SEVERAL PROBLEMS IN LINEAR ALGEBRAIC AND ADDITIVE COMBINATORICS

By

DANIEL SCHEINERMAN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Mathematics

Written under the direction of

Swastik Kopparty

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2019

ABSTRACT OF THE DISSERTATION

Several problems in linear algebraic and additive combinatorics

By DANIEL SCHEINERMAN

Dissertation Director: Swastik Kopparty

This thesis studies three problems in linear algebraic and additive combinatorics.

Our first result gives new upper bounds for the determinant of an $n \times n$ zero-one matrix containing kn ones. Our results improve upon a result of Ryser for $k = o(n^{1/3})$. For fixed $k \geq 3$ it was an open question [BR18] whether Hadamard's inequality could be exponentially improved. We answer this in the affirmative. Our approach revolves around studying $m \times n$ matrices whose rows sum to k and bounding their Gram determinants. For the class of $n \times n$ matrices whose rows sum to k we show that Ryser's result can be improved for $k \leq \sqrt{n/10}$. Our technique also allows us to give upper bounds when these matrices are perturbed.

Our second result concerns a question in additive combinatorics. For a prime $p > 2$, we say a nonempty set $A \subseteq \mathbb{F}_p$ is *unique sum free* (USF) if every element of the sumset $A+A$ can be written as a sum of two elements from A in at least two different ways. That is for any $s \in A + A$ there exist a, b, c, d with $\{a, b\} \neq \{c, d\}$ such that $s = a + b = c + d$. If $\mu(p)$ is the size of the smallest USF set in $\mathbb{F}_p$ it is straightforward to show that $\mu(p) = O(\sqrt{p})$. Kopparty [Kop17] conjectured that $\mu(p) = \Theta(\sqrt{p})$. However, we show constructively that $\mu(p) = O(\log^2 p)$.

Our third result concerns a graph theoretic problem on the Hamming cube, Q_n. For

a graph, G, we say a proper k-coloring of G is a fall k-coloring if each vertex is adjacent to a vertex in each of the $k-1$ other color classes. A result of Laskar and Lyle [LL09] shows that for $k \neq 3$ and n sufficiently large Q_n has a fall k-coloring. It is natural to identify the Hamming cube, Q_n, with the vector space $\mathbb{F}_2^n$. In this context we may seek fall k-colorings of $\mathbb{F}_2^n$ in which each color class is an affine subspace. Our main result is that for even k and n sufficiently large there exist affine fall k-colorings of $\mathbb{F}_2^n$. In particular, we show these exist for the same range of values of n as in the construction of Laskar and Lyle.

Acknowledgements

First and foremost I want to say thank you for the love and support of my wife, Leora. You have enriched my life more than I can put to words and I can say with certainty I would not be where I am today without you. I love you. Thank you to Carmella for being a source of endless enthusiasm and love. You are my pride and joy. Carmella's consultation on the diagrams in this thesis were invaluable, particularly her (strong) thoughts on the choice of colors.

I am very lucky to have been raised by an Ema and Abba who loved me unconditionally and instilled in me a love of learning. I am equally lucky to have grown up with and enjoyed (and continue to enjoy) the teasing of my siblings Chel, Gnomie and Jonah. Thank you to all my family and friends for your support.

It has been a privilege and a pleasure to have been advised by Swastik Kopparty. You bring a wonderful combination of insight, enthusiasm and humor to mathematics. I have learned so much from you and enjoyed all the problems you have presented to me, even if I have only made progress on an ε fraction.

I would like to thank Steven J. Miller, a mentor of mine at Brown University, who gave me my first taste of mathematical research. I am grateful to Jozsef Beck, Bhargav Narayanan and Vsevolod Lev for serving as members of my dissertation committee and to David Cash, Michael Saks and Shubhangi Saraf for serving on my oral exam committee.

In addition to those named elsewhere, this thesis benefited from discussions with a great number of mathematicians. Compiling their names was a daunting, but also fun exercise. As best as a I can manage to recall, thank you to Yonah Biers-Ariel, Reinier Broker, John Chiarelli, Tim Chow, Rebecca Coulson, Ron Fertig, Nathan Fox, Cole Franks, Surya Teja Gavva, Richard Gottesman, Ben Howard, Jonathan Jaquette, Katie

McKeon, Tim Naumovitz, Jinyoung Park, Aditya Potukuchi, Fei Qi, Rob Rhoades, Matthew Russell, Edward Scheinerman, Jonah Scheinerman, Justin Semonsen, Jeff VanderKam and Vidya Venkateswaran.

I have always been a kinetic learner and, as such, many of the key insights in this thesis came to me as I was taking Barney for a walk. So thank you for being a good walking partner and for sleeping on Leora's side of the bed.

Dedication

To Leora and Carmella

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis addresses three combinatorial questions. Combinatorics is a broad field that draws on a wide array of techniques. This thesis is no exception, however, the common thread of the topics herein are linear algebra and additive combinatorics. The tools we apply will, broadly speaking, come from linear algebra and graph theory.

1.1 Upper bounds for determinants of sparse zero-one matrices

Our first results, found in Chapter 2, regard the determinants of matrices of zeros and ones. For integers $1 \leq k \leq n$ we consider three classes of zero-one matrices containing kn ones:

$$S(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and the rows and columns of } A \text{ sum to } k\}$$
$$R(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and the rows of } A \text{ sum to } k\}$$
$$T(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and contains } kn \text{ ones}\}.$$

We denote the maximum determinant over matrices in these three classes by

$$M(n,k) = \max_{A \in S(n,k)} \det(A)$$
$$M_R(n,k) = \max_{A \in R(n,k)} \det(A)$$
$$M_T(n,k) = \max_{A \in T(n,k)} \det(A).$$

Clearly, $S(n,k) \subseteq R(n,k) \subseteq T(n,k)$ and thus $M(n,k) \leq M_R(n,k) \leq M_T(n,k)$. Since the rows of a matrix in $R(n,k)$ have norm $\sqrt{k}$, we have from Hadamard's inequality that $M_R(n,k) \leq k^{n/2}$. Since the rows of $A \in T(n,k)$ have average sum k applying Hadamard's inequality to the rows of A in conjunction with the AM-GM inequality

gives the same upper bound: $M_T(n,k) \leq k^{n/2}$. Ryser [Rys56] showed the following improvement. Let $\lambda = k(k-1)/(n-1)$. Then $M_T(n,k) \leq k(k-\lambda)^{(n-1)/2}$. Further, Ryser showed that this bound was tight if and only if A is the incidence matrix of an (n,k,λ) combinatorial design. In this case we would have $A \in S(n,k)$, so although $S(n,k)$ is a more constrained class than $T(n,k)$, Ryser's bound is tight even for $M(n,k)$ in many instances. Notice that if $k < \sqrt{n}$ then $\lambda < 1$ and thus in particular λ is not an integer so no (n,k,λ) combinatorial design exists. Thus we may hope to improve Ryser's result for $k < \sqrt{n}$. In particular, for fixed k we may seek upper bounds for the quantities

$$\limsup_{n\to\infty} M(n,k)^{1/n} \leq \limsup_{n\to\infty} M_R(n,k)^{1/n} \leq \limsup_{n\to\infty} M_T(n,k)^{1/n}.$$

Hadamard's inequality gives the upper bound $\sqrt{k}$ for each, and, since $\lambda \to 0$ as $n \to \infty$, Ryser's result gives the same. However, for $k = 2$, Bruhn and Rautenbach [BR18] show that $\limsup_{n\to\infty} M_R(n,2)^{1/n} \leq 2^{1/3}$ and $\limsup_{n\to\infty} M_T(n,2)^{1/n} \leq 6^{1/6}$. Note that $2^{1/3} < 6^{1/6} < \sqrt{2}$. Bruhn and Rautenbach ask: for $k > 2$ can Ryser's theorem be exponentially improved? We answer this question in the affirmative. We show for $k \geq 2$ that there exists a function $c_k < \sqrt{k}$ such that $M_T(n,k) \leq c_k^n$ and thus in particular $\limsup_{n\to\infty} M_T(n,k)^{1/n} \leq c_k$. For example, for $k = 3$ we show that $\limsup_{n\to\infty} M_T(n,3)^{1/n} \leq 24^{1/6} \approx 1.6984 < \sqrt{3}$.

