pi'(T_{i-1,j})$ and $\pi(T_{i,2j-1}) = \pi'(T_{i,2j-1})$, so

$$\mathrm{ind}_i(\pi(T_{i-1,j}), \pi(T_{i,2j-1})) = \mathrm{ind}_i(\pi'(T_{i-1,j}), \pi'(T_{i,2j-1})).$$

Thus we obtain that also $\mathrm{ind}_i(\pi(T_{i-1,j'}), \pi(T_{i,2j'-1})) = \mathrm{ind}_i(\pi'(T_{i-1,j'}), \pi'(T_{i,2j'-1}))$. Moreover, as we also know that $\pi(T_{i-1,j'}) = \pi'(T_{i-1,j'})$ and that $\mathrm{ind}_i(\pi(T_{i-1,j'}), \cdot)$ is a bijection to $\mathbb{Z}_{M_i}$, it must be the case that $\pi(T_{i,2j'-1}) = \pi'(T_{i,2j'-1})$ and hence also $\pi(T_{i,2j'}) = \pi'(T_{i,2j'})$. Thus we conclude that $\pi(T_{i,j}) = \pi'(T_{i,j})$ for all $j \in [2^i]$.

To conclude, as $\tau = \pi(\pi')^{-1}$ is a cycle, it must be contained in either $T_{i,2j'-1}$ or $T_{i,2j'}$. Thus, τ must fix all points in $T_{i,2j'-1}$ or all points in $T_{i,2j'}$. We set $j_i \in \{2j'-1, 2j'\}$ accordingly. □

Finally, we compute the size of A. As $|A_i| = |A_{i-1}|/M_i$ and $M_i = \binom{2^{m-i+1}}{2^{m-i}} \leq 2^{2^{m-i+1}}$ we obtain that

$$|A| \geq \frac{n!}{\prod_{i=1}^{m} 2^{2^i}} \geq \frac{n!}{2^{2^{m+1}}} = \frac{n!}{4^n}.$$

Chapter 1, in full, is a reprint of the material as it appears in IEEE Foundations of Computer Science FOCS 2017 and SIAM Journal on Computing SICOMP. Kane, Daniel; Lovett,

Shachar; Karingula, Sankeerth Rao. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Codes over integers, and the singularity of random matrices with large entries

The prototypical construction of error correcting codes is based on linear codes over finite fields. In this work, we make first steps in the study of codes defined over integers. We focus on Maximally Distance Separable (MDS) codes, and show that MDS codes with linear rate and distance can be realized over the integers with a constant alphabet size. This is in contrast to the situation over finite fields, where a linear size finite field is needed.

At the core is a new result on the singularity probability of random matrices. We show that a for a random $n \times n$ matrix with entries chosen independently from the range $\{-m, \ldots, m\}$, the probability that it is singular is at most m^{-cn} for some absolute constant $c > 0$.

2.1 Introduction

Coding theory is the study of error correction schemes. Codes are widely used in many applications, such as data storage, telecommunications and robust protocols. Algorithms for codes perform arithmetic operations over an underlying alphabet, and hence their computational complexity is constrained by this alphabet size. Thus, understanding the alphabet size needed to support a given code structure is a central question in coding theory. By far, the most common approach to design codes is to use linear codes over finite fields. The main focus of this paper is to investigate the possibility of designing codes over integers. In particular, we study the alphabet

size needed to support basic code structures, and focus on the most basic and well-studied family of codes - Maximally Distance Separable (MDS) codes.

2.1.1 Alphabet size for MDS codes

An MDS code is a code with the best possible tradeoff between the message length, codeword length and minimal distance. Concretely, an (n,k,d)-code is a code with message length k, codeword length n and minimal distance d. The Singleton bound [57] gives that $d \leq n-k+1$. MDS codes are codes achieving this bound, namely (n,k,d)-codes with $d = n-k+1$. If we consider linear codes, then it is well-known that MDS codes are generated by the row span of *MDS matrices*.

Definition 2.1.1 (MDS matrix)**.** *Let $n \geq k$. A $k \times n$ matrix is called an* MDS matrix *if any k columns in it are linearly independent. Equivalently, if any $k \times k$ minor of it is nonsingular.*

Note that MDS matrices can be defined over finite fields or over the integers. If we define them over a finite field $\mathbb{F}_q$, then it is well-known that a linear field size is needed to support MDS matrices. Concretely, if we assume $n \geq k+2$, then it is known that $q \geq \max(k, n-k+1)$ (see for example the introduction of [4] for a proof). In particular, this implies that $q \geq n/2$. Reed-Solomon codes can be constructed over fields of size $q \geq n-1$, which is tight up to a factor of two. The MDS conjecture of Segre [56] speculates that this is indeed the best possible (except for a few special cases), and Ball [4] proved this over prime finite fields. In summary, over finite fields a linear field size $q = \Theta(n)$ is both necessary and sufficient.

We show that over the integers, MDS matrices exist over much smaller alphabet sizes.

Theorem 2.1.2 (MDS matrices over integers)**.** *Let $n \geq k$. There exist $k \times n$ MDS matrices over integers whose entries are in $\{-m, \ldots, m\}$, where $m \leq (cn/k)^c$ for some absolute constant $c > 0$.*

The typical regime in coding theory is that of linear rate and linear distance; namely, where $k = \alpha n$ for some constant $\alpha \in (0,1)$. Note that in this regime Theorem 2.1.2 shows that MDS codes over the integers exist with a *constant* alphabet size, which is in stark contrast

with the case over finite fields. It is easy to see that Theorem 2.1.2 is best possible, up to the unspecified constant c.

Claim 2.1.3. *Let $n \geq k \geq 2$. Let M be a $k \times n$ MDS matrix whose entries are in an alphabet Σ. Then $|\Sigma| \geq \sqrt{n/k}$.*

Proof. Let $P_i = (M_{1,i}, M_{2,i}) \in \Sigma^2$ denote the first two elements in the i-th column of M. If $n > |\Sigma|^2 k$, then there must be k distinct columns $i_1, \ldots, i_k \in [n]$ such that $P_{i_1} = \ldots = P_{i_k}$. But then M cannot be an MDS matrix, as the $k \times k$ minor formed by taking these columns has the first two rows being a scalar multiple of each other, and hence cannot be nonsingular. □

We prove Theorem 2.1.2 by choosing the matrix M randomly, and showing that with high probability it will be an MDS matrix. This is another aspect in which codes over integers seem to be different from codes over finite fields. Constructing MDS matrices over finite fields seems to require algebraic constructions (such as Reed-Solomon codes), unless the field size is exponential in n; whereas over the integers, random matrices work well even for very small entries.

2.1.2 Singularity of random matrices

Our main result is a bound on the singularity probability of random $n \times n$ matrices with uniform integer entries in $\{-m, \ldots, m\}$. Note that the probability that such a matrix is singular is at least $(2m+1)^{-n}$, which is the probability that its first two rows are the same. We show that this bound is tight, up to polynomial factors.

Theorem 2.1.4 (Singularity of random matrices)**.** *Let $n, m \geq 1$. Let M be an $n \times n$ random matrix with random integer entries chosen uniformly in $\{-m, \ldots, m\}$. Then for some absolute constant $c > 0$,*

$$\Pr[M \text{ is singular}] \leq m^{-cn}.$$

Previous works studied this question in two regimes: fixed m and large n, or fixed n and large m. Ours is the first work that can achieve good dependence on both n and m. Before

discussing the connection of our result to previous works, we first show how Theorem 2.1.2 follows directly from Theorem 2.1.4.

Proof of Theorem 2.1.2. Let M be a random $k \times n$ matrix with entries chosen uniformly from $\{-m, \ldots, m\}$. The number of $k \times k$ minors for M is $\binom{n}{k}$, and the probability that each one is singular is at most m^{-ck} by Theorem 2.1.4. Thus

$$\Pr[M \text{ is not MDS}] \leq \binom{n}{k} m^{-ck} \leq \left(\frac{en}{k}\right)^k m^{-ck} = \left(\frac{en}{km^c}\right)^k.$$

In particular, this probability is at most 2^{-k} (say) whenever $m \geq (2en/k)^{1/c}$. □

Previous works in random matrix theory.

Most of the previous works in random matrix theory focused on the regime of fixed m and large n. Specifically, on $n \times n$ random matrices whose entries are sampled independently from distributions with bounded tail . The most studied case is that of random sign matrices, namely $\{-1, 1\}$ entries. Komlós [30] proved that the probability that such a matrix is singular is $o(1)$ as $n \to \infty$, which already is a nontrivial result. It took nearly 30 years until Kahn, Komlós and Szemerédi [23] improved the bound to c^n for some constant $c \in (0, 1)$. A sequence of works [7, 61, 62] improved the value of the constant c, and recently Tikhomirov [67] proved that $c = 1/2 + o(1)$, which is best possible, as the probability that the first two rows of the matrix are equal is 2^{-n}. For more general distributions, Rudelson and Vershynin [49] proved that if the entries of an $n \times n$ matrix are sampled from a sub-Gaussian distribution, then the probability it is singular is at most c^n for some $c \in (0, 1)$. See also [50] for a recent survey on these results.

The other regime, of large m and constant n, was less explored. The only work we are aware of is by Katznelson [25] which gave a bound of the form $c_n m^{-n}$ for some constant c_n depending on n. While having optimal dependence on m for constant n, it has a caveat - it only applies in the regime where m is much larger than n (more precisely, for every fixed n, in the limit of large m).

Random matrices over integers vs over finite fields.

Note that if instead we chose M to be a random $n \times n$ matrix over a finite field $\mathbb{F}_q$, then the probability that M is singular would be about $1/q$, independent of how large n is. This is the main point of difference between random matrices over integers and over finite fields - the singularity probability over integers decreases as the matrix becomes larger, whereas over finite fields it stabilizes.

2.1.3 Discussion

We view Theorem 2.1.2 as a first step towards the study of codes over integers. There are many intriguing questions that arise in coding theory, once we can show that random integer matrices are MDS with high probability. There are also interesting conjectures on the singularity probability of matrices with entries sampled from general distributions. We discuss both briefly below.

Explicit constructions.

A natural question is to give an explicit construction of MDS matrices over integers with small integer values. Concretely, when $k = \alpha n$ for some constant $\alpha \in (0,1)$, to give an explicit construction of a $k \times n$ MDS matrix with a constant alphabet size (namely, independent of n).

Algorithms.

The next natural question, once there are explicit constructions, is to design efficient decoding algorithms for such codes. In particular, it would be intriguing to see if the smaller alphabet size can be utilized to obtain improved runtime (even by logarithmic factors).

General code designs.

In this paper, we focus on MDS codes and the alphabet size needed to realize them over integers. Many other code designs have been studied, many of which have the following common form. Let P be a *pattern matrix* whose entries are $\{0, *\}$. A matrix M (over a finite field, or over

the integers) of the same dimensions as P, is said to *realize* P if it satisfies the following two conditions:

(i) If $P_{i,j} = 0$ then $M_{i,j} = 0$.

(ii) For any maximal minor in P, if it can be realized by some nonsingular matrix, then the corresponding minor in M is nonsingular.

Questions of this form, for various patterns P, have been studied in coding theory. For example, MDS matrices correspond to patterns P which are all $*$. In some applications, condition (ii) is replaced with the following stronger condition (in which case we say that M *strongly* realizes P):

(ii)' For any (maximal or not) minor in P, if it can be realized by some nonsingular matrix, then the corresponding minor in M is nonsingular.

Some areas where these questions arise are: MDS codes with sparse generating matrices (also known as GM-MDS) [14,21,22,34,68]; tree codes, used in coding for interactive communication [8,13,42,53,54]; and maximally recoverable codes, used in coding for distributed storage (there are too many results to list here, we refer to [2] for a recent survey).

Given a pattern P, it is not known when it can be realized (or strongly realized) over small finite fields. Some works show that an exponential field size is needed in some cases [20,24], whereas other works show that in other cases, a polynomial field size is sufficient, using an algebraic construction [34,68]. However, a general understanding is currently lacking. In contrast, we speculate that every pattern (except maybe some pathological cases) can be realized over integers with small entries.

To pose a concrete conjecture, let P_n be the $n \times n$ pattern with $*$s on and below the diagonal, and 0s above the diagonal. Such patterns underlie optimal tree codes. For example for

$n = 4$:

$$P_4 = \begin{pmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & 0 \\ * & * & * & * \end{pmatrix}$$

The best known construction (see [13]) of a matrix M realizing P_n is the binomial coefficients matrix, namely $M_{i,j} = \binom{i}{j}$, whose entries are integers of magnitude about 2^n. We conjecture that this cannot be improved much over finite fields, but can be reduced to $\text{poly}(n)$ over the integers.

Conjecture 2.1.5. *The following holds for the pattern P_n:*

1. *Any matrix M strongly realizing P_n over a finite field $\mathbb{F}_q$ requires exponential field size, namely $q \geq \exp(\Omega(n))$.*

2. *There exist matrices M strongly realizing P_n over the integers, with the nonzero entries in $\{-m, \ldots, m\}$ for $m = poly(n)$. In fact, random matrices of this form should work with high probability.*

Singularity of matrices over general distributions.

As we discussed above, most works on the singularity of random matrices give a bound on the singularity of c^n for some absolute constant $c \in (0,1)$. Theorem 2.1.4 shows that if the entries are uniformly sampled from $\{-m, \ldots, m\}$, we can take $c = 1/\text{poly}(m)$. We speculate that this is an instance of a much more general phenomena - the singularity probability is determined by the anti-concentration of the underlying entries distribution. Given a distribution $\mathscr{D}$ over $\mathbb{R}$, define its max-probability as $\|\mathscr{D}\|_\infty = \max_x \mathscr{D}(x)$. For example, if $\mathscr{D}$ is the uniform distribution over $\{-m, \ldots, m\}$, then $\|\mathscr{D}\|_\infty = 1/(2m+1)$.

Conjecture 2.1.6. *Let $\mathscr{D}$ be a distribution over $\mathbb{R}$ and set $p = \|D\|_\infty$. Let M be a random $n \times n$ matrix with independent entries from $\mathscr{D}$. Then for some absolute constant $c > 0$,*

$$\Pr[M \text{ is singular}] \leq p^{cn}.$$

One can even speculate a more general conjecture, where each entry comes from a different underlying distribution, as long as they all have bounded max-probability.

Chapter Outline.

We prove Theorem 2.1.4 in the remainder of the paper. We start with a high-level overview of our framework in Section 2.2. We compute some preliminary estimates in Section 2.3, define and study incompressible vectors in Section 2.4, define the LCD condition in Section 2.5, where we also prove some properties of it, and bound the LCD of random vectors in Section 2.6. We put all the ingredients together and complete the proof in Section 2.7.

2.2 General approach

We will follow the general approach of Rudelson [48] with several modifications needed to handle the case of large m effectively.

Notation.

It will be convenient to scale the entries to be in $[-1,1]$; we denote by $\mathscr{D}$ the uniform distribution over $\{a/m : a \in \{-m,\dots,m\}\}$. We denote by $\mathscr{D}^n$ the distribution over n-dimensional vectors with independent entries from $\mathscr{D}$, and by $\mathscr{D}^{n\times n}$ the distribution over $n\times n$ matrices with independent entries from $\mathscr{D}$. We denote by S^{n-1} the unit sphere in $\mathbb{R}^n$, namely $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. We will use the c, c', c_0, etc, to denote unspecified positive constants. Note that the same letter (e.g. c) can mean different unspecified constants in different lemmas.

We may assume that n, m are large enough.

We will assume throughout the proof that n, m are large enough; concretely, for any absolute constants n_0, m_0, we may assume that $n \geq n_0, m \geq m_0$, and this would only effect the value of the constant c in Theorem 2.1.4.

To see why, consider first the regime of constant m and large n. The distribution $\mathscr{D}$ is

symmetric and bounded in $[-1,1]$. The results of [49] show that in such a case,

$$\Pr[M \text{ is singular}] \leq c^n$$

for some absolute constant $c \in (0,1)$. This proves Theorem 2.1.4 for any constant m.

The other regime is that n is constant and m is large. While we may appeal to the result of Katznelson [25] in this regime, which gives a sharp bound of $c_n m^{-n}$, there is a much simpler argument that gives a bound of the form $1/m$ in this case, which is good enough to establish Theorem 2.1.4 in this regime. As the determinant of an $n \times n$ matrix is a polynomial of degree n, the Schwartz-Zippel lemma [55, 69] gives

$$\Pr[M \text{ is singular}] = \Pr[\det(M) = 0] \leq \frac{n}{m}.$$

In particular, for constant n and large m, this probability scales like $1/m$, which is consistent with the claimed bound of Theorem 2.1.4 (taking $c < 1/n$).

General approach.

Let $M \sim \mathscr{D}^{n \times n}$, and let $X_1, \ldots, X_n$ denote its rows. If M is singular, then one of the rows belongs to the span of the other rows. By symmetry we have

$$\Pr[M \text{ is singular}] \leq n \cdot \Pr[X_n \in \text{Span}(X_1, \ldots, X_{n-1})].$$

Let X^* be any unit vector orthogonal to $X_1, \ldots, X_{n-1}$ (if there are multiple ones, choose one in some deterministic way). We call it a *random normal vector*. We will shorthand $X = X_n$. Observe that X, X^* are independent. Thus we can bound

$$\Pr[M \text{ is singular}] \leq n \cdot \Pr[\langle X^*, X \rangle = 0].$$

To do so, we will show that unless X^* belongs to a set of "bad" vectors, then the above probability

is at most m^{-cn}, and that the probability that X^* is bad is also at most m^{-cn}.

2.3 Preliminary estimates

We establish some preliminary estimates in this section, which will be needed later in the proof.

Maximal eigenvalues of random matrices.

The first ingredient is bounding the spectral norm of M. In fact, we would need this bound for rectangular matrices. Given an $n \times k$ matrix R we denote its spectral norm as $\|R\| = \max\{\|Rx\|_2 : x \in S^{k-1}\}$. Note that $\|R\| = \|R^T\|$ since $\|R\| = \max_{x \in S^{k-1}, y \in S^{n-1}} y^T R x$.

The following claim is a special case of [48, proposition 4.4], who showed that it holds for any symmetric distribution $\mathscr{D}$ supported in $[-1,1]$.

Claim 2.3.1. *Let $R \sim \mathscr{D}^{n \times k}$ for $n \geq k$. Then for any $\lambda > 0$,*

$$\Pr[\|R\| \geq \lambda \sqrt{n}] \leq 2^{-c\lambda^2 k}.$$

Anti-concentration of projections.

Next, we need anti-concentration results for projections of $\mathscr{D}^n$. To begin with we consider projections of the uniform distribution over the solid cube $[-1,1]^n$.

Claim 2.3.2. *Let $U \sim [-1,1]^n$ be uniformly distributed. Then for every $x \in S^{n-1}$ and $\varepsilon > 0$,*

$$\Pr_u[|\langle U, x \rangle| \leq \varepsilon] \leq c\varepsilon.$$

Proof. The uniform distribution $U \sim [-1,1]^n$ is a log-concave distribution. Let $S = \langle U, x \rangle$ and note that S is a projection of U along the direction x. The Prékopa–Leindler inequality [33, 47] states that projections of log-concave distributions are log-concave, and so S is a log-concave distribution. Carbery and Wright [9, Theorem 8] show that the required anti-concentration bound holds for any log-concave distribution. □

We extend this anti-concentration to the discrete case using a coupling argument. Here and throughout, we denote by $\log(\cdot)$ logarithm in base 2.

Claim 2.3.3. *Let $X \sim \mathscr{D}^n$ and set $\varepsilon_0 = \frac{\sqrt{\log m}}{m}$. Then for every $x \in S^{n-1}$ and $\varepsilon \geq \varepsilon_0$,*

$$\Pr[|\langle X, x\rangle| \leq \varepsilon] \leq c\varepsilon.$$

Proof. We apply a coupling argument between the uniform distribution in $[-1,1]^n$ and $\mathscr{D}^n$. Sample $X \sim \mathscr{D}^n$, $Y \sim [-1,1]^n$ and set $Z = X + Y/2m$. Observe that Z is uniform in the solid cube $[-1-1/2m, 1+1/2m]^n$. Next, fix $\varepsilon > 0$ and observe that $\langle X, x\rangle = \langle Z, x\rangle - \langle Y, x\rangle/2m$. Thus we can bound

$$\Pr[|\langle X, x\rangle| \leq \varepsilon] \leq \Pr[|\langle Z, x\rangle| \leq 2\varepsilon] + \Pr[|\langle Y, x\rangle| \geq 2\varepsilon m].$$

For the first term, Claim 2.3.2 bounds its probability by $c_1\varepsilon$. For the second term, the Chernoff bound bounds its probability for $\varepsilon \geq \varepsilon_0$ by $1/m$. As we have $1/m \leq \varepsilon$, the claim follows. □

Tensorization lemma.

We also need the following "tensorization lemma" (Lemma 6.5 in [48]).

Claim 2.3.4. *Let $Y_1, \ldots, Y_n$ be independent real-valued random variables. Assume for some $K, \varepsilon_0 > 0$ that*

$$\Pr[|Y_i| \leq \varepsilon] \leq K\varepsilon \qquad \text{for all } \varepsilon \geq \varepsilon_0.$$

Then

$$\Pr\left[\sum_{i=1}^n Y_i^2 \leq \varepsilon^2 n\right] \leq (cK\varepsilon)^n \qquad \text{for all } \varepsilon \geq \varepsilon_0.$$

Nets.

A set of unit vectors $\mathscr{N} \subset S^{n-1}$ is called an ε-net, for $\varepsilon > 0$, if it satisfies:

$$\forall x \in S^{n-1}\ \exists y \in \mathscr{N}\ \|x - y\|_2 \leq \varepsilon.$$

The following claim bounds the size of such a net. For a proof see [41, Lemma 2.6].

Claim 2.3.5. *For any $\varepsilon > 0$, there exists a ε-net $\mathcal{N} \subset S^{n-1}$ of size $|\mathcal{N}| \leq (3/\varepsilon)^n$.*

Integer points in ball.

We need a bound on the number of integer vectors in a ball of a given radius. Let $B_n(r) = \{x \in \mathbb{R}^n : \|x\|_2 \leq r\}$ denote the ball of radius r in $\mathbb{R}^n$.

Claim 2.3.6. *The number of integer vectors in $B_n(r)$ is at most $\left(1 + \frac{3r}{\sqrt{n}}\right)^n$.*

2.4 Compressible vectors

The first set of "bad" vectors that we want to rule out are vectors which are close to sparse. A vector $u \in \mathbb{R}^n$ is k-sparse if it has at most k nonzero coordinates.

Definition 2.4.1 (Compressible vectors)**.** *Let $\alpha, \beta \in (0,1)$. A unit vector $x \in S^{n-1}$ is called* (α,β)-compressible *if it can be expressed as $x = u + v$, where u is (αn)-sparse and $\|v\|_2 \leq \beta$. Otherwise, we say that x is* (α,β)-incompressible*.*

We will later choose α, β, but we note here that α will be a small enough absolute constant and β a small polynomial in $1/m$. Concrete values that work are $\alpha = 1/50, \beta = 1/\sqrt{m}$. We will implicitly assume that both n, m are large enough; concretely, at various places we assume that $\alpha n \geq 2$.

The main lemma we prove in this section is the following.

Lemma 2.4.2. *Let $\alpha \in (0, 1/8), \beta \in (\varepsilon_0, 1/2)$ where $\varepsilon_0 = \frac{\sqrt{\log m}}{m}$. Then*

$$\Pr[X^* \text{ is } (\alpha,\beta)\text{-compressible}] \leq (c\beta)^{n/8}.$$

We need a bound on the smallest singular value of a rectangular matrix.

Claim 2.4.3. *Let $R \sim \mathscr{D}^{n \times k}$ for $n \geq k$. Then for every $x \in S^{k-1}$ and $\varepsilon \geq \varepsilon_0$,*

$$\Pr\left[\|Rx\|_2 \leq \varepsilon\sqrt{n}\right] \leq (c\varepsilon)^{n/2}.$$

Proof. Assume $\|Rx\|_2 < \varepsilon\sqrt{n}$. This implies that $|(Rx)_i| \le 2\varepsilon$ for at least $n/2$ coordinates $i \in [n]$. Note that for each fixed i, the value $(Rx)_i$ is distributed as $\langle X, x\rangle$ for some $X \sim \mathscr{D}^k$. Applying Claim 2.3.3 and the union bound over the choice of the $n/2$ coordinates gives

$$\Pr\left[\|Rx\|_2 \le \varepsilon\sqrt{n}\right] \le 2^n (c_1\varepsilon)^{n/2} = (c\varepsilon)^{n/2}.$$

□

Claim 2.4.4. *Let $R \sim \mathscr{D}^{n\times k}$ for $n \ge 8k$. Then for every $\varepsilon \ge \varepsilon_0$,*

$$\Pr\left[\min_{x\in S^{k-1}} \|Rx\|_2 \le \varepsilon\sqrt{n}\right] \le (c\varepsilon)^{n/4}.$$

Proof. We may assume that $\varepsilon \le 1$ by taking $c \ge 1$. Let $\mathscr{N}$ be an (ε^2)-net in S^{k-1} of size $|\mathscr{N}| \le (3/\varepsilon^2)^k$, as given by Claim 2.3.5. Let E_1 denote the event that there exists $y \in \mathscr{N}$ for which $\|Ry\|_2 \le 2\varepsilon\sqrt{n}$. Applying Claim 2.4.3 and a union bound gives

$$\Pr[E_1] \le (3/\varepsilon^2)^k \cdot (c_1\varepsilon)^{n/2} \le (c_2\varepsilon)^{n/4},$$

where we used the assumption $n \ge 8k$. Let E_2 denote the event that $\|R\| \ge \lambda\sqrt{n}$ for $\lambda = \sqrt{\log(1/\varepsilon)}$. Claim 2.3.1 shows that $\Pr[E_2] \le (c_3\varepsilon)^n$. We next show that if E_1, E_2 don't hold then the condition of the claim also doesn't hold, namely that $\|Rx\|_2 > \varepsilon\sqrt{n}$ for all $x \in S^{k-1}$.

Let $x \in S^{k-1}$ be arbitrary and let $y \in \mathscr{N}$ be such that $\|x-y\|_2 \le \varepsilon^2$. Then

$$\|Rx\|_2 \ge \|Ry\|_2 - \|R\| \cdot \|x-y\|_2 \ge (2\varepsilon - \varepsilon^2\lambda)\sqrt{n}.$$

It can be verified that for $\varepsilon \le 1$ we have $\varepsilon\lambda \le 1$, which implies that $\|Rx\|_2 \ge \varepsilon\sqrt{n}$. □

We will now use these two claims to prove Lemma 2.4.2.

Proof of Lemma 2.4.2. Let M' be the $(n-1)\times n$ matrix with rows $X_1, \ldots, X_{n-1}$. Assume that

there exists an (α,β)-compressible vector $x \in S^{n-1}$ in the kernel of M'. By definition, $x = u+v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta$. In particular, $M'(u+v) = 0$ and hence $\|M'u\|_2 = \|M'v\|_2$. In addition, $\|u\|_2 \geq \|x\|_2 - \|v\|_2 \geq 1/2$ since x is a unit vector and $\|v\|_2 \leq \beta \leq 1/2$.

Let E denote the event that $\|M'\| \geq \lambda\sqrt{n}$ for $\lambda = c_1\sqrt{\log(1/\beta)}$, where we choose $c_1 \geq 1$ large enough so that by Claim 2.3.1, $\Pr[E] \leq \beta^n$. Note that as we assume $\beta \leq 1/2$ we have $\lambda \geq c_1 \geq 1$. Assuming that E doesn't hold, we have

$$\|M'u\|_2 = \|M'v\|_2 \leq \|M'\| \cdot \|v\|_2 \leq \lambda\beta\sqrt{n}.$$

In particular, $y = u/\|u\|$ is an (αn)-sparse unit vector that satisfies $\|M'y\|_2 \leq 2\lambda\beta\sqrt{n}$. We next bound the probability that such a vector exists.

Let $\varepsilon = 2\lambda\beta$, and note that $\varepsilon \geq \varepsilon_0$ since $\beta \geq \varepsilon_0$ and $\lambda \geq 1$. There are $\binom{n}{\alpha n}$ options for the support of y. Let $I = \{i : y_i \neq 0\}$ denote a possible support, set $k = |I|$ and let R be an $(n-1)\times k$ matrix with columns $(Y_i : i \in I)$. As $\alpha < 1/8$ we have $n-1 \geq 8k$. Thus we can apply Claim 2.4.4 and obtain that

$$\Pr\left[\neg E \quad \wedge \quad \exists y \in S^{k-1},\ \|Ry\|_2 \leq \varepsilon\sqrt{n}\right] \leq (c_2\varepsilon)^{n/4} = \left(c_3\beta\sqrt{\log 1/\beta}\right)^{n/4}.$$

Note that for $\beta \leq 1$ we have $\beta\log(1/\beta) \leq 1$ and hence the above bound is at most $(c_4\beta)^{n/8}$.

To conclude, we union bound over the choices for I, the number of which is $\binom{n}{\alpha n} \leq 2^n$. Thus we can bound the total probability by $2^n(c_4\beta)^{n/8} = (c_5\beta)^{n/8}$. □

2.5 The LCD condition

Given $x \in \mathbb{R}^n$ let $x = [x] + \{x\}$ be its decomposition into integer and fractional parts, where $[x] \in \mathbb{Z}^n$ and $\{x\} \in [-1/2, 1/2]^n$. The following definition is a variant of the LCD definition of [48].

Definition 2.5.1 (Least common denominator (LCD))**.** *Let $\alpha,\beta \in (0,1)$. Given a unit vector*

$x \in S^{n-1}$, its least common denominator (LCD), denoted $LCD_{\alpha,\beta}(x)$, is the infimum of $D > 0$ such that we can decompose $\{Dx\} = u + v$, where u is (αn)-sparse and $\|v\|_2 \leq \beta \min(D, \sqrt{n})$.

Claim 2.5.2. *Assume $x \in S^{n-1}$ is $(5\alpha, \beta)$-incompressible. Then $LCD_{\alpha,\beta}(x) > \sqrt{\alpha n}$.*

Proof. Let $D = \mathrm{LCD}_{\alpha,\beta}(x)$ and assume towards a contradiction that $D \leq \sqrt{\alpha n}$. Let $y = Dx$. As $\|y\|_2^2 \leq \alpha n$ there are at most $4\alpha n$ coordinates $i \in [n]$ where $|y_i| \geq 1/2$. In all other coordinates $\{y_i\} = y_i$, and hence $y - \{y\}$ is $(4\alpha n)$-sparse. By assumption we can decompose $\{y\} = u + v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta D$. This implies that we can decompose $y = u' + v$ where u' is $(5\alpha n)$-sparse. Thus, we can decompose $x = y/D$ as $x = u'' + v''$, where $u'' = u/D$ is $(5\alpha n)$-sparse and $v'' = v/D$ satisfies $\|v''\|_2 \leq \beta$. This violates the assumption that x is $(5\alpha, \beta)$-incompressible. □

Our main goal in this section is to prove the following lemma, which extends Claim 2.3.3 assuming x has large LCD. To get intuition, we note that the lemma below is useful as long as $\beta \ll \gamma \ll 1$. We will later set $\gamma = \sqrt{\beta}$ to be such a choice. In particular, if we set $\beta = m^{-1/2}$ then we have $\gamma = m^{-1/4}$.

Lemma 2.5.3. *Let $X \sim \mathscr{D}^n$. Let $\alpha, \beta, \gamma \in (0, 1/2)$, $x \in S^{n-1}$ be (α, γ)-incompressible and set $D = LCD_{\alpha,\beta}(x)$. Then for every $\varepsilon \geq 1/2\pi mD$, it holds that*

$$\Pr\left[|\langle X, x\rangle| \leq \varepsilon\right] \leq c\left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right).$$

The proof of Lemma 2.5.3 relies on Esseen's lemma [15].

Lemma 2.5.4 (Esseen's Lemma)**.** *Let Y be a real-valued random variable. Let $\phi_Y(t) = \mathbb{E}[e^{itY}]$ denote the characteristic function of Y. Then for any $\varepsilon > 0$, it holds that*

$$\Pr[|Y| \leq \varepsilon] \leq c\varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} |\phi_Y(t)| dt.$$

Before proving Lemma 2.5.3, we need some auxiliary claims. Fix some $x \in S^{n-1}$, let $X \sim \mathscr{D}^n$ and let $Y = \langle X, x \rangle$. In order to apply Lemma 2.5.4, we need to compute the characteristic function of Y.

Claim 2.5.5. *Let $X \sim \mathscr{D}^n$, $x \in S^{n-1}$ and set $Y = \langle X, x \rangle$. For $t \in \mathbb{R}$ it holds that*

$$|\phi_Y(t)| = \prod_{k=1}^{n} F\left(\frac{x_k t}{2\pi m}\right)$$

where $F : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$F(y) = \left| \frac{\sin\left((2m+1)\pi y\right)}{(2m+1)\sin(\pi y)} \right|.$$

Proof. We have $Y = \sum x_i \xi_i$ where $\xi_1, \ldots, \xi_n \sim \mathscr{D}$ are independent. Hence

$$\phi_Y(t) = \prod_{k=1}^{n} \mathbb{E}[e^{i x_k \xi_k t}].$$

Next we compute

$$\mathbb{E}[e^{i x_k \xi_k t}] = \frac{1}{2m+1} \sum_{\ell=-m}^{m} e^{i x_k (\ell/m) t} = \frac{1}{2m+1} \cdot \frac{\sin(\frac{2m+1}{2m} x_k t)}{\sin(\frac{1}{2m} x_k t)}.$$

Hence

$$\left| \mathbb{E}[e^{i t x_k \xi_k}] \right| = F\left(\frac{x_k t}{2\pi m}\right).$$

□

The next claim proves some basic properties of the function F.