Our results stem from generalizing the class $R(n,k)$ to rectangular matrices and bounding their Gram determinants. We define

$$R(m,n,k) = \{A \ : \ A \text{ is } m \times n \text{ and the rows of } A \text{ sum to } k\},$$

and for any $m \times n$ matrix, A, we define $\mathrm{Vol}(A) = \sqrt{\det(AA^T)}$. Our main tool is the following generalization of Hadamard's inequality. If any $m \times n$ matrix, A, is partitioned into two horizontal blocks A_1 and A_2 with dimensions $m_1 \times n$ and $m_2 \times n$ respectively (thus $m_1 + m_2 = m$) then $\mathrm{Vol}(A) \leq \mathrm{Vol}(A_1)\,\mathrm{Vol}(A_2)$. Rather than considering the rows of $A \in R(m,n,k)$ individually we show that sets of size $q > 1$ will "overlap" and have Gram determinants smaller than given by Hadamard's inequality. Notice that $R(m,n,k)$ has a recursive structure; if we remove q rows from A the resulting matrix

lies in $R(m-q,n,k)$. Varying our strategy in selecting these sets of rows we will give a few different bounds. Although $T(n,k)$ does not have the same recursive structure the following intuition will prove true: for $A \in T(n,k)$ to have a large determinant it cannot have too many rows not summing to k. Therefore, we will be able to apply some of the results we develop for matrices in $R(n,k)$ to matrices in $T(n,k)$.

1.2 Unique sum free sets

Chapter 3 of this thesis regards a problem in additive combinatorics. For a prime $p > 2$, we say a nonempty set $A \subseteq \mathbb{F}_p$ is *unique sum free* (USF) if every element of the sumset $A + A$ can be written as a sum of two elements from A in at least two different ways. That is for any $s \in A + A$ there exist a, b, c, d with $\{a, b\} \neq \{c, d\}$ such that $s = a + b = c + d$. Notice that finding large USF sets is trivial. Indeed, for $p > 2$ if we take $A = \mathbb{F}_p$ this is a USF set. We are interested in how small a USF set can be. We define for $p > 2$,

$$\mu(p) = \min\{|A| \ : \ A \subseteq \mathbb{F}_p \text{ is USF}\}.$$

A pigeon hole argument shows that $\mu(p) > \log_4 p$ and a linear algebraic argument improves this lower bound to $\mu(p) > \log_2 p$. For $\varepsilon > 0$, a random subset of $\mathbb{F}_p$ of size $p^{1/2+\varepsilon}$ can be shown to be USF with high probability, so USF sets of size $p^{1/2+\varepsilon}$ abound. Furthermore, let $a = \lfloor \sqrt{p} \rfloor$ and let k be the largest integer so that $ka < p$. Then $A = \{0, 1, 2, \ldots, 2a, 3a, 4a, \ldots, ka\}$ taken as residues modulo p is easily seen to be USF. Thus $\mu(p) = O(\sqrt{p})$. Closing the gap between the lower bound $\log_2 p$ and the upper bound $O(\sqrt{p})$ was our goal. As we will show in Figure 3.1, a plot of p versus $\mu(p)$ for small primes closely matches a plot of the curve $2\sqrt{p}$. As such, Kopparty [Kop17] conjectured that $\mu(p) = \Theta(\sqrt{p})$. However, the main result of Chapter 3 is a construction demonstrating that $\mu(p) = O(\log^2 p)$.

Our construction has two main ingredients. The first is the observation that if A is an arbitrary subset of $\mathbb{F}_p$ then $B = A + A$ has many non-unique sums. This follows since if $a, b, c, d \in A$ are distinct then $s = a + b + c + d \in B + B$ and we can write s as a sum of elements from B in (at least) two distinct ways $s = (a + b) + (c + d)$

and $s = (a + c) + (b + d)$. Notice that, in general, B will not be USF. For example, the sum $s = a + a + a + a$ can not, in general, be obtained in a manner distinct from $2a + 2a$. However, if A has the property that $2a = b + c$ with $b, c \in A$ and $b \neq c$ then $s = (a + a) + (b + c) = (a + b) + (a + c)$ and we are in business. This is our second ingredient. We of course have not resolved all the cases necessary to prove B is USF, but these can be found in Chapter 3. A set $A \subseteq \mathbb{F}_p$ with the property that for any $a \in A$ there exist distinct $b, c \in A$ such that $2a = b + c$ is called balanced [NQ08] and the fact that there exist balanced sets of size $O(\log p)$ was originally shown in [Str76].

We will generalize the notion of balanced sets to sets with no unique triples (NUT). We say $A \subseteq \mathbb{F}_p$ is NUT if for all $a \in A$ there exist $b, c, d \in A$ not all equal so that $3a = b + c + d$. We generalize the result above to show that if A is NUT then $B = A + A$ is USF. We will develop the notion of *regular* balanced and NUT sets and give experimental evidence that small examples exist for infinitely many primes, p.

1.3 Affine fall k-colorings of the Hamming cube

Our final result, discussed in Chapter 4, regards coloring the Hamming cube, Q_n. For any simple graph, G, and integer, k, we say a proper k-coloring of G is a fall k-coloring if every vertex in G is adjacent to a vertex in each of the other $k - 1$ color classes. The question of for which k, n does Q_n have a fall k-coloring originates in [DHH$^+$00]. We note that if V_i is the i-th color class of a fall k-coloring that V_i is an independent set since the coloring is proper and further that V_i is dominating since every vertex not colored i is, by definition, adjacent to a vertex in V_i. So the color classes form an independent dominating partition of the vertices. As such, some authors [GH13] refer to fall k-colorings as idomatic partitions.

Since Q_n is bipartite it is trivial to see that for $n \geq 2$, Q_n is fall 2-colorable. In fact, it is straightforward to see that if k is a power of 2 then Q_n has a fall k-coloring for all $n \geq k - 1$. It is known [LL09] that Q_n does not have a fall 3-coloring for any n. In [LL09], Laskar and Lyle show that for $k \neq 3$, if $2^{a-1} < k \leq 2^a$ then Q_n has a fall k-coloring for $n \geq 2^a - 1$. For example, Q_n has a fall 20-coloring for $n \geq 31$.

It is natural to identify Q_n with the vector space $\mathbb{F}_2^n$. Denote by e_i the i-th standard basis vector. Then if $u, v \in \mathbb{F}_2^n$ we say $u \sim v$ if and only if $u + v = e_i$ for some i. In this context, we may ask algebraic questions about the color classes. In particular we are interested in colorings where each color class, V_i, is an affine subspace. We call such colorings affine fall k-colorings. The construction of Laskar and Lyle does not give affine fall k-colorings when k is not a power of 2. Our main result is that for even k such that $2^{a-1} < k \leq 2^a$, Q_n has an affine fall k-coloring for $n \geq 2^a - 1$. Thus for even k we can obtain the same minimum dimension as the construction of Laskar and Lyle, but with the additional property that the color classes are affine subspaces.