Claim 2.5.6. *The function F satisfies the following properties:*

1. *F is symmetric: $F(y) = F(-y)$ for all $y \in \mathbb{R}$.*
2. *F is invariant to shifts by integers: $F(y) = F(\{y\})$ for $y \in \mathbb{R}$.*

3. *F is bounded: $F(y) \in [0,1]$ for all $y \in \mathbb{R}$.*

4. *$F(y) \le G(my)$ for $y \in [0,1/2]$, where $G : \mathbb{R}_+ \to [0,1]$ is defined as follows:*

$$G(y) = \begin{cases} e^{-\eta y^2} & \text{if } y \in [0,1] \\ \frac{e^{-\eta}}{y} & \text{if } y \ge 1 \end{cases}$$

Here, $\eta > 0$ is an absolute constant. Note that G is decreasing.

Proof. The first three claims follow immediately from the definition of F in Claim 2.5.5. In order to prove the last claim, we will prove that $F(y) \le \frac{c_1}{my}$ for $y \in [1/m, 1/2]$ for some $c_1 \in (0,1)$; and that $F(y) \le \exp(-c_2(my)^2)$ for $y \in [0,1/m]$ for some $c_2 > 0$. The claim then follows by taking $\eta = \min(\ln(1/c_1), c_2)$.

First, note that $F(y) \le \frac{1}{(2m+1)|\sin(\pi y)|}$. Using Taylor expansion at 0, we get for $y \in [0,1/2]$ that

$$\sin(\pi y) \ge \pi y - \frac{\pi^3 y^3}{6} \ge \frac{\pi y}{2}.$$

In particular, $F(y) \le \frac{1}{\pi my}$, which gives the desired bound for $c_1 = 1/\pi$.

Next, note that $F(y) = \frac{1}{2m+1}|\sin((2m+1)\pi y) \cdot \csc(\pi y)|$. The Laurent series of $\csc(x)$ at $x \ne 0$ is $\csc(x) = \frac{1}{x} + \frac{x}{6} + \frac{7x^3}{360} + \frac{31x^5}{15120} + \Theta(x^7)$ and the Taylor series for $\sin(x)$ is $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \Theta(x^7)$. So for $y \in [0,1/m]$ we have $F(y) \le 1 - c_2(my)^2 \le \exp(-c_2(my)^2)$. □

We also need the following claim, which shows that incompressible vectors retain a large fraction of their norm when restricted to small coordinates. We use the following notation: given $x \in \mathbb{R}^n$ and a set of coordinates $J \subset [n]$, we denote by $x|_J \in \mathbb{R}^J$ the restriction of x to coordinates in J.

Claim 2.5.7. *Let $x \in S^{n-1}$ be (α,γ)-incompressible. Let $J = \left\{ i : x_i \le \frac{1}{\sqrt{\alpha n - 1}} \right\}$. Then*

$$\|x|_J\|_2^2 \ge \|x|_J\|_\infty^2 + \gamma^2.$$

Proof. Let $J^c = [n] \setminus J$. Since x is a unit vector, we have $|J^c| \leq \alpha n - 1$. Let $j \in J$ be such that $|x_j|$ is maximal and take $K = J \setminus \{j\}$. Then $|K^c| \leq \alpha n$, and since we assume that x is (α, γ)-incompressible, we have $\|x|_K\|_2 \geq \gamma$. This completes the proof, since

$$\|x|_J\|_2^2 - \|x|_J\|_\infty^2 = \|x|_J\|_2^2 - x_j^2 = \|x|_K\|_2^2 \geq \gamma^2.$$

□

We would need the following lemma in the computations later on.

Lemma 2.5.8. *Let $\gamma, \delta > 0$. Let $x \in \mathbb{R}^n$ be a vector such that $\|x\|_\infty \leq \delta$ and $\|x\|_2^2 \geq \|x\|_\infty^2 + \gamma^2$. Let $T = \pi m/\delta$. Then*

$$I = \int_0^T \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt \leq \frac{c}{\gamma}.$$

Proof. To simplify the proof, we may assume by Claim 2.5.6(1) that $x_i \geq 0$ for all i. Reorder the coordinates of x so that $x_1 \geq x_2 \geq \ldots \geq x_n \geq 0$. Observe that for $x_i \in [0, T]$ we have $\frac{x_i t}{2\pi m} \in [0, 1/2]$ and hence we can apply Claim 2.5.6(4) and bound each term by $F\left(\frac{x_i t}{2\pi m}\right) \leq G\left(\frac{x_i t}{2\pi}\right)$. Thus

$$I \leq \int_0^T \prod_{i=1}^n G\left(\frac{x_i t}{2\pi}\right) dt = 2\pi \int_0^{T/2\pi} \prod_{i=1}^n G(x_i t) dt \leq 2\pi \int_0^\infty \prod_{i=1}^n G(x_i t) dt.$$

We bound this last integral. Let $t_i = 1/x_i$ so that $t_1 \leq t_2 \leq \ldots \leq t_n$. For simplicity of notation set $t_0 = 0, t_{n+1} = \infty$. We break the computation of the integral to intervals $[t_k, t_{k+1})$ for $k = 0, \ldots, n$, and denote by I_k the integral in each interval:

$$I_k = \int_{t_k}^{t_{k+1}} \prod_{i=1}^n G(x_i t) dt = \int_{t_k}^{t_{k+1}} \prod_{i=1}^k \frac{e^{-\eta}}{x_i t} \cdot \prod_{i=k+1}^n e^{-\eta t^2 x_i^2} dt = e^{-\eta k} \int_{t_k}^{t_{k+1}} \frac{e^{-\eta t^2 \sum_{i=k+1}^n x_i^2}}{t^k \prod_{i=1}^k x_i} dt.$$

Fix k and consider first the case that $\sum_{i=k+1}^n x_i^2 \geq \gamma^2/2$. In this case, using the fact that $x_i t \geq 1$ for $i \in [k]$ and $t \in [t_k, t_{k+1}]$, we can bound I_k by

$$I_k \leq e^{-\eta k} \int_{t_k}^{t_{k+1}} e^{-\eta \gamma^2 t^2/2} dt \leq e^{-\eta k} \int_0^\infty e^{-\eta \gamma^2 t^2/2} dt \leq \frac{c_1 e^{-\eta k}}{\gamma}.$$

Next, consider the case that $\sum_{i=k+1}^{n} x_i^2 < \gamma^2/2$, which means that $\sum_{i=1}^{k} x_i^2 > \|x\|_2^2 - \gamma^2/2 \geq \|x\|_\infty^2 + \gamma^2/2$. Observe that this is impossible for $k = 0$ or $k = 1$, and hence we may assume $k \geq 2$. In this case we bound

$$I_k \leq e^{-\eta k} \int_{t_k}^{t_{k+1}} \frac{1}{t^k \prod_{i=1}^{k} x_i} dt \leq e^{-\eta k} \int_{t_k}^{\infty} \frac{1}{t^k \prod_{i=1}^{k} x_i} dt = \frac{e^{-\eta k} x_k^{k-1}}{(k-1)\prod_{i=1}^{k} x_i} \leq \frac{e^{-\eta k}}{(k-1)x_1}.$$

By our assumption, $\sum_{i=1}^{k} x_i^2 \geq \gamma^2/2$ and hence $x_1^2 \geq \gamma^2/2k$. Thus we can bound

$$I_k \leq \frac{e^{-\eta k}}{(k-1)\gamma/\sqrt{2k}} \leq \frac{c_2 e^{-\eta k}}{\gamma}.$$

Overall, we bounded the integral by

$$I \leq 2\pi \sum_{k=0}^{n} I_k \leq 2\pi \max(c_1, c_2) \sum_{k=0}^{n} \frac{e^{-\eta k}}{\gamma} \leq \frac{c}{\gamma},$$

where we used the fact that $c_1, c_2, \eta > 0$ are all absolute constants. □

Now we have all the ingredients to complete proof of Lemma 2.5.3.

Proof of Lemma 2.5.3. Let $Y = \langle X, x \rangle$. Lemma 2.5.4 and Claim 2.5.5 give the bound

$$\Pr[|Y| \leq \varepsilon] \leq c_1 \varepsilon I,$$

where I is the following integral:

$$I = \int_0^{1/\varepsilon} \prod_{i=1}^{n} F\left(\frac{x_i t}{2\pi m}\right) dt.$$

Let $T = \pi m \sqrt{\alpha n - 1}$. We will bound the integral in the regime $[0, T]$ and $[T, 1/\varepsilon]$, and denote the corresponding integrals by I_1, I_2.

Consider first the integral I_1 in $[0, T]$. Let $\delta = 1/\sqrt{\alpha n - 1}$ and take $J = \{i : x_i \leq \delta\}$.

Observe that by Claim 2.5.6(3), we can bound $F\left(\frac{x_i t}{2\pi m}\right) \leq 1$ for $i \notin J$. Thus

$$I_1 = \int_0^T \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt \leq \int_0^T \prod_{i \in J} F\left(\frac{x_i t}{2\pi m}\right) dt.$$

Next, as we assume that x is (α, γ)-incompressible, Claim 2.5.7 gives that $\|x|_J\|_2^2 \geq \|x|_J\|_\infty^2 + \gamma^2$. Applying Lemma 2.5.8 to $x|_J$, we bound the first integral by

$$I_1 \leq \frac{c_2}{\gamma}.$$

Next, consider the second integral I_2 in $[T, 1/\varepsilon]$. We will apply the LCD assumption to uniformly bound the integrand in this range. Given $t \in [T, 1/\varepsilon]$, let $y(t) = \left\{\frac{xt}{2\pi m}\right\} \in [-1/2, 1/2]^n$, $\beta(t) = \beta \min(t/\sqrt{n}, 1)$ and $J(t) = \{i \in [n] : |y(t)_i| \geq \beta(t)\}$. As $t \leq 1/\varepsilon \leq 2\pi m D$, we have that $\frac{t}{2\pi m} \leq D = \mathrm{LCD}_{\alpha,\beta}(x)$, and hence $|J(t)| \geq \alpha n$. Applying Claim 2.5.6, we bound the integrand by

$$\prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) = \prod_{i=1}^n F(y_i) \leq \prod_{i \in J(t)} F(y_i) \leq \prod_{i \in J(t)} G(m y_i) \leq \prod_{i \in J(t)} G(m\beta(t)) \leq G(m\beta(t))^{\alpha n}.$$

Following up on this, we have

$$\beta(t) \geq \beta(T) = \beta \frac{\sqrt{\alpha n - 1}}{\sqrt{n}} \geq \beta \sqrt{\alpha/2} \geq \alpha\beta,$$

where we used the assumptions that $\alpha n \geq 2$ and $\alpha \leq 1/2$. We may assume that $\alpha\beta m \geq 1$, otherwise the conclusion of the lemma is trivial. In that case we have by Claim 2.5.6(4) that

$$G(m\beta(t)) \leq G(\alpha\beta m) \leq \frac{1}{\alpha\beta m}.$$

Thus we can bound the integral I_2 by

$$I_2 = \int_T^{1/\varepsilon} \prod_{i=1}^{n} F\left(\frac{x_i t}{2\pi m}\right) dt \leq \frac{1/\varepsilon}{(\alpha\beta m)^{\alpha n}}.$$

Overall, we get

$$\Pr[|Y| \leq \varepsilon] \leq c_1 \varepsilon I = c_1 \varepsilon (I_1 + I_2) \leq \frac{c_1 c_2 \varepsilon}{\gamma} + \frac{c_1}{(\alpha\beta m)^{\alpha n}}.$$

□

2.6 Bounding the LCD

Our main goal in this section is to prove that a random normal vector X^* has large LCD with high probability. Let M' denote the $(n-1) \times n$ matrix with rows $X_1, \ldots, X_{n-1}$. Let $D_0 = \sqrt{\alpha n}$ and $D_1 = \beta(\alpha\beta m)^{\alpha n}$ in this section.

Lemma 2.6.1. *Let $\alpha \in (0, 1/40)$, $\beta \in (0, 1/2)$ and take $1 \leq D \leq D_1$. Then*

$$\Pr[LCD_{\alpha,\beta}(X^*) \leq D] \leq D^2 (1/\alpha c)^n \beta^{cn}$$

for some absolute constant $c \in (0,1)$.

We set $\gamma = \sqrt{\beta}$ throughout the section. We first condition on a number of bad events not holding. Define:

$$E_1 = \left[\|M\| \geq \sqrt{n \log(1/\beta)}\right]$$

$$E_2 = [X^* \text{ is } (5\alpha, \beta)\text{-compressible}]$$

$$E_3 = [X^* \text{ is } (\alpha, \gamma)\text{-compressible}]$$

Applying Claim 2.3.1 for E_1, and Lemma 2.4.2 for E_2, E_3, we get that

$$\Pr[E_1 \text{ or } E_2 \text{ or } E_3] \leq \beta^{cn}.$$

Thus, we will assume in this section that non of E_1, E_2, E_3 hold. Assuming $\neg E_2$, Claim 2.5.2 yield that $\mathrm{LCD}_{\alpha,\beta}(X^*) \geq D_0$. For $D \geq D_0$ define

$$\mathscr{S}_D = \left\{ x \in S^{n-1} : \mathrm{LCD}_{\alpha,\beta}(x) \in [D, 2D] \text{ and } x \text{ is } (\alpha,\gamma)\text{-incompressible} \right\}.$$

The following is an analog of Lemma 7.2 in [48].

Claim 2.6.2. *Let $D \geq D_0$ and set $\nu = 6\beta\sqrt{n}/D$. There exists a ν-net $\mathscr{N}_D \subset \mathscr{S}_D$ of size*

$$|\mathscr{N}_D| \leq (D/\beta) \left(\frac{26D}{\sqrt{\alpha n}} \right)^n (1/\beta)^{\alpha n}.$$

Namely, for each $x \in \mathscr{S}_D$ there exists $y \in \mathscr{N}_D$ that satisfies $\|x - y\|_2 \leq \nu$.

Proof. Let $x \in \mathscr{S}_D$ and shorthand $D(x) = \mathrm{LCD}_{\alpha,\beta}(x)$. By definition, we can decompose $\{D(x)x\} = u + v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta \min(D, \sqrt{n}) \leq \beta\sqrt{n}$.

Let W denote the set of (αn)-sparse vectors $w \in [-1/2, 1/2]^n$ such that each w_i is an integer multiple of β. Then $|W| \leq \binom{n}{\alpha n}(1/\beta)^{\alpha n}$, and there exists $w \in W$ such that $\|u - w\|_\infty \leq \beta$, which implies $\|u - w\|_2 \leq \beta\sqrt{n}$. This implies that

$$\|\{D(x)x\} - w\|_2 \leq 2\beta\sqrt{n}.$$

Next, consider $[D(x)x] \in \mathbb{Z}^n$. As $|[a]| \leq 2|a|$ for all $a \in \mathbb{Z}$, we have $\|[D(x)x]\|_2 \leq 2D(x)\|x\|_2 \leq 4D$. Let $Z = \{z \in \mathbb{Z}^n : \|z\|_2 \leq 4D\}$. Then $[D(x)x] \in Z$, and Claim 2.3.6 bounds $|Z| \leq \left(1 + \frac{12D}{\sqrt{n}}\right)^n$. So there is $z \in Z$ such that

$$\|D(x)x - z - w\|_2 \leq 2\beta\sqrt{n}.$$

Next, let R be set of integer multiples of β in the range $[D, 2D]$, so that $|R| \leq D/\beta$ and there exists $r \in R$ with $|D(x) - r| \leq \beta$. As $\|x\|_2 = 1$ we have

$$\|rx - z - w\|_2 \leq 2\beta\sqrt{n} + \beta \leq 3\beta\sqrt{n}.$$

Finally, define the set

$$Y = \{(z+w)/r : z \in Z, w \in W, r \in R\}.$$

Then there exists $y \in Y$ such that

$$\|x - y\|_2 \leq 3\beta\sqrt{n}/D = \nu/2.$$

Take a maximal set $\mathscr{N}_D \subset \mathscr{S}_D$ which is ν-separated. That is, for any $x', x'' \in \mathscr{N}_D$ we have $\|x' - x''\|_2 > \nu$. Note that by maximality, $\mathscr{N}_D$ is a ν-net in $\mathscr{S}_D$. Next, note that $|\mathscr{N}_D| \leq |Y|$, as any point $x \in \mathscr{N}_D$ must be $(\nu/2)$-close to a distinct point in Y. To conclude, we need to bound $|Y|$. We have

$$|Y| \leq |W||Z||R| \leq \binom{n}{\alpha n}(1/\beta)^{\alpha n} \cdot \left(1 + \frac{12D}{\sqrt{n}}\right)^n \cdot (D/\beta).$$

As $D \geq D_0 = \sqrt{\alpha n}$ we can simplify $1 + \frac{12D}{\sqrt{n}} \leq \frac{13D}{\sqrt{\alpha n}}$. We can trivially bound $\binom{n}{\alpha n} \leq 2^n$. Hence

$$|\mathscr{N}_D| \leq |Y| \leq (D/\beta)\left(\frac{26D}{\sqrt{\alpha n}}\right)^n (1/\beta)^{\alpha n}.$$

□

Claim 2.6.3. *For any $D \in [D_0, D_1]$ we have*

$$\Pr[X^* \in \mathscr{S}_D \text{ and } \neg E_1] \leq D^2 (c/\alpha)^n \beta^{n/8}.$$

Proof. First, note that we may assume $\beta \leq \beta_0$ for any absolute constant $\beta_0 \in (0,1)$, by choosing

the constant $c > 0$ large enough to compensate for that (namely, taking $c \geq 1/\beta_0$). In particular, setting $\beta_0 = 2^{-20}$ works.