Appropriate to the search for fall k-colorings, our result is based on trees. In particular, we construct an object similar to a parity decision tree [O'D14]. We construct a full binary tree with k leaves that classifies the vectors in $v \in \mathbb{F}_2^n$ into k color classes. To a decision node, t, we associate a vector $h(t)$ and the node proceeds to its right or left child depending on the parity of $\langle h(t), v \rangle$. For the class of decision trees we consider, we show that our construction achieves the minimum possible dimension, n.

Computer experimentation and computation played a large role in each of the three topics in this thesis. The primary tools used were Sage [S$^+$17] and Julia [BKSE12]. The figures in this thesis were generated using Matplotlib [Hun07]. Within Julia, the following packages were invaluable: Combinatorics, JuMP [DHL17], Memoize, Nemo [FHHJ17] and PyPlot. Fundamental experiments that lead to Chapter 4 were done using integer linear programming where the solver Gurobi [GO18] proved very useful.

Chapter 2

Upper bounds for determinants of sparse zero-one matrices

2.1 Introduction

Hadamard's maximum determinant problem [Had93] asks for each positive integer, n, what is the largest possible determinant over all 2^{n^2} zero-one matrices of order n. The problem is well studied [Syl67, Wil46, Rys56, BC72, Orr05, OS07] and open questions remain. In this chapter we consider a sparse version of this question. For a parameter k, we may consider the combinatorial class of $n \times n$ zero-one matrices containing kn ones. We are interested in giving an upper bound on their determinants. There are three natural (nested) classes of such matrices to consider.

Definition 2.1. *Let $1 \leq k \leq n$. We define the following three classes of zero-one matrices.*

$$S(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and the rows and columns of } A \text{ sum to } k\}$$

$$R(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and the rows of } A \text{ sum to } k\}$$

$$T(n,k) = \{A \ : \ A \text{ is } n \times n \text{ and contains } kn \text{ ones}\}.$$

We can ask a version of Hadamard's maximum determinant problem for each of these classes.

Definition 2.2.

$$M(n,k) = \max_{A \in S(n,k)} \det(A)$$

$$M_R(n,k) = \max_{A \in R(n,k)} \det(A)$$

$$M_T(n,k) = \max_{A \in T(n,k)} \det(A).$$

The notation $S(n,k)$, $M(n,k)$ can be found, for example, in [FvdD97, LLR99]. As $S(n,k) \subseteq R(n,k) \subseteq T(n,k)$ we have $M(n,k) \leq M_R(n,k) \leq M_T(n,k)$. Note that taking the absolute value of $\det(A)$ in Definition 2.2 does not alter the definition as for each of the three classes of matrices swapping two rows (or columns) of A maintains membership in the class and negates $\det(A)$.

We note as an aside that the related question of bounding the permanent of such matrices has received considerable attention. If A is an $n \times n$ zero-one matrix with rows summing to r_i, then we can associate to A a bipartite graph, G, whose partition classes are $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_n\}$ where $a_{i,j} = 1$ precisely when $u_i \sim v_j$. For each i the degree of u_i is r_i. Then $\operatorname{perm}(A)$ counts the number of perfect matchings in this bipartite graph. Minc [Min63] conjectured, and Bregman [Bre73] first showed the following tight inequality:

$$\operatorname{perm}(A) \leq \prod_{i=1}^{n} (r_i!)^{1/r_i}.$$

Schrijver [Sch78] gave a short proof of the Bregman-Minc inequality and a probabilistic proof is given by Alon and Spencer [AS00].

Returning to bounding determinants, the easiest upper bound for $M_R(n,k)$ (and thus $M(n,k)$) comes from Hadamard's inequality [Had93] which gives $M_R(n,k) \leq k^{n/2}$ since each row has norm exactly $\sqrt{k}$. If $A \in T(n,k)$ then its rows have average sum k and so using the AM-GM inequality and Hadamard's inequality the bound $\det(A) \leq k^{n/2}$ still applies. Thus $M_T(n,k) \leq k^{n/2}$. Ryser [Rys56] proved a strengthening of this result. First we recall the definition of an (n,k,λ) combinatorial design [BJL85].

Definition 2.3. *An (n,k,λ) combinatorial design is a collection of sets $S_1, \ldots, S_n$ such that $\bigcup_i S_i = \{1, \ldots, n\}$ and the following hold.*

1. *$|S_i| = k$ for $i = 1, \ldots, n$.*
2. *For all $i \neq j$, $|S_i \cap S_j| = \lambda$.*
3. *For all $i = 1, \ldots, n$, there are exactly k values of j such that $i \in S_j$.*

We note that the third criterion follows from the previous, but mention it for emphasis. One can show that for such a design to exist we must have $\lambda = k(k-1)/(n-1)$. The incidence matrix of an (n,k,λ) combinatorial design is the $n \times n$ matrix, A, so

that $a_{i,j} = 1$ if $i \in S_j$ and $a_{i,j} = 0$ otherwise. Notice that the Gram matrix AA^T is independent of our choice of presentation. If I_n is the $n \times n$ identity matrix and J_n is the $n \times n$ all ones matrix then $AA^T = (k - \lambda)I_n + \lambda J_n$. Matrices of this type will reappear throughout this chapter. Now we present Ryser's result as it appears in [Rys56].

Theorem 2.4 (Ryser's Theorem)**.** *Let A be an $n \times n$ zero-one matrix with a total of t ones. Let $k = t/n$ and $\lambda = k(k-1)/(n-1)$. Then*

$$\det(A) \leq k(k - \lambda)^{\frac{1}{2}(n-1)}$$

with equality holding if and only if A is the incidence matrix of an (n, k, λ) combinatorial design. □

Thus we have $M_T(n, k) \leq k(k - \lambda)^{\frac{1}{2}(n-1)}$. Notice that when Theorem 2.4 is tight we have $A \in S(n, k)$ as the incidence matrix of an (n, k, λ) combinatorial design has constant row and column sums. Thus in this case $M(n, k), M_R(n, k)$ and $M_T(n, k)$ coincide. Note that if, for example, $k = \Theta(n)$ then $\lambda = \Theta(n)$ and Theorem 2.4 gives a large improvement upon Hadamard's inequality. However, if, for example, k is fixed then λ is tending to zero and this gives a more modest improvement. We note that if $k \leq \sqrt{n}$ then $\lambda < 1$ and so λ is not an integer. Therefore, we may hope to improve Theorem 2.4 for matrices that are sufficiently sparse. In particular, for fixed k we may seek upper bounds for the quantities

$$\limsup_{n\to\infty} M(n, k)^{1/n} \leq \limsup_{n\to\infty} M_R(n, k)^{1/n} \leq \limsup_{n\to\infty} M_T(n, k)^{1/n}.$$

Hadamard's inequality gives the upper bound $\sqrt{k}$ for each, and, since $\lambda \to 0$ as $n \to \infty$, Ryser's result gives the same. Our main result is that for $k = o(n^{1/3})$ we can improve the bound given in Theorem 2.4. We show that for $k \geq 2$, there exists $c_k < \sqrt{k}$ depending only on k such that $M_T(n, k) \leq c_k^n$. Thus for k fixed we give an exponential improvement to the bound given by Hadamard's inequality. The existence of such a $c_k < \sqrt{k}$ was only known for $k = 2$ [BR18]. More details for the case $k = 2$ can be found in Section 2.2. Furthermore, if we restrict to studying $M_R(n, k)$ which the majority of this chapter considers, we can give further improvements and, in particular, for $k < \sqrt{n/10}$ we can improve the bound given in Theorem 2.4.

We generalize the notions $R(n,k)$ and $M_R(n,k)$ to non-square matrices.

Definition 2.5. *Let $R(m,n,k)$ be the set of $m \times n$ zero-one matrices whose rows sum to k.*

Definition 2.6. *For any $m \times n$ real matrix, A, where $m \leq n$, let $\mathrm{Vol}(A) = \sqrt{\det(AA^T)}$.*

The matrix AA^T is called the Gram matrix of A and the quantity $\det(AA^T)$ is known as the Gram determinant. See for example [HJ13]. If $m = n$ we of course have $\mathrm{Vol}(A) = |\det(A)|$. For any $m \times n$ real matrix, A, with $m \leq n$, $\mathrm{Vol}(A)$ is the volume of the parallelepiped formed by the rows of A. Gram's inequality tells us that $\det(AA^T) \geq 0$ with equality if and only if the rows of A are linearly dependent in which case we consider the parallelepiped to be degenerate which is consistent with zero volume.

Definition 2.7. *Let $M_R(m,n,k) = \max_{A \in R(m,n,k)} \mathrm{Vol}(A)$.*

We will repeatedly use the following generalization of Hadamard's inequality. Let A be an $m \times n$ real matrix. If A is partitioned into two horizontal blocks A_1 and A_2 with dimensions $m_1 \times n$ and $m_2 \times n$ respectively (thus $m_1 + m_2 = m$) then we have the inequality

$$\mathrm{Vol}(A) \leq \mathrm{Vol}(A_1)\,\mathrm{Vol}(A_2). \tag{2.1}$$

This follows, for example, by Fischer's inequality applied to the Gram matrix

$$AA^T = \begin{pmatrix} A_1A_1^T & A_1A_2^T \\ A_2A_1^T & A_2A_2^T \end{pmatrix}.$$

In developing bounds for $M_R(n,k)$ we show more general bounds for $M_R(m,n,k)$. Our basic approach stems from the following. If $A \in R(n,k)$ then it contains kn ones and therefore the columns have average sum k. Thus there exists a collection of at least k rows that share a column of ones. Thus there exists an $k \times n$ submatrix, A_1, of A that contains a column of ones. Due to the presence of a column of ones, the rows of A_1 are pairwise non-orthogonal. Intuitively, this suggests that the volume of the parallelepiped spanned by those rows is noticeably smaller than what is implied by

Hadamard's inequality. We bound $\mathrm{Vol}(A_1)$ and consider the remaining rows of A. Since the row sums are constant the remaining rows form a matrix in $R(n-k,n,k)$. We can compute the column averages and iterate this process to give an improved bound.

This chapter is organized as follows. In Section 2.2, we give background on the special case $k=2$ where $M_R(n,k)$ is known up to a constant factor and is exponentially smaller than $2^{n/2}$. We also give lower bounds for $M(n,k)$. In Section 2.3, we give an upper bound for $M_R(n,k)$ given by taking the rows in pairs. In Section 2.4, we improve this bound by taking the rows in sets of size $q \leq k$. In Section 2.5, we give, for small k, our best bound for $M_R(n,k)$ by greedily selecting the rows for removal. In Section 2.6, we establish some determinant inequalities we will need repeatedly. We use these to prove a generalization of Ryser's theorem for matrices in $R(m,n,k)$. We also give a counterexample to a conjecture of Li, Lin and Rodman [LLR99]. In Section 2.7, we show that the bound found in Section 2.3 applies to $M_T(n,k)$ for integral k thus answering a question of Bruhn and Rautenbach [BR18]. Furthermore, we generalize this to $M_T(n,\tilde{k})$ where $\tilde{k}$ is a rational number. In Section 2.8, we show that these techniques give upper bounds for perturbations of matrices in $R(n,k)$. We conclude with some open questions. Several of the messier calculations are deferred to Section 2.10.

2.2 Special case $k=2$ and lower bounds for $M_R(n,k)$

In [BR18] Bruhn and Rautenbach study zero-one matrices with at most $2n$ ones. To begin, they prove the following upper bound for $M_R(n,2)$.

Theorem 2.8. *If A is an $n \times n$ zero-one matrix, and each row of A contains at most two ones then $|\det(A)| \leq 2^{n/3}$.* □

Thus, in particular $M_R(n,2) \leq 2^{n/3}$. This gives an exponential improvement to the bound given by Theorem 2.4. This can be seen to be tight up to a constant factor from the following result found in [FvdD97].

Theorem 2.9. *$M(4,2)=2$. For $n \neq 4$, if $n=3\ell$ or $n=3\ell+2$ then $M(n,2)=2^{\ell}$. If $n=3\ell+1$ then $M(n,2)=2^{\ell-1}$.* □

Furthermore, the following bound for $M_T(n,2)$ is found in [BR18].

Theorem 2.10. $M_T(n,2) \leq 6^{n/6} \approx 1.348^n$. □

This gives an exponential improvement over the bound $M_T(n,2) \leq 2^{n/2}$ given by Hadamard's inequality. Bruhn and Rautenbach ask if a similar exponential improvement holds for matrices with $3n$ ones. We answer this question in the affirmative in Section 2.7.

Curiously, even for $k = 2$ we need not have $M(n,k) = M_R(n,k)$. From Theorem 2.9 we see that $M(7,2) = 2$. However, $M_R(7,2) = 4$. For example, if

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

then $\det(A) = 4$. Notice that the rows of A do indeed sum to 2 however not all columns have sum 2. Thus $A \notin S(7,2)$. So we pose the following question. For which values of n, k is $M(n,k) = M_R(n,k)$? Similarly, when do we have $M_R(n,k) = M_T(n,k)$? We know from Theorem 2.4 that equality holds when $\lambda = k(k-1)/(n-1)$ and there is an (n,k,λ) combinatorial design.

Next we discuss lower bounds for $M(n,k)$ (and thus $M_R(n,k)$ and $M_T(n,k)$). The theorem below follows from basic facts about projective planes which can be found for example in [BJL85].

Theorem 2.11. *Let $\varepsilon > 0$. For a prime power q, let $k = q + 1$. Then,*

$$\limsup_{n\to\infty} M(n,k)^{1/n} \geq \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2-\varepsilon}}\right).$$

Proof. Let $k = q + 1$ as given and let $n = q^2 + q + 1$. Then there exists a projective plane of order q. The incidence matrix, A, of this projective plane is an $n \times n$ matrix with row and column sums equal to k. Thus, $A \in S(n,k)$. This is a case where Ryser's

theorem is tight. We have $\lambda = 1$ and $\det(A) = M(n,k) = k(k-1)^{(n-1)/2}$. Now for any positive integer t let $N = tn$ and form $A^{(t)}$ as the block diagonal matrix with t copies of A along the diagonal. Then $A^{(t)} \in S(N,k)$ and we have

$$\begin{aligned}\det(A^{(t)}) &= \det(A)^t \\ &= k^t(k-1)^{t(n-1)/2} \\ &= k^{N/n}(k-1)^{(N-N/n)/2} \\ &= \left(k^{\frac{1}{k^2-k+1}}(k-1)^{\frac{1}{2}-\frac{1}{2(k^2-k+1)}}\right)^N.\end{aligned}$$

Thus,

$$\limsup_{n\to\infty} M(n,k)^{1/n} \geq k^{\frac{1}{k^2-k+1}}(k-1)^{\frac{1}{2}-\frac{1}{2(k^2-k+1)}} := f(k).$$

We have

$$\begin{aligned}\log(f(k)) &= \frac{\log k}{k^2-k+1} + \left(\frac{1}{2} - \frac{1}{k^2-k+1}\right)\log(k-1) \\ &= \frac{1}{2}\log(k-1) + O\left(\frac{1}{k^{2-\varepsilon}}\right).\end{aligned}$$