If $X^* \in \mathscr{S}_D$ then there exists $y \in \mathscr{N}_D$ such that $\|X^* - y\|_2 \leq \nu$ for $\nu = 6\beta\sqrt{n}/D$. By definition of X^* we have $M'X^* = 0$, and as we assume that $\neg E_1$ hold, we have

$$\|M'y\|_2 \leq \|M'\|\|X^* - y\|_2 \leq \nu\sqrt{n\log(1/\beta)}.$$

Set $\beta_1 = 6\beta\sqrt{\log(1/\beta)}$. The assumption $\beta \leq \beta_0$ implies that $\beta_1 \leq \beta^{3/4}$. Set $\delta = \beta^{3/4}\sqrt{n}/D$. We will bound the probability that there exists $y \in \mathscr{N}_D$ such that $\|M'y\|_2 \leq \delta\sqrt{n}$.

Fix $y \in \mathscr{N}_D$, let $X \sim \mathscr{D}^n$, and define $p(\varepsilon) = \Pr[|\langle X, y\rangle|] \leq \varepsilon$. As $y \in \mathscr{N}_D \subset \mathscr{S}_D$ we have that y is (α, γ)-incompressible, and hence we can apply Lemma 2.5.3, which gives

$$p(\varepsilon) \leq c_1\left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right) \qquad \text{for all } \varepsilon \geq 1/2\pi mD.$$

Next, we restrict attention to only $\varepsilon \geq \delta$, and note that in this regime the first term is dominant (since $D \leq D_1$ we have $\delta \geq \beta^{3/4}\sqrt{n}/D_1 \geq 1/(\alpha\beta m)^{\alpha n}$). We can then simplify the bound as

$$p(\varepsilon) \leq \frac{c_2\varepsilon}{\gamma} \qquad \text{for all } \varepsilon \geq \delta.$$

Applying Claim 2.3.4, and recalling that we set $\gamma = \sqrt{\beta}$, gives

$$\Pr\left[\|M'y\|_2 \leq \delta\sqrt{n}\right] \leq \left(\frac{c_3\delta}{\gamma}\right)^{n-1} = \left(\frac{c_4\beta^{1/4}\sqrt{n}}{D}\right)^{n-1}.$$

Union bounding over all $y \in \mathscr{N}_D$, using Claim 2.6.2 to bound its size, gives

$$\begin{aligned}\Pr[\exists y \in \mathscr{N}_D, \|M'y\|_2 \leq \delta\sqrt{n}] &\leq (D/\beta)\left(\frac{26D}{\sqrt{\alpha n}}\right)^n (1/\beta)^{\alpha n} \cdot \left(\frac{c_4\beta^{1/4}\sqrt{n}}{D}\right)^{n-1} \\ &\leq D^2\left(c_5/\sqrt{\alpha}\right)^n \beta^{n/4-\alpha n-2}.\end{aligned}$$

Our assumption $\alpha < 1/40$ and the implicit assumption $\alpha n \geq 2$ imply that $\alpha n + 2 \leq n/8$, which simplifies the above bound to the claimed bound. □

We are now in place to prove Lemma 2.6.1.

Proof of Lemma 2.6.1. We may assume that non of E_1, E_2, E_3 hold, as the probability that any of them hold is at most $\beta^{c_1 n}$ for some absolute constant $c_1 \in (0,1)$. This in particular implies that $\mathrm{LCD}_{\alpha,\beta}(X^*) \geq D_0$. Fix $D \in [D_0, D_1]$. As $D \leq D_1$ we can applying Claim 2.6.3 to $D_i = 2^i D_0$ as long as $D_i \leq D/2$. Summing the results we get

$$\Pr\left[\mathrm{LCD}_{\alpha,\beta}(X^*) \leq D \text{ and } \neg E_1, \neg E_2, \neg E_3\right] \leq (2D)^2 (c_2/\alpha)^n \beta^{n/8}.$$

Thus overall we have

$$\Pr\left[\mathrm{LCD}_{\alpha,\beta}(X^*) \leq D\right] \leq \beta^{c_1 n} + (2D)^2 (c_2/\alpha)^n \beta^{n/8}.$$

The lemma follows by taking $c \in (0,1)$ small enough. □

2.7 Completing the proof

We now prove Theorem 2.1.4.

Proof of Theorem 2.1.4. Fix $\alpha = 1/50, \beta = 1/\sqrt{m}$ and assume $m \geq m_0$ for a large enough constant m_0 to be determined soon. Let D to be determined soon. Lemma 2.6.1 gives

$$\Pr[\mathrm{LCD}_{\alpha,\beta}(X^*) \leq D] \leq D^2 (1/\alpha c_1)^n \beta^{c_1 n}.$$

As α is constant, and using the choice $\beta = 1/\sqrt{m}$, we can simplify the bound as follows. For a small enough constant $c \in (0,1)$, setting $D = m^{cn}$ and $c_2 = 1/\alpha c_1$, we have

$$\Pr[\mathrm{LCD}_{\alpha,\beta}(X^*) \leq m^{cn}] \leq m^{2cn} c_2^n m^{-(c_1/2)n} \leq c_2^n m^{-(c_1/2 - 2c)n} \leq c_2^n m^{-cn}.$$

Assuming $m \geq m_0$ for a large enough constant m_0, we can simplify this bound further as

$$\Pr[\mathrm{LCD}_{\alpha,\beta}(X^*) \leq m^{cn}] \leq m^{-(c/2)n}.$$

Next, assume $D = \mathrm{LCD}_{\alpha,\beta}(X^*) \geq m^{cn}$. In this case, Lemma 2.5.3 for $\varepsilon = 1/2\pi mD$ gives that

$$\Pr[\langle X^*, X\rangle = 0] \leq \Pr[|\langle X^*, X\rangle| \leq \varepsilon] \leq c_3 \left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right) \leq m^{-c'n}$$

for some $c' \in (0,1)$. Overall we obtain the desired bound. □

Chapter 2, in full, is currently under review for publication of the material. Karingula, Sankeerth Rao; Lovett, Shachar. The dissertation author was the primary investigator and author of this material.

Chapter 3

Combinatorial designs

A new probabilistic technique for establishing the existence of certain regular combinatorial structures has been introduced by Kuperberg, Lovett, and Peled (STOC 2012). Using this technique, it can be shown that under certain conditions, a randomly chosen structure has the required properties of a t-(n,k,λ) combinatorial design with tiny, yet positive, probability.

Herein, we strengthen both the method and the result of Kuperberg, Lovett, and Peled as follows. We modify the random choice and the analysis to show that, under the same conditions, not only does a t-(n,k,λ) design exist but, in fact, with positive probability there exists a *large set* of such designs — that is, a partition of the set of k-subsets of $[n]$ into t-(n,k,λ) designs. Specifically, using the probabilistic approach derived herein, we prove that for all sufficiently large n, large sets of t-(n,k,λ) designs exist whenever $k > 9t$ and the necessary divisibility conditions are satisfied. This resolves the existence conjecture for large sets of designs for all $k > 9t$.

3.1 Introduction

Let $[n] = \{1,2,\ldots,n\}$. A k-set is a subset of $[n]$ of size k. A t-(n,k,λ) *combinatorial design* is a collection $\mathscr{D}$ of distinct k-sets of $[n]$, called *blocks*, such that every t-set of $[n]$ is contained in exactly λ blocks. A *large set of designs* of size l, denoted $\mathbf{LS}(l;t,k,n)$, is a set of l disjoint t-(n,k,λ) designs $\mathscr{D}_1,\mathscr{D}_2,\ldots,\mathscr{D}_l$ such that $\mathscr{D}_1 \cup \mathscr{D}_2 \cup \cdots \cup \mathscr{D}_l$ is the set of all k-sets of

$[n]$. That is, $\mathbf{LS}(l;t,k,n)$ is a partition of the set of k-sets of $[n]$ into t-(n,k,λ) designs, where necessarily $\lambda = \binom{n-t}{k-t}/l$.

The existence problem for large sets of designs can be phrased as follows: for which values of l,t,k,n do $\mathbf{LS}(l;t,k,n)$ large sets exist? The existence conjecture for large sets, formulated for example in [66, Con-jecture 1.4], asserts that for every fixed l,t,k with $k \geq t+1$, a large set $\mathbf{LS}(l;t,k,n)$ exists for all sufficiently large n that satisfy the obvious divisibility constraints (see Section 3.1.2). However, according to [66, p. 564] as well as more recent surveys, "not many results about $\mathbf{LS}(l;t,k,n)$ with $k > t+1$ are known." One of our main results herein is a proof of the foregoing existence conjecture for all $k > 9t$.

3.1.1 Large sets of designs

Combinatorial design theory can be traced back to the work of Euler, who introduced the famous "36 officers problem" in 1782. Euler's ideas were further developed in the mid-19th century by Cayley, Kirkman, and Steiner. In particular, the existence problem for large sets of designs was first considered in 1850 by Cayley [10], who found two disjoint 2-$(7,3,1)$ designs and showed that no more exist. The first nontrivial large set, namely $\mathbf{LS}(7;2,3,9)$, was constructed by Kirkman [29] in the same year. Following these results, the existence problem for large sets of type $\mathbf{LS}(n-2;2,3,n)$ — that is, large sets of Steiner triple systems — attracted considerable research attention. Nevertheless, this problem remained open until the 1980s, when it was settled by Lu [35, 36] and Teirlinck [65]. Specifically, it is shown in [35, 36, 65] that $\mathbf{LS}(n-2;2,3,n)$ exist for all $n \geq 9$ with $n \equiv 1,3 \pmod 6$. In 1987, came the celebrated work of Teirlinck [63], who proved that nontrivial t-(n,k,λ) designs exist for all values of t. In fact, Teirlinck's proof of this theorem in [63] proceeds by constructing for all $t \geq 1$, a large set $\mathbf{LS}(l;t,t+1,n)$, where $l = (n-t)/(t+1)!^{(2t+1)}$. His results in [63, 64] further imply that for all fixed t,k with $k \geq t+1$, nontrivial large sets $\mathbf{LS}(l;t,k,n)$ exist for infinitely many values of n. However, as mentioned earlier, it is unknown whether such large sets exist for *all sufficiently large values of n* that satisfy the necessary divisibility constraints. For much more on the history

of the problem and the current state of knowledge, see the surveys [27, 28, 66] and references therein.

There are numerous applications of large sets of designs in discrete mathematics and computer science. For example, large sets of Steiner systems were used to construct perfect secret-sharing schemes by Stinson and Vanstone [58], and others [16, 52]. An application of general large sets of designs to threshold secret-sharing schemes was proposed by Chee [11]. As another example, Chee and Ling [12] showed how large sets can be used to construct infinite families of optimal constant weight codes. As yet another example, large sets of 1-designs (also known as one-factorizations) have been used extensively in various kinds of scheduling problems — see [40, pp. 51–53] and references therein.

3.1.2 Divisibility constraints and our existence theorem

Consider a t-(n,k,λ) design with N blocks. It is very easy to see that every such design must satisfy certain natural divisibility constraints. For instance, every k-set of $[n]$ contains exactly $\binom{k}{t}$ many t-sets, and since every t-set is covered exactly λ times by the N blocks, we have $N\binom{k}{t} = \lambda\binom{n}{t}$. In particular, this implies that $\binom{k}{t}$ should divide $\lambda\binom{n}{t}$. Now let us fix a positive integer $s \leq t-1$ and restrict our attention only to those N' blocks that contain a specific s-set of $[n]$. Since the fixed s-set can be extended to a t-set in $\binom{n-s}{t-s}$ ways and each of these t-sets is covered λ times by the N' blocks, a similar argument yields $N'\binom{k-s}{t-s} = \lambda\binom{n-s}{t-s}$. Thus $\binom{k-s}{t-s}$ should divide $\lambda\binom{n-s}{t-s}$. Altogether, this simple counting argument produces t divisibility constraints:

$$\binom{k-s}{t-s} \;\Bigg|\; \lambda\binom{n-s}{t-s} \qquad \text{for all } \; s = 0, 1 \ldots, t-1. \tag{3.1}$$

The above leads to the following natural question. Are these t divisibility conditions also *sufficient* for the existence of t-(n,k,λ) designs, at least when n is large enough? This is one of the central questions in combinatorial design theory. In a remarkable achievement, Keevash [26] was able to answer this question positively, thereby settling the *existence conjecture*

for combinatorial designs. Specifically, Keevash proved that for any $k > t \geq 1$ and $\lambda \geq 1$, there is a sufficiently large $n_0 = n_0(t,k,\lambda)$ such that the following holds: for all $n \geq n_0$ such that n,t,k,λ satisfy the divisibility conditions in (3.1), there exists a t-(n,k,λ) design.

Let us now consider the divisibility conditions for large sets. A large set $\mathbf{LS}(l;t,k,n)$ is a partition of all k-sets of $[n]$ into t-(n,k,λ) designs. Clearly, each of these designs consists of $N = \binom{n}{k}/l = \lambda\binom{n}{t}/\binom{k}{t}$ blocks. This can be used to specify λ in terms of n,t,k,l as follows:

$$\lambda \;=\; \frac{\binom{n}{k}\binom{k}{t}}{l\binom{n}{t}} \;=\; \frac{1}{l}\binom{n-t}{k-t} \tag{3.2}$$

With this, the divisibility constraints (3.1) for the l component designs of a large set $\mathbf{LS}(l;t,k,n)$ can be re-written in terms of n,t,k,l. Altogether, we conclude that the parameters of a large set $\mathbf{LS}(l;t,k,n)$ must satisfy the following $t+1$ divisibility constraints:

$$l\binom{k-s}{t-s} \;\Bigg|\; \binom{n-t}{k-t}\binom{n-s}{t-s} \qquad \text{for all } \; s = 0,1\ldots,t. \tag{3.3}$$

Note that the constraint for $s = t$ simply refers to the condition that l must divide $\binom{n-t}{k-t}$, which is clearly necessary in view of (3.2). Once again, this leads to the following natural question. Are these $t+1$ divisibility conditions also *sufficient* for the existence of $\mathbf{LS}(l;t,k,n)$ large sets, at least when n is large enough?

One of our main results in this paper is a positive answer to this question for all $k > 9t$, which settles the existence conjecture for large sets for such values of k. We formulate this result as the following theorem.

Theorem 3.1.1. *For any $t \geq 1, k > 9t$ and $l \geq 1$, there is an $n_0 = n_0(t,k,l)$ such that the following holds: for all $n \geq n_0$ such that n,t,k,l satisfy the divisibility conditions in* (3.3)*, there exists an $\mathbf{LS}(l;t,k,n)$ large set.*

In fact, Theorem 3.1.1 follows as a special case of a more general statement — namely,

Theorem 3.1.9 of Section 3.1.4. Theorem 3.1.9 itself follows by extending and strengthening the probabilistic argument of Kuperberg, Lovett, and Peled [31]. We begin by describing the general framework for this probabilistic argument below.

3.1.3 General framework

Throughout this work, we will use the notation of the Kuperberg, Lovett, and Peled paper [31], which we shorthand as KLP. Let B, A be finite sets and let $\phi : B \to \mathbb{Z}^A$ be a vector valued function. One can think of ϕ as described by a $|B| \times |A|$ matrix where the rows correspond to the evaluation of the function ϕ on the elements in B. In this setting [31] gives sufficient conditions for the existence of a small set $T \subset B$ such that

$$\frac{1}{|T|} \sum_{t \in T} \phi(t) = \frac{1}{|B|} \sum_{b \in B} \phi(b). \tag{3.4}$$

In the context of designs we can think of B as all the k-sets of $[n]$ and A as all the t-sets of $[n]$. ϕ denotes the inclusion function, that is $\phi(b)_a = 1_{a \subset b}$ where b is a k-set of $[n]$ and a is a t-set of $[n]$. Equation (3.4) is then equivalent to T being a t-(n,k,λ) design for an appropriate λ.