Note that $\sqrt{1-x} = 1 - x/2 + O(x^2)$ and thus

$$\sqrt{k-1} = \sqrt{k}\sqrt{1-\frac{1}{k}} = \sqrt{k}\left(1 - \frac{1}{2k} + O\left(\frac{1}{k^2}\right)\right) = \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^2}\right). \quad (2.2)$$

Exponentiating $\log(f(k))$ and using equation (2.2) we have

$$\begin{aligned}f(k) &= \sqrt{k-1}e^{O(1/k^{2-\varepsilon})} \\ &= \left(\sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^2}\right)\right)\left(1 + O\left(\frac{1}{k^{2-\varepsilon}}\right)\right) \\ &= \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2-\varepsilon}}\right)\end{aligned}$$

as desired. □

As a consequence of Theorem 2.11, we cannot hope to find a general upper bound for $M_R(n,k)$ of the form $c_k^{n/2}$ with $\sqrt{k} - c_k = \omega(1/\sqrt{k})$. We have, for example, if $k = 3$ then the construction via the Fano plane gives $\limsup_{n\to\infty} M(n,3)^{1/n} \geq 24^{1/7} \approx 1.5746$. If $p = 151$ which is prime, then let $k = 152$ and A the incidence matrix of the projective plane of order p gives the lower bound $\limsup_{n\to\infty} M(n,152)^{1/n} \geq \det(A)^{1/n} \approx 12.28955$. In this case $\sqrt{k} - \frac{1}{2\sqrt{k}} \approx 12.28827$.

2.3 Taking rows in pairs

The goal of this section is to prove the following theorem.

Theorem 2.12. *For all positive integers $m \leq n$ and $k \leq n$,*

$$M_R(m,n,k) \leq \left(\sqrt{k^2-1}\right)^{\frac{m}{2}-\frac{n}{2k}} k^{\frac{n}{2k}}.$$

If $m = n$ let $c_k = \left(\sqrt{k^2-1}\right)^{\frac{1}{2}\left(1-\frac{1}{k}\right)} k^{\frac{1}{2k}}$. Then $M_R(n,k) \leq c_k^n$. Note that $c_k < \sqrt{k}$.

Suppose that $A \in R(m,n,k)$ and there are two rows r and s that overlap in a ones, i.e. $\langle r, s\rangle = a$ where $\langle \cdot, \cdot \rangle$ is the dot product. Then if we let A_1 be the $2 \times n$ matrix formed by these rows we have

$$A_1 A_1^T = \begin{pmatrix} k & a \\ a & k \end{pmatrix}$$

and thus

$$\mathrm{Vol}(A_1) = \sqrt{k^2 - a^2} \leq \sqrt{k^2-1}$$

which improves on just using Hadamard's inequality for these rows. Hadamard's inequality tells us that $M_R(m,n,k) \leq k^{m/2}$. We now use these ideas to prove Theorem 2.12.

Proof of Theorem 2.12. Any $A \in R(m,n,k)$ contains mk ones. If $mk > n$ then by the pigeon hole principle there is a column with at least two ones. Thus there exist rows r and s such that $\langle r, s\rangle \geq 1$. Let A_1 be the $2 \times n$ submatrix consisting of rows r and s and let A_2 be the submatrix consisting of the remaining $m-2$ rows. Then $\mathrm{Vol}(A_1) \leq \sqrt{k^2-1}$. Note that $A_2 \in R(m-2,n,k)$ and thus by equation (2.1), $M_R(m,n,k) \leq \sqrt{k^2-1} M_R(m-2,n,k)$. Iterating this procedure t times we have

$$M_R(m,n,k) \leq \left(\sqrt{k^2-1}\right)^t M_R(m-2t,n,k)$$

with the process halting once $(m-2t)k \leq n$. Thus $m - 2t \leq n/k$. So $M_R(m-2t,n,k) \leq k^{\frac{n}{2k}}$ by Hadamard's inequality. Further $t \geq \dfrac{m}{2} - \dfrac{n}{2k}$ so we obtain

$$M_R(m,n,k) \leq \left(\sqrt{k^2-1}\right)^{\frac{m}{2}-\frac{n}{2k}} \sqrt{k}^{n/k}$$

as desired. Substituting $m = n$ gives the bound for $M_R(n,k)$. □

Recall that in Theorem 2.11 we showed the lower bound

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \geq \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2-\varepsilon}}\right).$$

Noting that $(1-x^2)^{1/4} = 1-x^2/4+O(x^4)$ it is straightforward to see that Theorem 2.12 improves the upper bound for $\limsup_{n\to\infty} M_R(n,k)^{1/n}$ from $\sqrt{k}$ to

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \leq \sqrt{k} - \frac{1}{4k^{3/2}} + O\left(\frac{1}{k^3}\right)$$

which leaves a sizable gap. We will close this gap to one of the form

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \leq \sqrt{k} - \Theta\left(\frac{1}{\sqrt{k}}\right)$$

with an explicit constant in Section 2.4.

If k is not fixed with respect to n then Theorem 2.12 does not always give a better bound for $M(n,k)$ than Ryser's theorem. However for k small it does. This is summarized in Theorem 2.13.

Theorem 2.13. *Let c_k be defined as in Theorem 2.12 and $\lambda = k(k-1)/(n-1)$ as in Theorem 2.4. If $k = o(n^{1/3})$ then for n sufficiently large, $c_k^n < k(k-\lambda)^{(n-1)/2}$.*

The proof of Theorem 2.13 is straightforward, but tedious. It can be found in Section 2.10. We just sketch the heuristics here. The growth of c_k^n is, roughly, $\sqrt{k^2-1}^{n/2}$. Ryser's bound is, roughly, $(k-\lambda)^{n/2}$. Since $\sqrt{k^2-1} < k - \frac{1}{2k}$ the result is achieved provided $k - \frac{1}{2k} < k - \lambda$ and thus $\frac{1}{2k} > \lambda = k(k-1)/n$ which holds when $k = o(n^{1/3})$.

Example, $k = 3$

Let $k = 3$ and $n = 1000$. We give three bounds for $M_R(n,k)$.

1. Using Hadamard's inequality $M_R(n,k) \leq k^{n/2} = 3^{500} \approx 3.64 \times 10^{238}$.
2. Ryser's result has $\lambda = 2/333 = 0.\overline{006}$ and gives the bound $M(n,k) \leq 3(3-\lambda)^{\frac{1000-1}{2}} = 3(2.99399\ldots)^{499.5} \approx 2.31 \times 10^{238}$.
3. Theorem 2.12 gives the bound c_k^n where $c_k \approx 1.6984 < \sqrt{3}$ and thus $M_R(n,k) \leq c_k^n \approx 1.08 \times 10^{230}$.

2.4 Taking rows in sets of size q

In this section we generalize our approach in Section 2.3 to removing from $A \in R(m,n,k)$ rows in sets of size q. If we have q rows that as a submatrix have a column of ones, that is they share a one in a single common coordinate, then their Gram matrix will have elements k on the diagonal and elements greater than or equal to one off the diagonal. Thus we have the following definition.

Definition 2.14. *Let $S_{n,a,k}$ be the $n \times n$ matrix with diagonal elements equal to k and off-diagonal elements equal to a. If I_n is the $n \times n$ identity matrix and J_n is the $n \times n$ all ones matrix we can write $S_{n,a,k} = aJ_n + (k-a)I_n$.*

Notice that if A is the incidence matrix of an (n,k,λ) combinatorial design then $AA^T = S_{n,\lambda,k}$. We will make use of the following lemma which will be proved in Section 2.6.

Lemma 2.15. *We have* $\det(S_{n,a,k}) = (a(n-1)+k)(k-a)^{n-1}$ *and $S_{n,a,k}$ is positive definite if $a < k$. Further, for any positive definite $n \times n$ matrix A such that A has diagonal elements k and $A \geq S_{n,a,k}$ coordinatewise we have* $\det(A) \leq \det(S_{n,a,k})$.

In particular, we will make use of the special case of Lemma 2.15 that $\det(S_{q,1,k}) = (q+k-1)(k-1)^{q-1}$ which has maximum determinant over all $q \times q$ positive definite matrices with diagonal elements k and non-diagonal elements at least one. This generalizes the trivial fact, used in Section 2.3, that if $A = \begin{pmatrix} k & a \\ a & k \end{pmatrix}$ with $a \geq 1$ then $\det(A) \leq k^2 - 1$.

Theorem 2.16. *Let q be an integer with $1 \leq q \leq k$. We have,*

$$M_R(m,n,k) \leq \left(\sqrt{(q+k-1)(k-1)^{q-1}}\right)^{\frac{m}{q} - \frac{n}{k}\frac{q-1}{q}} k^{\frac{n(q-1)}{2k}}.$$

If $m = n$, let

$$c_{q,k} = (q+k-1)^{\frac{1}{2q}\left(1-\frac{q-1}{k}\right)}(k-1)^{\frac{1}{2}\frac{q-1}{q}\left(1-\frac{q-1}{k}\right)}k^{\frac{(q-1)}{2k}}. \tag{2.3}$$

Then $M_R(n,k) \leq c_{q,k}^n$.

Proof. Suppose we have $A \in R(m,n,k)$. The number of ones in A is mk. The average number of ones in a column is mk/n. So if $mk/n > q-1$ then there is some column containing at least q ones. Let R_q be an arbitrary submatrix formed by taking q rows that have a column of ones. Then $R_q R_q^T \geq S_{q,1,k}$ coordinatewise with equality if all other column sums of R_q are 0 or 1. Thus Lemma 2.15 tells us that $\mathrm{Vol}(R_q) \leq \sqrt{(q+k-1)(k-1)^{q-1}}$. We remove these rows and iterate t times. So we have

$$M_R(m,n,k) \leq \left(\sqrt{(q+k-1)(k-1)^{q-1}}\right)^t M_R(m-qt,n,k)$$

where t must satisfy $(m-qt)k/n > q-1$. Thus $m-qt > \frac{n}{k}(q-1)$ and $t < \frac{m}{q} - \frac{n}{k}\frac{q-1}{q}$. Thus we have

$$M_R(m,n,k) \leq \left(\sqrt{(q+k-1)(k-1)^{q-1}}\right)^{\frac{m}{q}-\frac{n}{k}\frac{q-1}{q}} k^{\frac{n(q-1)}{2k}}.$$

If we let $m=n$, then we have

$$\begin{aligned} M_R(m,n,k) &\leq \left((q+k-1)(k-1)^{q-1}\right)^{\frac{n}{2q}\left(1-\frac{q-1}{k}\right)} k^{\frac{n(q-1)}{2k}} \\ &= (q+k-1)^{\frac{n}{2q}\left(1-\frac{q-1}{k}\right)}(k-1)^{\frac{n}{2}\frac{q-1}{q}\left(1-\frac{q-1}{k}\right)} k^{\frac{n(q-1)}{2k}} \\ &= c_{q,k}^n \end{aligned}$$

with $c_{q,k}$ as defined in equation (2.3). □

Notice that c_k as defined in Theorem 2.12 is equivalent to $c_{2,k}$. In Theorem 2.32 in Section 2.10 we show that, for large k, $c_{q,k}$ is minimized when $q = sk$ where $s \approx 0.44$. For example, when $k = 49$, we computed $c_{q,k}$ for $q = 1, 2, \ldots, k$. In this case $c_{1,k} = \sqrt{k} = 7$. To visualize we plotted q versus $\sqrt{k} - c_{q,k}$. The peak of this graph tells us the optimal choice of q. See Figure 2.1. In this case, the optimal choice of q is $q_* = \mathrm{argmin}_q\, c_{q,k} = 23$ and we have $q_*/k \approx 0.47$. We can calculate $c_{23,49} \approx 6.9931$. The plot shows that, in terms of a discrepancy from $\sqrt{k}$, using $q = 23$ versus the simpler approach using $q = 2$ outlined in Section 2.3 gives substantial improvement. Furthermore, we will show in Theorem 2.33 that for a real number $t \approx 0.096$,

$$c_{sk,k} = \sqrt{k} - \frac{t}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2}}\right). \tag{2.4}$$

Recall in Theorem 2.11 we showed that for $\varepsilon > 0$, and for p^ℓ a prime power, if $k = p^\ell + 1$ then

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \geq \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2-\varepsilon}}\right).$$

So there is a gap to be resolved.

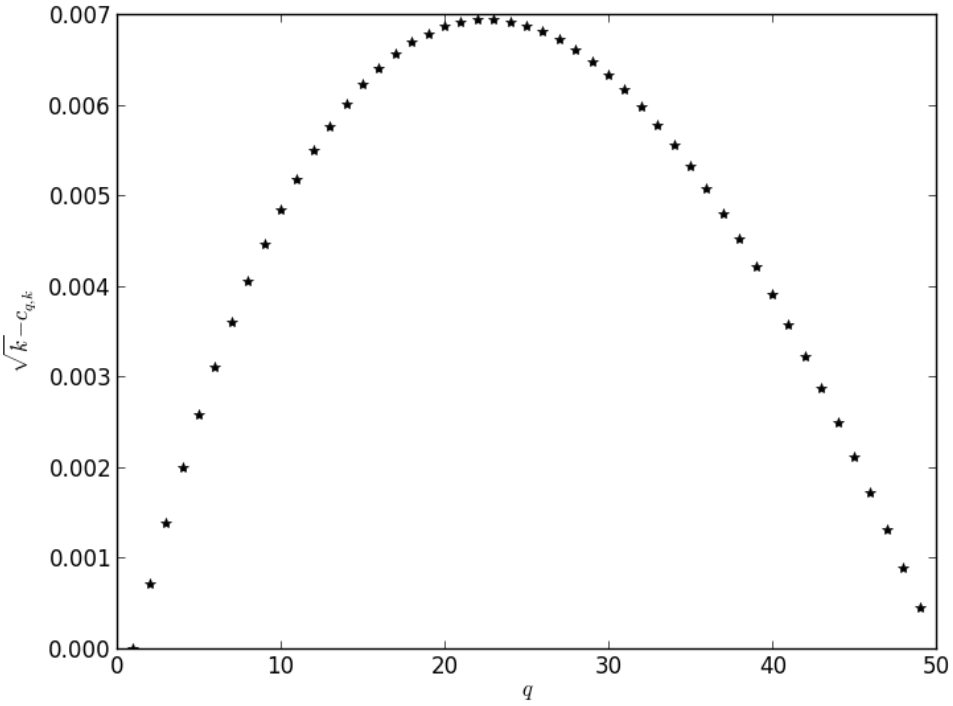


Figure 2.1: q versus $\sqrt{k} - c_{q,k}$ for $k = 49$. The peak is at $(23, 6.9931)$.

To visualize how quickly $c_{q_*,k}$ approaches $\sqrt{k} - \frac{t}{2\sqrt{k}}$ we plotted k versus $c_{q_*,k}$ for $k = 3, \ldots, 20$. See Figure 2.2. We overlay the curve $\frac{t}{2\sqrt{k}}$ to illustrate how quickly they converge.