Next, we present the conditions under which KLP showed that there is a solution for (3.4). We start with a few useful notations. For $a \in A$ we denote by $\phi_a \in \mathbb{Z}^B$ the a-column of the matrix described by ϕ, namely $(\phi_a)_b = \phi(b)_a$. Let $V \subset \mathbb{Q}^B$ be the vector space spanned by the columns of this matrix $\{\phi_a : a \in A\}$. Observe that (3.4) depends only on V and not on $\{\phi_a : a \in A\}$, which is a specific choice of basis for V. We identify $f \in V$ with a function $f : B \to \mathbb{Q}$. Thus, we may reformulate (3.4) as

$$\frac{1}{|T|} \sum_{t \in T} f(t) = \frac{1}{|B|} \sum_{b \in B} f(b) \quad \forall f \in V. \tag{3.5}$$

In particular, we may assume without loss of generality that $\dim(V) = |A|$.

The conditions and results outlined below will depend only on the subspace V. However, it will be easier to present some of them with a specific choice of basis. We may assume this to

be an integer basis. Thus, we assume throughout that $\phi : B \to \mathbb{Z}^A$ is a map whose coordinate projections $\phi_a : B \to \mathbb{Z}$ are a basis for V.

Divisibility conditions

For T to be a valid set for (3.5) with $|T| = N$, we must have

$$\sum_{t \in T} f(t) = \frac{N}{|B|} \sum_{b \in B} f(b) \quad \forall f \in V.$$

In particular there must exist $\gamma \in \mathbb{Z}^B$ such that

$$\sum_{b \in B} \gamma_b f(b) = \frac{N}{|B|} \sum_{b \in B} f(b) \quad \forall f \in V. \tag{3.6}$$

The set of integers N satisfying (3.6) for some $\gamma \in \mathbb{Z}^B$ consists of all integer multiples of some minimal positive integer c_1. This is because if N_1 and N_2 are solutions then so is $N_1 - N_2$. Thus it follows that $|T|$ must be an integer multiple of c_1. This is the divisibility condition and c_1 is the divisibility parameter of V.

We can rephrase (3.6) as $\frac{N}{|B|} \sum_{b \in B} \phi(b)$ belongs to the lattice spanned by $\{\phi(b) : b \in B\}$.

Definition 3.1.2 (Lattice spanned by ϕ)**.** *We define $L(\phi)$ to be the lattice spanned by $\{\phi(b) : b \in B\}$.*

$$L(\phi) = \Big\{ \sum_{b \in B} n_b \cdot \phi(b) : n_b \in \mathbb{Z} \Big\} \subset \mathbb{Z}^A.$$

Note that since we assume that $\dim(V) = |A|$ we have that $L(\phi)$ is a full rank lattice.

Definition 3.1.3 (Divisibility parameter c_1)**.** *The divisibility parameter of V is the minimal integer $c_1 \geq 1$ that satisfies $\frac{c_1}{|B|} \sum_{b \in B} \phi(b) \in L(\phi)$. Note that it does not depend on the choice of basis for V which defines ϕ.*

Boundedness conditions

The second condition is about boundedness conditions for integer vectors which span V and its orthogonal dual. We start with some general definitions. Let $1 \leq p \leq \infty$. The ℓ_p norm of

a vector $\gamma \in \mathbb{Z}^B$ is $\|\gamma\|_p = (\sum_{b\in B} |\gamma_b|^p)^{1/p}$. Below we restrict our attention to $\|\gamma\|_1 = \sum_{b\in B} |\gamma_b|$ and $\|\gamma\|_\infty = \max_{b\in B} |\gamma_b|$.

Definition 3.1.4 (Bounded integer basis)**.** *Let $W \subset \mathbb{Q}^B$ be a vector space. For $1 \le p \le \infty$, we say that W has a c-bounded integer basis in ℓ_p if W is spanned by integer vectors whose ℓ_p norm is at most c. That is, if*

$$Span(\{\gamma \in W \cap \mathbb{Z}^B : \|\gamma\|_p \le c\}) = W.$$

Recall that $V \subset \mathbb{Q}^B$ is the vector space spanned by $\{\phi_a : a \in A\}$. We denote by $V^\perp$ the orthogonal complement of V in $\mathbb{Q}^B$, that is,

$$V^\perp := \{g \in \mathbb{Q}^B : \sum_{b\in B} f(b)g(b) = 0 \quad \forall f \in V\}.$$

Definition 3.1.5 (Boundedness parameters c_2, c_3)**.** *We impose two boundedness conditions:*

- *Let $c_2 \ge 1$ be such that V has a c_2-bounded integer basis in ℓ_∞.*
- *Let $c_3 \ge 1$ be such that $V^\perp$ has a c_3-bounded integer basis in ℓ_1.*

Symmetry conditions

Next we require some symmetry conditions from the space V. Given a permutation $\pi \in S_B$ and a vector $f \in \mathbb{Q}^B$, we denote by $\pi(f) \in \mathbb{Q}^B$ the vector obtained by permuting the coordinates of f, namely $\pi(f)_b = f_{\pi(b)}$.

Definition 3.1.6 (Symmetry group of V)**.** *The symmetry group of V, denoted $Sym(V)$, is the set of all permutations $\pi \in S_B$ which satisfy that $\pi(f) \in V$ for all $f \in V$.*

It is easy to verify that $Sym(V)$ is a subgroup of S_B, the symmetric group of permutations on B. Note that the condition $\pi \in Sym(V)$ can be equivalently case as the existence of an invertible linear map $\tau : \mathbb{Q}^A \to \mathbb{Q}^A$ such that

$$\phi(\pi(b)) = \tau(\phi(b)) \qquad \forall\, b \in B.$$

Definition 3.1.7 (Transitive symmetry group)**.** *The symmetry group of V is said to be transitive if it acts transitively on B. That is, for every $b_1, b_2 \in B$ there is $\pi \in Sym(V)$ such that $\pi(b_1) = b_2$.*

Constant functions condition

The last condition is very simple: we require that the constant functions belong to V.

Main theorem of KLP

We are now at a position to state the main theorem of KLP [31].

Theorem 3.1.8 (KLP Theorem)**.** *Let B be a finite set and let $V \subset \mathbb{Q}^B$ be the subspace of functions. Assume that the following holds for some integers $c_1, c_2, c_3 \geq 1$:*

- *Divisibility: c_1 is the divisibility parameter of V.*
- *Boundedness of V: V has a c_2-bounded integer basis in ℓ_∞.*
- *Boundedness of $V^\perp$: $V^\perp$ has a c_3-bounded integer basis in ℓ_1.*
- *Symmetry: The symmetry group of V is transitive.*
- *Constant functions: The constant functions belong to V.*

Let N is an integer multiple of c_1 satisfying

$$min(N, |B| - N) \geq C \cdot c_2 c_3^2 dim(V)^6 \log(2c_3 dim(V))^6,$$

where $C > 0$ is an absolute constant. Then there exists a subset $T \subset B$ of size $|T| = N$ satisfying

$$\frac{1}{|T|} \sum_{t \in T} \phi(t) = \frac{1}{|B|} \sum_{b \in B} \phi(b).$$

3.1.4 Our main theorem

Our main result is an extension of the KLP theorem (Theorem 3.1.8) to large sets. It will have many of the same conditions, except that we need to update the divisibility condition to require the size of each design to be $N = |B|/\ell$. Thus the new divisibility condition is

$$\frac{1}{l}\sum_{b\in B}\phi(b) \in L(\phi).$$

Note that as before, this condition depends only on V; it does not depend on the choice of basis for V which defines ϕ.

Theorem 3.1.9 (Main theorem)**.** *Let B be a finite set and let $V \subset \mathbb{Q}^B$ be the subspace of functions. Let also $l \geq 1$ be an integer. Assume that the following holds for some integers $c_2, c_3 \geq 1$:*

- *Divisibility: $\frac{1}{l}\sum_{b\in B}\phi(b) \in L(\phi)$.*
- *Boundedness of V: V has a c_2-bounded integer basis in ℓ_∞.*
- *Boundedness of $V^\perp$: $V^\perp$ has a c_3-bounded integer basis in ℓ_1.*
- *Symmetry: The symmetry group of V is transitive.*
- *Constant functions: The constant functions belong to V.*

Assume furthermore that

$$|B| \geq C\dim(V)^6 l^6 c_3^3 \log^3(\dim(V)c_2c_3l),$$

for some absolute constant $C > 0$. Then there exists a partition of B to $T_1, \ldots, T_l$, each of size $|T_i| = |B|/l$ such that

$$\sum_{t\in T_i}\phi(t) = \frac{1}{l}\sum_{b\in B}\phi(b) \quad \textit{for all} \quad i = 1,\ldots,l.$$

Theorem 3.1.1 follows as a special case of Theorem 3.1.9.

Proof of Theorem 3.1.1. To recall, in this setting we have B the set of all k-sets of $[n]$, A the set of all t-sets of $[n]$, $\phi : B \to \{0,1\}^A$ given by inclusion $\phi(b)_a = 1_{a \subset b}$ for $a \in A, b \in B$ and V the subspace spanned by $\{\phi_a : a \in A\}$.

KLP [31] showed (see Section 3.3 in the arxiv version) that in this setting, the subspace V has a transitive symmetric group, it contains the constant functions, and it has boundedness parameters $c_2 = 1, c_3 \leq (4en/t)^t$. Furthermore, the condition that the vector $\bar{\lambda} = (\lambda, \ldots, \lambda) \in L(\phi)$ is equivalent to the set of conditions

$$\binom{k-s}{t-s} \Bigg| \lambda \binom{n-s}{t-s} \quad \text{for all} \quad s = 0, \ldots, t.$$

(see Theorem 3.7 in [31]). In particular in our case $\lambda = \binom{n-t}{k-t}/l$ and hence the divisibility conditions in Theorem 3.1.9 are equivalent to the necessary divisibility conditions given in (3.3). To obtain the lower bound on $|B|$, lets fix k, t, l and let n be large enough. Then $|B| \approx n^k$, $\dim(V) \approx n^t$ and $c_3 \approx n^t$. Then if $k > 9t$ and n is large enough the lower bound on B holds. □

3.1.5 Proof overview

The high level idea, similar to [31], is to analyze the natural random process and show that with positive (yet exponentially small) probability a desired event occurs.

Say that a subset $T \subset B$ is "uniform" if

$$\frac{1}{|T|} \sum_{b \in T} \phi(b) = \frac{1}{|B|} \sum_{b \in B} \phi(b).$$

Equivalently, the "tests" defined by V cannot distinguish the uniform distribution over T from the uniform distribution over B.

Let $\tau : B \to [l]$ be a uniform partition of B into l sets. Let $T_i = \tau^{-1}(i)$ be the induced partition for $i = 1, \ldots, l$. We would like to analyze the event that each part is uniform. That is,

we would like to show that

$$\Pr[T_1, \ldots, T_l \text{ are uniform}] > 0. \tag{3.7}$$

Notice that under the same notations, the main result of [31] can be formulated as

$$\Pr[T_1 \text{ is uniform}] > 0.$$

The random process can be modeled as a random walk on a lattice. For $i = 1, \ldots, l$ let $X_i = \sum_{b \in T_i} \phi(b)$ be random variables taking values in $\mathbb{Z}^A$. Let $\lambda = \mathbb{E}[X_1] = \ldots = \mathbb{E}[X_l] \in \mathbb{Q}^{|A|}$. Note that if $X_1 = \ldots = X_{l-1} = \lambda$ then also $X_l = \lambda$. Let $X = (X_1, \ldots, X_{l-1}) \in \mathbb{Z}^{(l-1)|A|}$. Thus we can reformulate (3.7) as

$$\Pr[X = \mathbb{E}[X]] > 0. \tag{3.8}$$

Recall that each random variable X_i takes values in a full-dimensional sub-lattice of $\mathbb{Z}^A$ which we denoted $L(\phi)$. One can show that X takes values in the lattice $L(\phi)^{\otimes(l-1)}$, which is a full dimensional lattice in $\mathbb{Q}^{(l-1)|A|}$. In order to study the distribution of X, we apply a local central limit theorem. The same approach was applied in [31] in order to analyze the individual distribution of each X_i. Here, we extend the method to analyze their joint distribution, namely the distribution of X. This is accomplished by a careful analysis of the Fourier coefficients of X, which in turn relies on "coding theoretic" properties of the space V. Given this coding theoretic properties, we show that $\Pr[X = \mathbb{E}[X]]$ can be approximated by the density of a gaussian process with the same first and second moment as X at the point $\mathbb{E}[X]$. In particular, it is positive, which establishes the existence result.

3.1.6 Broader perspective

The current work falls into the regime of "rare events" in probabilistic analysis. It is very common that the probabilistic method, when applied to show that certain combinatorial objects exist (such as expander graphs, error correcting codes, etc) shows that a random sample

succeeds with high probability. The challenge then shifts to obtaining explicit constructions of such objects, with efficient algorithmic procedures whenever relevant (e.g. efficient decoding algorithms for codes).

However, there are several scenarios where the "vanilla" probabilistic method fails, and one is forced to develop much more fine tuned techniques to prove existence of the desired combinatorial objects. The current work falls into the regime where the random process is the natural one, but the analysis is much more delicate. Other examples of similar instances are the constructive proof of the Lovász local lemma (see e.g. [43, 44]), the works on interlacing families of polynomials (see e.g. [37, 38]), and the entire field of discrepancy theory (see e.g. the book [39]). In each such instance, new methods were developed to prove existence of the relevant objects, that go beyond simple probabilistic analysis.

There are several families of problems in combinatroics, for which the only known constructions are explicit and of algebraic or combinatorial nature. For example, this is the case for all types of local codes (such as locally testable, decodable, or correctable codes; PIR schemes; batch codes, and so on). It is also the case for Zarenkiewicz-type Ramsey problems in graph theory, about maximal bipartite graphs without certain induced subgraphs. Another well known example is the existence of Hadamard matrices. The lack of a probabilistic model for a solution may be seen as the reason why the existential results known for these problems are very sparse and ad-hoc.

In the current work, we show that for the problem of existence of large sets, one can move beyond explicit ad-hoc constructions, such as the one of Teirlinck [65], to a more rigorous understanding of when existence of large sets is possible. Of course, the next step in this line of research, after existence has been established, is to find explicit constructions. We leave this question for future research. Another question is whether the existence result can be established to the full spectrum of parameters, namely $k \geq t+1$ and any $\ell \geq 1$ (recall that our result requires that $k > 9t$). This seems to be possible by replacing the gaussian estimate by an estimate which uses higher moments of the distribution of the random variable being analyzed. We leave this

also for future research.

3.2 Preliminaries

Recall that $\phi : B \to \mathbb{Z}^A$ is a map, whose coordinate projections are $\phi_a : B \to \mathbb{Z}$. We defined V to be the subspace of $\mathbb{Q}^B$ spanned by $\{\phi_a : a \in A\}$. We may assume that that these form a basis for V, and hence $\dim(V) = |A|$.

Let $\tau : B \to [l]$ be a mapping that partitions B into l bins. Let $T_i := \{b \in B : \tau(b) = i\}$ for $i \in [\ell]$ be the induced partition of B. In order to prove Theorem 3.1.9 we are looking for a τ for which

$$\sum_{b \in T_i} \phi(b) = \frac{1}{l} \sum_{b \in B} \phi(b) \quad \text{for all} \quad i = 1, \ldots, l. \tag{3.9}$$

Note that it suffices to require that (3.9) holds for $i = 1, \ldots, l-1$, as then it automatically also holds for $i = l$. So from now on we only require that (3.9) holds for the first $l-1$ bins. We will choose a uniformly random mapping τ, and show that (3.9) holds with a positive probability.

We start with some definitions. Let $\Phi : B \times [l] \to \mathbb{Z}^{(l-1)|A|}$ be defined as follows. $\Phi(b,i) = (x_1, \ldots, x_{l-1})$, where $x_1, \ldots, x_{l-1} \in \mathbb{Z}^A$ are given by $x_j = \phi(b) \cdot 1_{i=j}$. Note that in particular $\Phi(b,l) = 0$. Next, define a random variable $X \in \mathbb{Z}^{(l-1)|A|}$ as

$$X := \sum_{b \in B} \Phi(b, \tau(b)).$$

The mean of X is

$$\mathbb{E}[X] = \left(\frac{1}{l} \sum_{b \in B} \phi(b), \ldots, \frac{1}{l} \sum_{b \in B} \phi(b) \right) \in \mathbb{Q}^{(l-1)|A|}.$$

Thus, proving Theorem 3.1.9 is equivalent to showing that

$$\Pr_{\tau}[X = \mathbb{E}[X]] > 0. \tag{3.10}$$

We start by computing the covariance matrix of X.