Example, $k = 17$

From Theorem 2.16 we have $M_R(n, 17) \leq c_{q,17}^n$. We give the following progressively better upper bounds for $\limsup_{n\to\infty} M_R(n, 17)^{1/n}$.

1. Hadamard's inequality gives $c_{1,17} = \sqrt{17} \approx 4.1241$.
2. Using $q = 2$ rows at a time we have $c_{2,17} \approx 4.1197$.
3. For $q \in [17]$, the minimum $c_{q,17}$ occurs when $q = 8$. We have $c_{8,17} \approx 4.1111$.

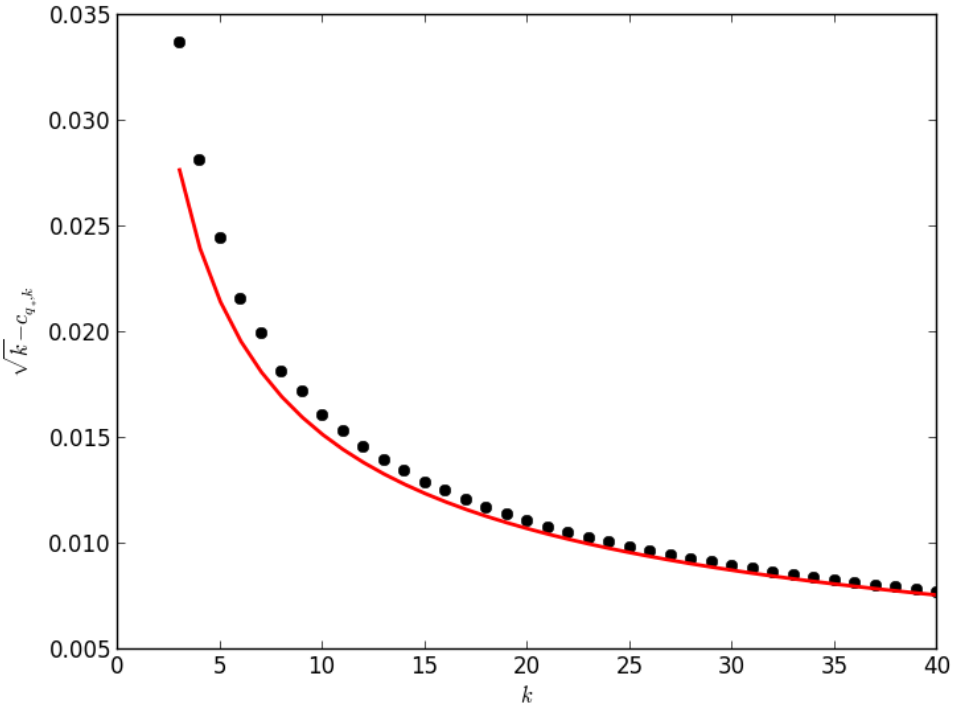


Figure 2.2: k versus $\sqrt{k} - c_{q_*,k}$ for $k = 3, \ldots, 20$. We draw in red the curve $\frac{t}{2\sqrt{k}}$ where $t \approx 0.096$ as given in Theorem 2.33.

In Section 2.5, we show that we can further improve our bound on $\limsup_{n\to\infty} M_R(n, 17)^{1/n}$.

2.5 Greedily selecting rows for removal

In Sections 2.3 and 2.4 we chose a value q and, for as many iterations as possible, removed rows in sets of size q. Then we used Hadamard's inequality to bound the remaining rows. In this section we vary the number of rows removed at a given iteration by greedily selecting as many as possible so as to assure a column of ones in each removal. The main result of this section is Theorem 2.17 below. As in the previous sections, we show this bound for $M_R(n,k)$ by establishing a more general bound for $M_R(m,n,k)$. This more general bound is given in Theorem 2.18. For constant k, the bound in Theorem 2.17 is asymptotically better than that in Theorem 2.12 and one can numerically check is better than Theorem 2.16 for $k \leq 27$. See Theorem 2.35 in Section 2.10. Experimentally, greedily selecting the rows gives a better bound than Theorem 2.16 for all k, however, they are quite close and due to the uncertainty in our estimates Theorem 2.17 does not give a better bound for all k. Perhaps a tighter analysis will demonstrate the superiority of this approach for all k.

Theorem 2.17. *Let*

$$\alpha_k = \sqrt{\frac{(2k-1)!}{(k-1)!}}(k-1)^{\frac{1}{4}(k^2-k)}$$

and

$$\beta_k = \left(k + \frac{k}{H_k} - 1\right)^{\frac{1}{2}(H_k/k)} (k-1)^{\frac{1}{2}(1-H_k/k)}$$

where $H_j = \sum_{i=1}^{j} 1/i$ is the j-th harmonic number. Then $M_R(n,k) \leq \alpha_k \beta_k^n$.

Suppose we have $A \in R(m,n,k)$. The number of ones in A is mk. Thus the column averages are mk/n. Thus if we let $r = \lceil mk/n \rceil$ we can find r rows that share a column of ones and thus by Lemma 2.15 their volume is at most $\sqrt{\det(S_{r,1,k})} = (r+k-1)^{1/2}(k-1)^{(r-1)/2}$. Recursively, we will then use the bound

$$M_R(m,n,k) \leq (r+k-1)^{1/2}(k-1)^{(r-1)/2} M_R(m-r,n,k).$$

We will begin by removing r rows but, as the number of rows in A diminishes, the number of rows we can remove at each iteration will ultimately diminish to one in which case we are using Hadamard's inequality. For example, if $m = 100$, $n = 200$ and

$k = 17$ we will begin by removing $\lceil 100 \cdot 17/200 \rceil = 9$ rows. We now have a matrix with $100 - 9 = 91$ rows and next we greedily remove $\lceil (100 - 9) \cdot 17/200 \rceil = 8$ rows. The sequence of removals, Q, in this case is

$$Q = (9, 8, 8, 7, 6, 6, 5, 5, 4, 4, 4, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1).$$

Let a_i be the number of times i appears in Q. In the above example $a_9 = 1$ and $a_8 = 2$. Let $m_r = m$ and for $i < r$ let m_i be the number of rows remaining just prior to removing a_i sets of i rows. Thus $m_0 = 0$. As above we have

$$r = \left\lceil \frac{mk}{n} \right\rceil.$$

For $i = 1, \ldots, r$ we have

$$m_{i-1} = m_i - ia_i.$$

For $i \leq r$ if we have m_{i-1} rows we just removed ia_i rows. Thus the column average is at most $i - 1$. However, if we had $m_{i-1} + i$ rows then the column average must have exceeded $i - 1$ as we were able to remove i rows. Thus we have

$$\frac{m_{i-1}k}{n} \leq i - 1 < \frac{(m_{i-1} + i)k}{n}.$$

Rearranging, we have

$$\frac{(i-1)n}{k} - i < m_{i-1} \leq \frac{(i-1)n}{k} \tag{2.5}$$

We stress that a similar bound need not hold for $m_r = m$ as this does not arise from just having removed sets of $r + 1$ rows. However, we will note momentarily that the bound does hold for m_r when $m = n$. For $2 \leq i \leq r$ we have

$$a_{i-1} = \frac{m_{i-1} - m_{i-2}}{i - 1}. \tag{2.6}$$

Taking the upper bound for m_{i-1} minus the lower bound for m_{i-2} from equation (2.5) and substituting into equation (2.6) gives an upper bound for a_{i-1}. Similarly we subtract from the lower bound for m_{i-1} the upper bound for m_{i-2} to get a lower bound for a_{i-1}. We obtain

$$\frac{n}{k(i-1)} - \frac{i}{i-1} < a_{i-1} < \frac{n}{k(i-1)} + 1. \tag{2.7}$$

So we see that for $i < r$, the approximation $a_i \approx \frac{n}{ki}$ is quite good. Finally, we seek a bound for a_r. We have

$$\begin{aligned} a_r &= \frac{m - m_{r-1}}{r} \\ &< \frac{m - \left(\frac{(r-1)n}{k} - r\right)}{r} \\ &= \frac{1}{r}\left(m + \frac{n}{k}\right) - \frac{n}{k} + 1 \\ &\leq \frac{1}{mk/n}\left(m + \frac{n}{k}\right) - \frac{n}{k} + 1 \\ &= \frac{n^2}{k^2 m} + 1 \end{aligned}$$

We note that if $n|mk$, for example when $n = m$ then this approximation is quite precise since $r = mk/n$. In the case $m = n$, we have $r = k$ and $a_k \leq n/k^2 + 1$ which is consistent with equation (2.7).