Claim 3.2.1. *The covariance matrix of X is the $(l-1)|A| \times (l-1)|A|$ matrix*

$$\Sigma[X] = R \otimes M$$

where R is the $|A| \times |A|$ matrix

$$R_{a,a'} = \sum_{b \in B} \phi(b)_a \phi(b)_{a'}$$

and M is the $(l-1) \times (l-1)$ matrix

$$M = \frac{1}{l^2} \begin{bmatrix} (l-1) & -1 & \dots & -1 \\ -1 & (l-1) & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & (l-1) \end{bmatrix}.$$

Proof. The random variables $\{\Phi(b, \tau(b)) : b \in B\}$ are independent, thus their contribution to the covariance matrix of X is additive. Fix $b \in B$. We compute the contribution of $\Phi(b, \tau(b))$ to the $(a,i),(a',i')$ entry of $\Sigma[X]$, where $a, a' \in A$ and $i, i' \in [l-1]$. The second moment is

$$\mathbb{E}_\tau[\Phi(b, \tau(b))_{a,i} \cdot \Phi(b, \tau(b))_{a',i'}] = \frac{1}{l} \phi(b)_a \phi(b)_{a'} \cdot 1_{i=i'}.$$

The expectation product is

$$\mathbb{E}_\tau[\Phi(b, \tau(b))_{a,i}] \cdot \mathbb{E}_\tau[\Phi(b, \tau(b))_{a',i'}] = \frac{1}{l^2} \phi(b)_a \phi(b)_{a'}.$$

Thus

$$\Sigma[X]_{(a,i),(a',i')} = \sum_{b \in B} \phi(b)_a \phi(b)_{a'} \left(\frac{1}{l} \cdot 1_{i=i'} - \frac{1}{l^2} \right) = R_{a,a'} \cdot M_{i,i'} = (R \otimes M)_{(a,i),(a',i')}.$$

□

Similar to the proof in KLP we would be interested in the lattice in which X resides. Recall that $L(\phi)$ is the lattice in $\mathbb{Z}^{|A|}$ spanned by the image of ϕ. We similarly define $L(\Phi)$.

Definition 3.2.2 (Lattice spanned by Φ)**.** *We define $L(\Phi)$ to be the lattice spanned by $\{\Phi(b,i) : b \in B, i \in [l]\}$. Namely,*

$$L(\Phi) := \left\{ \left(\sum_{b_1 \in B} n_{b_1} \cdot \phi(b_1), .., \sum_{b_j \in B} n_{b_j} \cdot \phi(b_j), .., \sum_{b_{l-1} \in B} n_{b_{l-1}} \cdot \phi(b_{l-1}) \right) : n_{b_j} \in \mathbb{Z}, j \in [l-1] \right\}.$$

Note that since $\dim(V) = |A|$ then $L(\phi)$ is a full rank lattice in $\mathbb{Z}^{|A|}$. Hence $L(\Phi) = L(\phi)^{\otimes(l-1)}$ is a full rank lattice in $\mathbb{Z}^{(l-1)|A|}$.

Similar to KLP we use Fourier analysis to study the distribution of X. The Fourier transform of X is the function $\widehat{X} : \mathbb{R}^{(l-1)|A|} \to \mathbb{C}$ defined by

$$\widehat{X}(\Theta) = \mathbb{E}_X[e^{2\pi i \langle X, \Theta \rangle}].$$

Note that $\widehat{X}$ is periodic. Concretely, let $L(\Phi)$ denote the dual lattice to $L(\Phi)$,

$$L(\Phi) := \left\{ \Theta \in \mathbb{R}^{(l-1)A} : \langle \Lambda, \Theta \rangle \in \mathbb{Z} \quad \forall \Lambda \in L(\Phi) \right\}.$$

Note that if $\Theta \in L(\Phi)$ then $\widehat{X}(\Theta + \Theta') = \widehat{X}(\Theta')$ for all $\Theta' \in \mathbb{R}^{(l-1)|A|}$, and $\widehat{X}(\Theta) = 1$ iff $\Theta \in L(\Phi)$. As $L(\Phi)$ is a full rank lattice it follows that $L(\Phi)$ is also be a full rank lattice and $\det(L(\Phi))\det(L(\Phi)) = 1$. Thus studying $\widehat{X}$ on any fundamental domain of $L(\Phi)$ would be sufficient to study the behavior of $\widehat{X}$ on $\mathbb{R}^{(l-1)|A|}$. Similar to KLP we work with a natural fundamental domain defined by a norm related to the covariance matrix of X.

Definition 3.2.3 (R-norm)**.** *For $\Theta = (\theta_1, ..., \theta_{l-1}) \in \mathbb{R}^{(l-1)A}$ we define the norm $\|\cdot\|_R$ as*

$$\|\Theta\|_R := \max_{j \in [l-1]} \left(\frac{1}{|B|} \theta_j^t R \theta_j \right)^{1/2} = \max_{j \in [l-1]} \left(\frac{1}{|B|} \sum_{b \in B} \langle \phi(b), \theta_j \rangle^2 \right)^{1/2}.$$

We define two related notions. Balls around zero in the R-norm are defined as

$$\mathscr{B}_R(\varepsilon) := \{\Theta \in \mathbb{R}^{(l-1)A} : \|\Theta\|_R \leq \varepsilon\}.$$

The Voronoi cell of 0 in the R-norm, with respect to the dual lattice $L(\Phi)$, is

$$D := \Big\{\Theta \in \mathbb{R}^{(l-1)A} : \|\Theta\|_R < \|\Theta - \alpha\|_R \quad \forall \alpha \in L(\Phi) \setminus \{0\}\Big\}.$$

Observe that D is a fundamental domain of $L(\Phi)$ up to a set of measure zero (its boundary), which we can ignore in our calculations. Then we have the following Fourier inversion formula over lattices: for every $\Gamma \in L(\Phi)$ it holds that

$$\Pr[X = \Gamma] = \frac{1}{vol(D)} \int_D \widehat{X}(\Theta) e^{-2\pi i \langle \Gamma, \Theta \rangle} d\Theta = \det(L(\Phi)) \int_D \widehat{X}(\Theta) e^{-2\pi i \langle \Gamma, \Theta \rangle} d\Theta. \tag{3.11}$$

Note that this formula holds true for any fundamental region of $L(\Phi)$ but we chose it to be the Voronoi cell D arising from the norm $\|\cdot\|_R$ because it would help in the computations later on. In order to prove (3.10), we specialize (3.11) to $\Gamma = \mathbb{E}[X]$ and obtain

$$\Pr[X = \mathbb{E}[X]] = \det(L(\Phi)) \int_D \widehat{X}(\Theta) e^{-2\pi i \langle \mathbb{E}[X], \Theta \rangle} d\Theta. \tag{3.12}$$

In the next section, we approximate this by a Gaussian estimate.

3.3 Gaussian estimate

In order to estimate (3.12), let Y be a Gaussian random variable in $\mathbb{R}^{(l-1)|A|}$ with the same mean and covariance as X. The density f_Y of Y is given by

$$f_Y(x) = \frac{\exp(-\frac{1}{2}(x - \mathbb{E}[X])^t \Sigma[X]^{-1}(x - \mathbb{E}[X]))}{(2\pi)^{\frac{(l-1)|A|}{2}} \sqrt{\det(\Sigma[X])}}. \tag{3.13}$$

The Fourier transform of Y equals

$$\widehat{Y}(\Theta) := \mathbb{E}[e^{2\pi i\langle Y,\Theta\rangle}] = e^{2\pi i\langle \mathbb{E}[X],\Theta\rangle - 2\pi^2\Theta^t\Sigma[X]\Theta}. \tag{3.14}$$

The inverse Fourier transform applied to Y yields

$$f_Y(x) = \int_{\mathbb{R}^{(l-1)A}} \widehat{Y}(\Theta)e^{-2\pi i\langle x,\Theta\rangle}d\Theta \qquad \forall x \in \mathbb{R}^{(l-1)A}. \tag{3.15}$$

We show that $\Pr[X = \mathbb{E}[X]]$ can be approximated by an appropriate scaling of $f_Y(\mathbb{E}[X])$. By (3.12) we have

$$\frac{\Pr[X = \mathbb{E}[X]]}{\det(L(\Phi))} - f_Y(\mathbb{E}[X]) = \int_D \widehat{X}(\Theta)e^{-2\pi i\langle \mathbb{E}[X],\Theta\rangle}d\Theta - \int_{\mathbb{R}^{(l-1)A}} \widehat{Y}(\Theta)e^{-2\pi i\langle \mathbb{E}[X],\Theta\rangle}d\Theta.$$

Note that by plugging $x = \mathbb{E}[X]$ in (3.13) we obtain that

$$f_Y(\mathbb{E}[X]) = \frac{1}{(2\pi)^{\frac{(l-1)|A|}{2}}\sqrt{\det(\Sigma[X])}}. \tag{3.16}$$

We will show that $|\frac{\Pr[X=\mathbb{E}[X]]}{\det(L(\Phi))} - f_Y(\mathbb{E}[X])| \ll f_Y(\mathbb{E}[X])$. For $\varepsilon > 0$ to be chosen later, we will bound it by

$$\begin{aligned}&\left|\frac{\Pr[X = \mathbb{E}[X]]}{\det(L(\Phi))} - f_Y(\mathbb{E}[X])\right| \leq \\ &\underbrace{\int_{\mathscr{B}_R(\varepsilon)} |\widehat{X}(\Theta) - \widehat{Y}(\Theta)|d\Theta}_{=I_1} + \underbrace{\int_{D\setminus\mathscr{B}_R(\varepsilon)} |\widehat{X}(\Theta)|d\Theta}_{=I_2} + \underbrace{\int_{\mathbb{R}^{(l-1)A}\setminus\mathscr{B}_R(\varepsilon)} |\widehat{Y}(\Theta)|d\Theta}_{=I_3}.\end{aligned} \tag{3.17}$$

At a high level, the upper bound is obtained by comparing $\widehat{X}(\Theta)$ and $\widehat{Y}(\Theta)$ in a small enough ball; and upper bounding their absolute value outside this ball. Observe that we need ε to be small enough so that $\mathscr{B}_R(\varepsilon) \subset D$.

3.3.1 Norms on $\mathbb{R}^{|A|}$ induced by ϕ

The following key technical lemmas from [31] are very useful in bounding the integrals. We begin with defining few norms which are all functions of ϕ.

Definition 3.3.1 (Norms on $\mathbb{R}^{|A|}$ induced by ϕ)**.** *For $\theta \in \mathbb{R}^{|A|}$ define the following norms:*

- $\|\theta\|_{\phi,\infty} = \max_{b\in B} |\langle \phi(b), \theta\rangle|$.
- $\|\theta\|_{\phi,2} = \left(\frac{1}{|B|}\sum_{b\in B} |\langle \phi(b), \theta\rangle|^2\right)^{1/2}$.

Furthermore, for $b \in B$ let $\langle \phi(b), \theta\rangle = n_b + r_b$ where $n_b \in \mathbb{Z}$ and $r_b \in [-1/2, 1/2)$. Define

- $|||\theta|||_{\phi,\infty} = \max_{b\in B} |r_b|$.
- $\|\theta\|_{\phi,2} = \left(\frac{1}{|B|}\sum_{b\in B} |r_b|^2\right)^{1/2}$.

Note that if $\theta' \in L(\phi)$ then $\langle \phi(b), \theta+\theta'\rangle - \langle \phi(b), \theta\rangle \in \mathbb{Z}$ for all $b \in B$. In particular, $\|\theta+\theta'\|_{\phi,\infty} = \|\theta\|_{\phi,\infty}$ and $\|\theta+\theta'\|_{\phi,2} = \|\theta\|_{\phi,2}$. The following lemmas from [31] relates the above norms.

Lemma 3.3.2 (Lemma 4.4 in [31])**.** *For every $\theta \in \mathbb{R}^A$ it holds that*

$$\|\theta\|_{\phi,\infty} \le M\|\theta\|_{\phi,2}$$

and

$$\|\theta\|_{\phi,\infty} \le M\|\theta\|_{\phi,2}.$$

Here, $M := C\left(|A|\log(2c_2|A|)\right)^{3/2}$ for some absolute constant $C > 0$.

Lemma 3.3.3 (Claim 4.12 in [31])**.** *Assume that for $\theta \in \mathbb{R}^A$ it holds that*

$$\|\theta\|_{\phi,\infty} \le \frac{1}{c_3}.$$

Then there exists $\theta' \in L(\phi)$ such that $\langle \theta - \theta', \phi(b) \rangle \in [-1/2, 1/2]$ for all $b \in B$. In particular

$$\|\theta - \theta'\|_{\phi,2} = \|\theta\|_{\phi,2}.$$

3.3.2 Norms on $\mathbb{R}^{(l-1)|A|}$ induced by Φ

We extend the previous definitions to norms on $\mathbb{R}^{(l-1)|A|}$ induced by Φ, and prove related lemmas relating the different norms.

Definition 3.3.4 (Generalizing the norms to $\mathbb{R}^{(l-1)|A|}$)**.** *For $\Theta = (\theta_1, \ldots, \theta_{l-1}) \in \mathbb{R}^{(l-1)|A|}$ define the following norms:*

- $\|\Theta\|_{\Phi,\infty} = \max_{j \in [l-1]} \|\theta_j\|_{\phi,\infty}$
- $\|\Theta\|_{\Phi,2} = \max_{j \in [l-1]} \|\theta_j\|_{\phi,2}$
- $\|\Theta\|_{\Phi,\infty} = \max_{j \in [l-1]} \|\theta_j\|_{\phi,\infty}$
- $\|\Theta\|_{\Phi,2} = \max_{j \in [l-1]} \|\theta_j\|_{\phi,2}$

Observe that $\|\cdot\|_{\Phi,2}$ is the same as the R-norm $\|\cdot\|_R$ we defined before. Similar to before, if $\Theta' \in L(\Phi)$ then $\|\Theta + \Theta'\|_{\Phi,\infty} = \|\Theta\|_{\Phi,\infty}$ and $\|\Theta + \Theta'\|_{\Phi,2} = \|\Theta\|_{\Phi,2}$.

The following extends Lemma 3.3.2 and Lemma 3.3.3 to the norms induced by Φ.

Lemma 3.3.5. *For the same M defined in Lemma 3.3.2, for every $\Theta \in \mathbb{R}^{(l-1)|A|}$ it holds that*

$$\|\Theta\|_{\Phi,\infty} \leq M \|\Theta\|_{\Phi,2}$$

and

$$\|\Theta\|_{\Phi,\infty} \leq M \|\Theta\|_{\Phi,2}.$$

Proof. Let $\Theta = (\theta_1, \ldots, \theta_{l-1})$. Then using Lemma 3.3.2 we have

$$\|\Theta\|_{\Phi,\infty} = \max_{j \in [l-1]} \|\theta_j\|_{\phi,\infty} \leq \max_{j \in [l-1]} M \|\theta_j\|_{\phi,2} = M \|\Theta\|_{\Phi,2}$$

and

$$\|\Theta\|_{\Phi,\infty} = \max_{j\in[l-1]} \|\theta_j\|_{\phi,\infty} \leq \max_{j\in[l-1]} M\|\theta_j\|_{\phi,2} = M\|\Theta\|_{\Phi,2}.$$

□

Lemma 3.3.6. *Assume that for $\Theta \in \mathbb{R}^{(l-1)A}$ it holds that*

$$\|\Theta\|_{\Phi,\infty} \leq \frac{1}{c_3}.$$

Then there exists $\Theta' \in L(\Phi)$ such that $\langle \Theta - \Theta', \Phi(b,j)\rangle \in [-1/2, 1/2]$ for all $b \in B, j \in [l-1]$. In particular

$$\|\Theta - \Theta'\|_{\Phi,2} = \|\Theta\|_{\Phi,2}.$$

Proof. Let $\Theta = (\theta_1, \ldots, \theta_{l-1})$. We have $\|\theta_j\|_{\phi,\infty} \leq \frac{1}{c_3}$ for all $j \in [l-1]$. Then using Lemma 3.3.3 we get that there exist $\theta'_1, \ldots, \theta'_{l-1} \in L(\phi)$ such that $\langle \theta_j - \theta'_j, \phi(b)\rangle \in [-1/2, 1/2]$ for all $b \in B$. The lemma follows for $\Theta' = (\theta'_1, \ldots, \theta'_{l-1}) \in L(\Phi)$. □

3.3.3 Estimates for balls in the Voronoi cell

To recall, we need $\varepsilon > 0$ to be small enough so that $\mathscr{B}_R(\varepsilon)$ is contained in the Voronoi cell D. The following Lemma utilizes Lemma 3.3.5 to achieve that.

Lemma 3.3.7. *If $\varepsilon < \frac{1}{2M}$ then $\mathscr{B}_R(\varepsilon) \subset D$.*

Proof. Let $\Theta = (\theta_1, \ldots, \theta_{l-1}) \in L(\Phi) \setminus \{0\}$. By definition $\langle \phi(b), \theta_j\rangle \in \mathbb{Z}$ for all $b \in B, j \in [l-1]$. Since $L(\phi)$ is of full rank and $\Theta \neq 0$, there exists some $b \in B, j \in [l-1]$ for which $|\langle \phi(b), \theta_j\rangle| \geq 1$. Thus

$$\|\Theta\|_{\Phi,\infty} \geq 1.$$

By Lemma 3.3.5 if follows that

$$\|\Theta\|_R = \|\Theta\|_{\Phi,2} \geq 1/M.$$

Thus, if $\Theta' \in \mathscr{B}_R(\varepsilon)$ for $\varepsilon < 1/2M$ then

$$\|\Theta - \Theta'\|_R \geq \|\Theta\|_R - \|\Theta'\|_R \geq 1/M - \varepsilon > 1/2M \geq \|\Theta'\|_R.$$

Hence $\mathscr{B}_R(\varepsilon) \subset D$ for any $\varepsilon < \frac{1}{2M}$. □

Let $\Theta \in D \setminus \mathscr{B}_R(\varepsilon)$. Clearly, its $\|\cdot\|_{\Phi,2}$ norm is noticeable (at least ε). We show that also its $\|\cdot\|_{\Phi,2}$ norm is noticeable. This will later be useful in bounding $\hat{X}(\Theta)$ in $D \setminus \mathscr{B}_R(\varepsilon)$.

Lemma 3.3.8. *Assume that $c_3 \geq 2$ and $\varepsilon < 1/c_3 M$. Let $\Theta \in D \setminus \mathscr{B}_R(\varepsilon)$. Then $\|\Theta\|_{\Phi,2} > \varepsilon$.*

Proof. Note that the conditions of Lemma 3.3.7 hold, and so $\mathscr{B}_R(\varepsilon) \subset D$. Assume towards contradiction that $\|\Theta\|_{\Phi,2} \leq \varepsilon$. Applying Lemma 3.3.5 gives $\|\Theta\|_{\Phi,\infty} \leq \varepsilon M \leq \frac{1}{c_3}$. Applying Lemma 3.3.6, this implies that there exists $\Theta' \in L(\Phi)$ for which $\|\Theta - \Theta'\|_{\Phi,2} = \|\Theta\|_{\Phi,2} \leq \varepsilon$. However, as $\Theta \in D$ we have $\|\Theta\|_{\Phi,2} \leq \|\Theta - \Theta'\|_{\Phi,2} \leq \varepsilon$, which gives that $\Theta \in \mathscr{B}_R(\varepsilon)$, a contradiction. □

3.3.4 Bounding the integrals

The following lemmas provide the necessary bounds on the integrals I_1, I_2, I_3, as defined in (3.17). The proofs are deferred to Section 3.4.

Lemma 3.3.9. *Assume that $\varepsilon \leq (CM|B|)^{-1/3}$. Then*

$$I_1 \leq \frac{Cl^3 M|A|^{3/2}}{|B|^{1/2}} \cdot f_Y(\mathbb{E}[X]).$$

Here $C > 0$ is some large enough absolute constant.

Lemma 3.3.10. *Assume that $c_3 \geq 2$ and $\varepsilon \leq 1/c_3 M$. Then*

$$I_2 \leq \frac{1}{\det(L(\Phi))} \exp\left(-\frac{|B|\varepsilon^2}{l^2}\right)$$

Lemma 3.3.11. *For any $\varepsilon > 0$ it holds that*

$$I_3 \leq f_Y(\mathbb{E}[X]) \cdot (l-1)2^{|A|/2} \exp\left(-\frac{\pi^2 |B| \varepsilon^2}{l^2}\right).$$

3.3.5 Putting it all together

Let $C_1, C_2, \ldots$ be unspecified absolute constants below. By choosing an appropriate basis for V which is c_2-bounded in ℓ_∞, we may assume that $\phi : B \to \mathbb{Z}^A$ where $|\phi(b)|_a \leq c_2$ for all $a \in A, b \in B$.

Set $\varepsilon = (C_1 MB)^{-1/3}$ so that we may apply Lemma 3.3.9, and assume that $\varepsilon \leq 1/c_3 M$ so that we may apply Lemma 3.3.10. We thus have

$$\Pr[X = \mathbb{E}[X]] = \det(L(\Phi)) f_Y(\mathbb{E}[X])(1 + \alpha_1 + \alpha_3) + \alpha_2,$$

where

$$\begin{aligned}
|\alpha_1| &= \frac{C_1 l^3 M |A|^{3/2}}{|B|^{1/2}}, \\
|\alpha_2| &= \exp\left(-\frac{|B|\varepsilon^2}{l^2}\right) = \exp\left(-C_2 \frac{|B|^{1/3}}{l^2 M^{2/3}}\right), \\
|\alpha_3| &= (l-1)2^{|A|/2} \exp\left(-\frac{\pi^2 |B| \varepsilon^2}{l^2}\right) \leq l 2^{|A|} \exp\left(-C_3 \frac{|B|^{1/3}}{l^2 M^{2/3}}\right).
\end{aligned}$$

We would like that $|\alpha_1|, |\alpha_3| \leq 1/4$, which requires that

$$|B| \geq C_4 |A|^3 M^2 l^6 c_3^3$$

Thus

$$\Pr[X = \mathbb{E}[X]] \geq \frac{1}{2} \det(L(\Phi)) f_Y(\mathbb{E}[X]) + \alpha_2.$$

We assume that $\phi : B \to \mathbb{Z}^A$, so $L(\Phi)$ is an integer lattice and hence $\det(L(\Phi)) \geq 1$. We next

lower bound $f_Y(\mathbb{E}[X])$. We have by (3.16) that

$$f_Y(\mathbb{E}[X]) = \frac{1}{(2\pi)^{\frac{(l-1)|A|}{2}}\sqrt{\det(\Sigma[X])}}.$$

We assume that ϕ is spanned by integer vectors of maximum entry c_2, so we can bound each entry of $\Sigma[X]$ by

$$|\Sigma[X]_{(a,i),(a',i')}| \leq \sum_{b\in B} |\phi(b)_a \phi(b)_{a'}| \leq |B|c_2^2.$$

Thus

$$\det(\Sigma[X]) \leq (|A||B|c_2^2)^{|A|}.$$

In order to require $\alpha_2 \leq (1/4) f_Y(\mathbb{E}[X])$, say, we need to require that

$$|B| \geq C_5 |A|^3 M^2 l^6 \log(|A|Ml).$$

Putting it all together, and plugging in the value of M from Lemma 3.3.2, as long as

$$|B| \geq C|A|^6 l^6 c_3^3 \log^3(|A|c_2 c_3 l),$$

we have that

$$\Pr[X = \mathbb{E}[X]] \geq \frac{1}{4}\det(L(\Phi)) f_Y(\mathbb{E}[X]) > 0.$$

3.4 Bounding the integrals

3.4.1 Bounding I_1

Recall that $I_1 = \int_{\mathscr{B}_R(\varepsilon)} |\hat{X}(\Theta) - \hat{Y}(\Theta)| d\Theta$. We will bound it by bounding pointwise the difference $|\hat{X}(\Theta) - \hat{Y}(\Theta)|$ and integrating it.

We first compute an exact formula for $\widehat{X}(\Theta)$. Recall that $X = \sum_{b\in B} \Phi(b, \tau(b))$ where

$\tau(b) \in [l]$ are independently chosen. Thus

$$\widehat{X}(\Theta) = \mathbb{E}_X\left[e^{2\pi i\langle X,\Theta\rangle}\right] = \prod_{b\in B}\left[\frac{1}{l}\left(1+\sum_{j=1}^{l-1} e^{2\pi i\langle \phi(b),\theta_j\rangle}\right)\right]. \tag{3.18}$$

Fix $\Theta = (\theta_1,\ldots,\theta_{l-1})$. To simplify notations, let $x_{b,j} = 2\pi\langle\phi(b),\theta_j\rangle$ and $x_b = (x_{b,1}\ldots x_{b,l-1}) \in \mathbb{R}^{l-1}$. Define the function $f:\mathbb{R}^{l-1}\to\mathbb{C}$ given by $f(x) = \frac{1}{l}\left(1+\sum_{j=1}^{l-1} e^{ix_j}\right)$. Then we can simplify (3.18) as

$$\widehat{X}(\Theta) = \prod_{b\in B} f(x_b). \tag{3.19}$$

We next approximate $\log f(x)$. We use the shorthand $O(z)$ to denote a (possible complex) value, whose absolute value is bounded by Cz for some unspecified absolute constant $C > 0$. For $x = (x_1,\ldots,x_{l-1})$ we denote $|x| = \max_j |x_j|$.

Claim 3.4.1. *Let $x = (x_1,\ldots,x_{l-1}) \in \mathbb{R}^{l-1}$ with $|x| \le 1$. Then*

$$f(x) = \exp\left(i\frac{1}{l}\sum_j x_j - \frac{1}{2l}\left(1-\frac{1}{l}\right)\sum_j x_j^2 + \frac{1}{2l^2}\sum_{j\neq j'} x_j x_{j'} + O\left(|x|^3\right)\right).$$

Proof. Let $y = \frac{1}{l}\sum_{j=1}^{l-1}(e^{ix_j}-1)$ so that $f(x) = 1+y$. The condition $|x|\le 1$ guarantees that $|y| < 1$, so the Taylor expansion for $\log(1+y)$ converges and gives

$$\log(f(x)) = \log(1+y) = y - \frac{y^2}{2} + O(|y|^3).$$

One can verify that $|y| \le O(|x|)$, that

$$y = i\frac{1}{l}\sum_j x_j - \frac{1}{2l}\sum_j x_j^2 + O\left(|x|^3\right).$$

and that

$$y^2 = -\frac{1}{l^2}\left(\sum_j x_j\right)^2 + O\left(|x|^3\right).$$

Combining these gives the required result. □

Applying Claim 3.4.1 to (3.19) allows us to approximate $\widehat{X}(\Theta)$ as equal to

$$\exp\left(\frac{2\pi i}{l}\sum_{\substack{b\in B\\ j\in[l-1]}}\langle\phi(b),\theta_j\rangle-\frac{2\pi^2(l-1)}{l^2}\sum_{\substack{b\in B\\ j\in[l-1]}}\langle\phi(b),\theta_j\rangle^2+\frac{2\pi^2}{l^2}\sum_{\substack{b\in B\\ j\neq j'}}\langle\phi(b),\theta_j\rangle\langle\phi(b),\theta_{j'}\rangle+\delta(\Theta)\right)$$

which can be rephrased as

$$\widehat{X}(\theta)=\exp\left(2\pi i\langle\mathbb{E}[X],\Theta\rangle-2\pi^2\Theta^t\Sigma[X]\Theta+\delta(\Theta)\right). \tag{3.20}$$

The error term $\delta(\Theta)$ is bounded by

$$\begin{aligned}\delta(\Theta)&\leq O\left(\sum_{b\in B}|x_b|^3\right)=O\left(\sum_{b\in B}\max_{j\in[l-1]}|\langle\phi(b),\theta_j\rangle|^3\right)\\&\leq\left(\max_{b\in B,j\in[l-1]}|\langle\phi(b),\theta_j\rangle|\right)\left(\sum_{b\in B}\max_{j\in[l-1]}|\langle\phi(b),\theta_j\rangle|^2\right)\\&=\|\Theta\|_{\Phi,\infty}\cdot|B|\|\Theta\|_{\Phi,2}^2.\end{aligned}$$

By Lemma 3.3.5 we have $\|\Theta\|_{\Phi,\infty}\leq M\|\Theta\|_{\Phi,2}$, and hence as $\|\Theta\|_{\Phi,2}=\|\Theta\|_R$ we conclude that

$$\delta(\Theta)\leq C_1M|B|\|\Theta\|_R^3, \tag{3.21}$$

where $C_1>0$ is some absolute constant.

Next, we apply these estimates to bound the integral I_1. Recall that by (3.14) we have

$$\widehat{Y}(\Theta):=\exp(2\pi i\langle\mathbb{E}[X],\Theta\rangle-2\pi^2\Theta^t\Sigma[X]\Theta).$$

Thus we can bound I_1 by

$$I_1=\int_{\mathscr{B}_R(\varepsilon)}|\hat{X}(\Theta)-\hat{Y}(\Theta)|d\Theta\leq\int_{\mathscr{B}_R(\varepsilon)}e^{-2\pi^2\Theta^t\Sigma[X]\Theta}|e^{\delta(\Theta)}-1|d\Theta.$$

We assume that $\varepsilon > 0$ is small enough so that $C_1 M|B|\varepsilon^3 \leq 1$, so that for all for $\Theta \in \mathscr{B}_R(\varepsilon)$ we have

$$|e^{\delta(\Theta)} - 1| \leq 2\delta(\Theta) \leq 2C_1 M|B|\|\Theta\|_R^3.$$

Thus

$$I_1 \leq 2C_1 M|B| \int_{\mathscr{B}_R(\varepsilon)} e^{-2\pi^2 \Theta^t \Sigma[X]\Theta} \|\Theta\|_R^3 d\Theta \leq 2C_1 M|B| \int_{\mathbb{R}^{(l-1)A}} e^{-2\pi^2 \Theta^t \Sigma[X]\Theta} \|\Theta\|_R^3 d\Theta.$$

Next, we evaluate the integral on the right. Let Z be a Gaussian random variable in $\mathbb{R}^{(l-1)|A|}$ with mean zero and covariance matrix $\frac{1}{4\pi^2}\Sigma[X]^{-1}$. Then the density of Z is

$$f_Z(\Theta) = (2\pi)^{\frac{(l-1)|A|}{2}} \sqrt{\det(\Sigma)} e^{-2\pi^2 \Theta^t \Sigma[X]\Theta} = \frac{1}{f_Y(\mathbb{E}[X])} e^{-2\pi^2 \Theta^t \Sigma[X]\Theta},$$

where we have used (3.16). Hence

$$\int_{\mathbb{R}^{(l-1)A}} e^{-2\pi^2 \Theta^t \Sigma[X]\Theta} \|\Theta\|_R^3 d\Theta = f_Y(\mathbb{E}[X]) \cdot \mathbb{E}[\|Z\|_R^3].$$

Let $G \in \mathbb{R}^{(l-1)|A|}$ be standard multivariate Gaussian with mean zero and identity covariance matrix. Recall that by Claim 3.2.1 we have $\Sigma[X] = R \otimes M$. In particular, $\Sigma[X]$ is positive definite, so its root exists. So we may take $Z = \frac{1}{2\pi}\Sigma[X]^{-1/2} G$. We have

$$\Sigma[X] = R \otimes M = R \otimes (U^t D U)$$

where D is a diagonal matrix with diagonal $(1/l^2, 1/l, \ldots, 1/l)$ and U is an orthogonal matrix. Thus

$$\Sigma[X]^{-1/2} = R^{-1/2} \otimes (U^t D^{-1/2} U).$$

Note that $D^{-1/2}$ is a diagonal matrix with diagonal $(l, \sqrt{l}, \ldots, \sqrt{l})$.

Let $G = (G_1, \ldots, G_{l-1})$ with $G_i \in \mathbb{R}^{|A|}$ and similarly $Z = (Z_1, \ldots, Z_{l-1})$ with $Z_i \in \mathbb{R}^{|A|}$.

We can express $Z_1, \ldots, Z_{l-1}$ as

$$Z_1 = \frac{l}{2\pi} R^{-1/2} \sum_{k=1}^{l-1} U_{1,k} G_k$$
$$Z_j = \frac{\sqrt{l}}{2\pi} R^{-1/2} \sum_{k=1}^{l-1} U_{j,k} G_k \qquad j = 2, \ldots, l-1.$$

Let $G^j = \sum_{k=1}^{l-1} U_{j,k} G_k$. Since U is an orthogonal matrix, we have that $(G^1, \ldots, G^{l-1})$ is also a standard multivariate Gaussian $\mathbb{R}^{(l-1)|A|}$ with mean zero and identity covarince matrix. Thus we have

$$Z_1 = \frac{l}{2\pi} R^{-1/2} G^1$$
$$Z_j = \frac{\sqrt{l}}{2\pi} R^{-1/2} G^j \qquad j = 2, \ldots, l-1.$$

That is, $Z_1, \ldots, Z_{l-1}$ are independent Gaussian random variables with mean zero, where Z_1 has covariance matrix $\frac{l^2}{4\pi^2} R^{-1}$ and for $j = 2, \ldots, l-1$ we have that Z_j has covariance matrix $\frac{l}{4\pi^2} R^{-1}$. We may thus bound

$$\begin{aligned}
\mathbb{E}_Z \left[\|Z\|_R^3 \right] &= \mathbb{E}_Z \left[\max_j \left(\frac{1}{|B|} Z_j^t R Z_j \right)^{3/2} \right] \\
&\leq \mathbb{E}_Z \left[\sum_j \left(\frac{1}{|B|} Z_j^t R Z_j \right)^{3/2} \right] = \sum_j \mathbb{E}_Z \left[\left(\frac{1}{|B|} Z_j^t R Z_j \right)^{3/2} \right] \\
&= \left(\left(\frac{l^2}{4\pi^2 |B|} \right)^{\frac{3}{2}} + (l-2) \left(\frac{l}{4\pi^2 |B|} \right)^{\frac{3}{2}} \right) \mathbb{E} \left[\|G'\|_2^3 \right] \\
&\leq \frac{2l^3}{(4\pi^2)^{3/2} |B|^{3/2}} \mathbb{E} \left[\|G'\|_2^3 \right]
\end{aligned}$$

where G' is a standard multivariate Gaussian random vector in $\mathbb{R}^A$ with mean zero and identity covariance matrix. Note that by Jensen's inequality $\mathbb{E}[\|G'\|_2^3] \leq \mathbb{E}[\|G'\|_2^4]^{3/4} \leq 4^{3/4} |A|^{3/2}$. Thus

we can summarize that

$$I_1 \leq O\left(\frac{l^3 M|A|^{3/2}}{|B|^{1/2}}\right) \cdot f_Y(\mathbb{E}[X]).$$

3.4.2 Bounding I_2

Recall that $I_2 = \int_{D\setminus\mathscr{B}_R(\varepsilon)} |\widehat{X}(\Theta)| d\Theta$. We upper bound I_2 by proving an upper bound on $|\hat{X}(\Theta)|$ in $D \setminus \mathscr{B}_R(\varepsilon)$.

Fix $\Theta = (\theta_1, \ldots, \theta_{l-1}) \in D$ where we assume $\|\Theta\|_{\Phi,2} = \|\Theta\|_R \geq \varepsilon$. Our goal is to upper bound $\hat{X}(\Theta)$. Let $\langle \phi(b), \theta_j \rangle = n_{b,j} + r_{b,j}$ where $n_{b,j} \in \mathbb{Z}$ and $r_b \in [-1/2, 1/2)$. By (3.19) we have

$$\widehat{X}(\Theta) = \prod_{b\in B}\left[\frac{1}{l}\left(1 + \sum_{j=1}^{l-1} e^{2\pi i\langle\theta_j,\phi(b)\rangle}\right)\right] = \prod_{b\in B}\left[\frac{1}{l}\left(1 + \sum_{j=1}^{l-1} e^{2\pi i\cdot r_{b,j}}\right)\right] = \prod_{b\in B} f(2\pi\cdot r_b),$$

where $f(x) = \frac{1}{l}\left(1 + \sum_{j=1}^{l-1} e^{ix_j}\right)$ and $r_b = (r_{b,1}, \ldots, r_{b,l-1})$. Recall that $|x| = \max_j |x_j|$.

Claim 3.4.2. *Let $x \in \mathbb{R}^{l-1}$ be with $|x| \leq \pi$. Then $|f(x)| \leq \exp(-|x|^2/8l)$.*

Proof. Let $x_j = |x|$. Then $|f(x)| \leq \frac{l-2}{l} + \frac{2}{l}|\frac{1+e^{ix_j}}{2}|$. If $z \in [-\pi, \pi]$ then $|\frac{1+e^{iz}}{2}| \leq e^{-z^2/8}$. One can verify that

$$\log|f(x)| \leq \log\left(1 - \frac{2}{l}\left(e^{-|x|^2/8} - 1\right)\right) \leq -\frac{|x|^2}{8l}.$$

□

Thus we have

$$\log|\widehat{X}(\Theta)| \leq -\frac{4\pi^2}{8l}\sum_{b\in B}|r_b|^2 \leq -\frac{1}{l^2}\sum_{b\in B, j\in[l-1]} r_{b,j}^2 = -\frac{|B|}{l^2}\|\Theta\|_{\Phi,2}^2.$$

Next, assume that $\varepsilon \leq 1/c_3 M$. By Lemma 3.3.8 we have that $\|\Theta\|_{\Phi,2} \geq \varepsilon$. Thus

$$|\widehat{X}(\Theta)| \leq \exp(-|B|\varepsilon^2/l^2).$$

Thus we may bound

$$I_2 \leq vol(D)\exp(-|B|\varepsilon^2/l^2) = \frac{1}{\det(L(\Phi))}\exp(-|B|\varepsilon^2/l^2).$$

3.4.3 Bounding I_3

Recall that

$$I_3 = \int_{\mathbb{R}^{(l-1)A}\setminus\mathscr{B}_R(\varepsilon)} |\hat{Y}(\Theta)|d\Theta = \int_{\mathbb{R}^{(l-1)A}\setminus\mathscr{B}_R(\varepsilon)} e^{-2\pi^2\Theta^t\Sigma[X]\Theta}d\Theta.$$

As in the calculation of the bound for I_1, let $Z \in \mathbb{R}^{(l-1)|A|}$ be Gaussian random variable with mean zero and covariance matrix $\frac{1}{4\pi^2}\Sigma[X]^{-1}$. Then

$$I_3 = f_Y(\mathbb{E}[X]) \cdot \Pr\left[\|Z\|_R > \varepsilon\right].$$

Recall that we showed that if we set $Z = (Z_1,\ldots,Z_{l-1})$, then $Z_1,\ldots,Z_{l-1} \in \mathbb{R}^A$ are independent Gaussian random variables in with mean zero, where Z_1 has covariance matrix $\frac{l^2}{4\pi^2}R^{-1}$ and Z_j has covariance matrix $\frac{l}{4\pi^2}R^{-1}$ for $j = 2,\ldots,l-1$. We may thus bound

$$\begin{aligned}
\Pr\left[\|Z\|_R > \varepsilon\right] &= \Pr_Z\left[\max_j\left(\frac{1}{|B|}Z_j^t R Z_j\right) > \varepsilon^2\right] \leq \sum_j \Pr_{Z_j}\left[\left(\frac{1}{|B|}Z_j^t R Z_j\right) > \varepsilon^2\right] \\
&= \Pr_{G'}\left[\|G'\|_2^2 > \frac{4\pi^2|B|\varepsilon^2}{l^2}\right] + (l-2)\Pr_{G'}\left[\|G'\|_2^2 > \frac{4\pi^2|B|\varepsilon^2}{l}\right] \\
&\leq (l-1)\Pr_{G'}\left[\|G'\|_2^2 > \frac{4\pi^2|B|\varepsilon^2}{l^2}\right],
\end{aligned}$$

where $G' \in \mathbb{R}^A$ is a Gaussian random variable with mean zero and identity covariance matrix.

In order to bound $\Pr_{G'}\left[\|G'\|_2^2 > \rho\right]$ we note that for any $t < 1/2$, it holds that $\mathbb{E}\left[e^{t\|G'\|_2^2}\right] =$

$(1-2t)^{-|A|/2}$. Fixing $t = 1/4$ and apply the Markov inequality gives

$$\Pr_{G'}\left[\|G'\|_2^2 > \rho\right] \leq \frac{\mathbb{E}\left[e^{\|G'\|_2^2/4}\right]}{e^{\rho/4}} = 2^{|A|/2}e^{-\rho/4}.$$

So

$$I_3 \leq f_Y(\mathbb{E}[X]) \cdot (l-1)2^{|A|/2}e^{-\frac{\pi^2|B|\varepsilon^2}{l^2}}.$$

Chapter 3, in full, is a reprint of the material as it appears in SIAM Symposium on Discrete Algorithms SODA 2018 and Journal of Combinatorial Theory Series A JCTA. Karingula, Sankeerth Rao; Lovett, Shachar; Vardy, Alexander. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=67b493e60ba9.

[2] S. Balaji, M. N. Krishnan, M. Vajha, V. Ramkumar, B. Sasidharan, and P. V. Kumar. Erasure coding for distributed storage: An overview. *Science China Information Sciences*, 61(10):100301, 2018.

[3] S. Balaji and P. V. Kumar. On partial maximally-recoverable and maximally-recoverable codes. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1881–1885. IEEE, 2015.

[4] S. 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.

[5] A. Barvinok. *A course in convexity*, volume 54. American Mathematical Society Providence, 2002.

[6] L. J. Billera and A. Sarangarajan. The combinatorics of permutation polytopes. In *Formal power series and algebraic combinatorics*, volume 24, pages 1–23, 1994.

[7] J. Bourgain, V. H. Vu, and P. M. Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

[8] M. Braverman. Towards deterministic tree code constructions. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 161–167, 2012.

[9] A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

[10] A. Cayley. On the triadic arrangements of seven and fifteen things. *London Edinburgh and Dublin Philos. Mag. and J. Sci.*, 37:50–53, 1850.

[11] Y. Chee. The basis reduction algorithm and existence of combinatorial designs. *M. Math. Thesis, Department of Computer Science, University of Waterloo, Ontario, Canada*, 1989.

[12] Y. Chee and S. Ling. Constructions for q-ary constant-weight codes. *IEEE Transactions on Information Theory*, 53:135–146, 2007.

[13] G. Cohen, B. Haeupler, and L. J. Schulman. Explicit binary tree codes with polylogarithmic size alphabet. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 535–544, 2018.

[14] S. H. Dau, W. Song, and C. Yuen. On the existence of MDS codes over small fields with constrained generator matrices. In *2014 IEEE International Symposium on Information Theory*, pages 1787–1791. IEEE, 2014.

[15] C. Esseen. On the Kolmogorov-Rogozin inequality for the concentration function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):210–216, 1966.

[16] T. Etzion. On threshold schemes from large sets. *Journal on Combinatorial Designs*, 4:323–338, 1996.

[17] S. Fenner, R. Gurjar, and T. Thierauf. Bipartite perfect matching is in quasi-nc. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 754–763. ACM, 2016.

[18] P. Gopalan, G. Hu, S. Kopparty, S. Saraf, C. Wang, and S. Yekhanin. Maximally recoverable codes for grid-like topologies. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2092–2108. SIAM, 2017.

[19] P. Gopalan, C. Huang, B. Jenkins, and S. Yekhanin. Explicit maximally recoverable codes with locality. *IEEE Transactions on Information Theory*, 60(9):5245–5256, 2014.

[20] S. Gopi, V. Guruswami, and S. Yekhanin. Maximally recoverable LRCs: A field size lower bound and constructions for few heavy parities. *IEEE Transactions on Information Theory*, 2020.

[21] A. Heidarzadeh and A. Sprintson. An algebraic-combinatorial proof technique for the GM-MDS conjecture. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 11–15. IEEE, 2017.

[22] S. Hoang Dau, W. Song, and C. Yuen. On the existence of MDS codes over small fields with constrained generator matrices. *arXiv*, pages arXiv–1401, 2014.

[23] J. Kahn, J. Komlós, and E. Szemerédi. On the probability that a random ± 1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240, 1995.

[24] D. Kane, S. Lovett, and S. Rao. The independence number of the Birkhoff polytope graph, and applications to maximally recoverable codes. *SIAM Journal on Computing*, 48(4):1425–1435, 2019.

[25] Y. Katznelson. Singular matrices and a uniform bound for congruence groups of $SL_n(\mathbb{Z})$. *Duke Mathematical Journal*, 69(1):121–136, 1993.

[26] P. Keevash. The existence of designs. *arXiv:1401.3665*, 2014.

[27] G. Khosrovshahi and R. Laue. t-designs with $t \geq 3$,. *CRC Handbook of Combinatorial Designs*, 2:79–100, 2006.

[28] G. Khosrovshahi and B. Tayfeh-Rezaie. Large sets of t-designs through partitionable sets: A survey. *Discrete Math.*, 306:2993–3004, 2006.

[29] T. Kirkman. Note on an unanswered prize question. *Cambridge and Dublin Mathematical Journal*, 5:255–262, 1850.

[30] J. Komlós. On determinant of $(0,1)$ matrices. *Studia Science Mathematics Hungarica*, 2:7–21, 1967.

[31] G. Kuperberg, S. Lovett, and R. Peled. Probabilistic existence of regular combinatorial structures. *Proc. 44-th ACM Symp.Theory of Computing (STOC)*, pages 1091–1106, 2013.

[32] V. Lalitha and S. V. Lokam. Weight enumerators and higher support weights of maximally recoverable codes. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 835–842. IEEE, 2015.

[33] L. Leindler. On a certain converse of Hölder's inequality. In *Proceedings of the 1971 Oberwolfach Conference, BirkhHuser Verlag. Basel-Stuttgart*, 1972.

[34] S. Lovett. MDS matrices over small fields: A proof of the GM-MDS conjecture. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 194–199. IEEE, 2018.

[35] J. Lu. On large sets of disjoint steiner triple systems: Parts i, ii, and iii. *Journal of Combinatorial Theory Series A (JCTA)*, 34:140–182, 1983.

[36] J. Lu. On large sets of disjoint steiner triple systems: Parts iv, v, and vi. *Journal of Combinatorial Theory Series A (JCTA)*, 37:136–192, 1984.

[37] A. Marcus, D. Spielman, and N. Srivastava. Interlacing families i: Bipartite ramanujan graphs of all degrees. *Proc. of the 54-th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 529–537, 2013.

[38] A. Marcus, D. Spielman, and N. Srivastava. Interlacing families ii: Mixed characteristic polynomials and the kadison-singer problem. *Annals of Mathematics*, 2013.

[39] J. Matousek. Geometric discrepancy: An illustrated guide. *Springer Science & Business Media, 2009*, 18.

[40] E. Mendelsohn and A. Rosa. One-factorizations of the complete graph: A survey. *Journal on Graph Theory*, pages 43–65, 1985.

[41] V. D. Milman and G. Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer, 2009.

[42] C. Moore and L. J. Schulman. Tree codes and a conjecture on exponential sums. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 145–154, 2014.

[43] R. Moser. A constructive proof of the lovász local lemma. *Proceedings of the 41-st annual ACM Symposium on Theory of computing(STOC)*, pages 343–350, 2009.

[44] R. Moser and G. Tardos. A constructive proof of the general lovász local lemma. *Journal of the ACM (JACM)*, 2010.

[45] S. Onn. Geometry, complexity, and combinatorics of permutation polytopes. *Journal of Combinatorial Theory, Series A*, 64(1):31–49, 1993.

[46] J. S. Plank, M. Blaum, and J. L. Hafner. Sd codes: erasure codes designed for how storage systems really fail. In *FAST*, pages 95–104, 2013.

[47] A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[48] M. Rudelson. Lecture notes on non-asymptotic theory of random matrices. 2013.

[49] M. Rudelson and R. Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[50] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

[51] B. Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, volume 203. Springer Science & Business Media, 2013.

[52] P. Schellenberg and D. Stinson. Threshold schemes from combinatorial designs. *J. Combin. Math. Combin. Comput.*, 1989.

[53] L. J. Schulman. Deterministic coding for interactive communication. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 747–756, 1993.

[54] L. J. Schulman. Coding for interactive communication. *IEEE transactions on information theory*, 42(6):1745–1756, 1996.

[55] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *Journal of the ACM (JACM)*, 27(4):701–717, 1980.

[56] B. Segre. Curve razionali normali ek-archi negli spazi finiti. *Annali di Matematica Pura ed Applicata*, 39(1):357–379, 1955.

[57] R. Singleton. Maximum distance q-nary codes. *IEEE Transactions on Information Theory*, 10(2):116–118, 1964.

[58] D. Stinson and S. Vanstone. A combinatorial approach to threshold schemes. *SIAM J. Discrete Math.*, 1988.

[59] I. Tamo and A. Barg. Bounds on locally recoverable codes with multiple recovering sets. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 691–695. IEEE, 2014.

[60] I. Tamo and A. Barg. A family of optimal locally recoverable codes. *IEEE Transactions on Information Theory*, 60(8):4661–4676, 2014.

[61] T. Tao and V. Vu. On random ± 1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006.

[62] T. Tao and V. Vu. On the singularity probability of random Bernoulli matrices. *Journal of the American Mathematical Society*, 20(3):603–628, 2007.

[63] L. Teirlinck. Non-trivial t-designs without repeated blocks exist for all t. *Discrete Math.*, 1987.

[64] L. Teirlinck. Locally trivial t-designs and t-designs without repeated blocks. *Discrete Math.*, 1989.

[65] L. Teirlinck. A completion of lu's determination of the spectrum of large sets of disjoint steiner triple systems. *Journal of Combinatorial Theory Series A (JCTA)*, 1991.

[66] L. Teirlinck. Large sets of disjoint designs and related structures. *Contemporary Design Theory*, 1992.

[67] K. Tikhomirov. Singularity of random Bernoulli matrices. *Annals of Mathematics*, 191(2):593–634, 2020.

[68] H. Yildiz and B. Hassibi. Further progress on the GM-MDS conjecture for Reed-Solomon codes. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 16–20. IEEE, 2018.

[69] R. Zippel. Probabilistic algorithms for sparse polynomials. In *International symposium on symbolic and algebraic manipulation*, pages 216–226. Springer, 1979.