Now that we have bounded a_i for $i = 1, \ldots, r$ we can give an upper bound for $M_R(m, n, k)$. We have

$$\begin{aligned} M_R(m,n,k) &\leq \prod_{i=1}^{r} \left(\sqrt{(i+k-1)(k-1)^{i-1}}\right)^{a_i} \\ &= \left(\prod_{i=1}^{r-1} \left((i+k-1)(k-1)^{i-1}\right)^{a_i/2}\right) \left((r+k-1)(k-1)^{r-1}\right)^{a_r/2} \\ &= \left(\prod_{i=1}^{r-1} (i+k-1)^{a_i/2}\right) \left((k-1)^{\frac{1}{2}\sum_{i=1}^{r-1}(i-1)a_i}\right) \left((r+k-1)(k-1)^{r-1}\right)^{\frac{1}{2}a_r} \\ &\leq X_{r-1} \cdot Y_{r-1} \cdot Z_r \end{aligned} \tag{2.8}$$

where

$$X_r = \prod_{i=1}^{r} (i+k-1)^{\frac{1}{2}\left(\frac{n}{ki}+1\right)} \tag{2.9}$$

$$Y_r = (k-1)^{\frac{1}{2}\sum_{i=1}^{r}(i-1)a_i} \tag{2.10}$$

$$Z_r = \left((r+k-1)(k-1)^{r-1}\right)^{\frac{1}{2}\left(\frac{n^2}{k^2 m}+1\right)} \tag{2.11}$$

Note that in the case $m = n$, we have $r = k$ and the estimate $a_k \leq \frac{n}{k^2} + 1$ agrees with the bound $a_i \leq \frac{n}{ik} + 1$ and thus

$$M_R(n,k) \leq X_k Y_k. \tag{2.12}$$

We begin by bounding X_r.

$$\begin{aligned} X_r &= \prod_{i=1}^{r} (i+k-1)^{\frac{1}{2}\left(\frac{n}{ki}+1\right)} \\ &= \sqrt{\frac{(r+k-1)!}{(k-1)!}} \left(\prod_{i=1}^{r} (i+k-1)^{1/i} \right)^{\frac{n}{2k}}. \end{aligned}$$

Let $F(r,k) = \prod_{i=1}^{r} (i+k-1)^{1/i}$. Then $\log(F(r,k)) = \sum_{i=1}^{r} \frac{\log(i+k-1)}{i}$. Denote by $H_j = \sum_{i=1}^{j} 1/i$ the j-th harmonic number. Since log is a concave function we have, using Jensen's inequality,

$$\begin{aligned} \frac{\sum_{i=1}^{r} \frac{\log(i+k-1)}{i}}{\sum_{i=1}^{r} \frac{1}{i}} &\leq \log \left(\frac{\sum_{i=1}^{r} \frac{i+k-1}{i}}{\sum_{i=1}^{r} \frac{1}{i}} \right) \\ &= \log \left(\frac{r + (k-1)H_{r-1}}{H_r} \right) \\ &= \log \left(k + \frac{r}{H_r} - 1 \right) \end{aligned}$$

and therefore

$$\log(F(r,k)) \leq \log \left(k + \frac{r}{H_r} - 1 \right) H_r.$$

So

$$F(r,k) \leq \left(k + \frac{r}{H_r} - 1 \right)^{H_r}.$$

Finally, we see that

$$X_r \leq \sqrt{\frac{(r+k-1)!}{(k-1)!}} \left(k + \frac{r}{H_r} - 1 \right)^{\frac{nH_r}{2k}}.$$

Next, we study the second factor in equation (2.8). Let $T_r = \sum_{i=1}^{r} (i-1)\left(\frac{n}{ik}+1\right)$. Then $Y_r = (k-1)^{T_r/2}$. We have

$$\begin{aligned} T_r &= \sum_{i=1}^{r} \frac{n}{k} - 1 + i - \frac{n}{ik} \\ &= r\left(\frac{n}{k} - 1\right) + \frac{r(r+1)}{2} - \frac{n}{k} H_r \\ &= (r - H_r)\frac{n}{k} + \frac{1}{2}(r^2 - r). \end{aligned}$$

Thus,

$$Y_r = (k-1)^{\frac{n}{2k}(r-H_r)} (k-1)^{\frac{1}{4}(r^2-r)}.$$

If we substitute our bound for X_{r-1} and Y_{r-1} and Z_r into equation (2.8) we obtain the following theorem.

Theorem 2.18.

$$M_R(m,n,k) \leq \sqrt{\frac{(r+k-2)!}{(k-1)!}}(k-1)^{\frac{1}{4}(r^2-3r+2)} \times \left(k+\frac{r-1}{H_{r-1}}-1\right)^{\frac{nH_{r-1}}{2k}} (k-1)^{\frac{n}{2k}(r-H_{r-1}-1)} \left((r+k-1)(k-1)^{r-1}\right)^{\frac{1}{2}\left(\frac{n^2}{k^2m}+1\right)} \tag{2.13}$$

where we have arranged the terms that depend on r and k only in the first row and the terms that depend on n and m in the second.

If we have a square matrix, $m = n$, then equation (2.12) gives us

$$\begin{aligned} M_R(n,k) &\leq X_k Y_k \\ &\leq \sqrt{\frac{(2k-1)!}{(k-1)!}}\left(k+\frac{k}{H_k}-1\right)^{\frac{nH_k}{2k}} (k-1)^{\frac{n}{2k}(k-H_k)}(k-1)^{\frac{1}{4}(k^2-k)} \\ &= \sqrt{\frac{(2k-1)!}{(k-1)!}}(k-1)^{\frac{1}{4}(k^2-k)}\left(\left(k+\frac{k}{H_k}-1\right)^{\frac{H_k}{2k}}(k-1)^{\frac{1}{2k}(k-H_k)}\right)^n \end{aligned}$$

establishing Theorem 2.17 above.

Examples, $k = 3$ and $k = 17$

For $k = 3$ we have the following,

1. In Section 2.3 we saw $c_{2,3} = 1.6984$. So $M_R(n,3) \leq 1.6984^n$.
2. Theorem 2.17 tells us that $\alpha_3 \approx 21.91$ and $\beta_3 = (40/11)^{11/36}2^{7/36} \approx 1.6977$ and $M_R(n,3) \leq 21.91 \times 1.6977^n$. In this case the strategy is, roughly, to use $n/9$ sets of three rows, $n/6$ sets of two rows, and apply Hadamard's inequality to the remaining $n/3$ rows.

For $k = 17$ we have the following progressively better upper bounds. These are visualized in Figure 2.3.

1. $\limsup_{n\to\infty} M_R(n,17) \leq c_{2,17} \approx 4.1197$.
2. $\limsup_{n\to\infty} M_R(n,17) \leq c_{8,17} \approx 4.1111$.

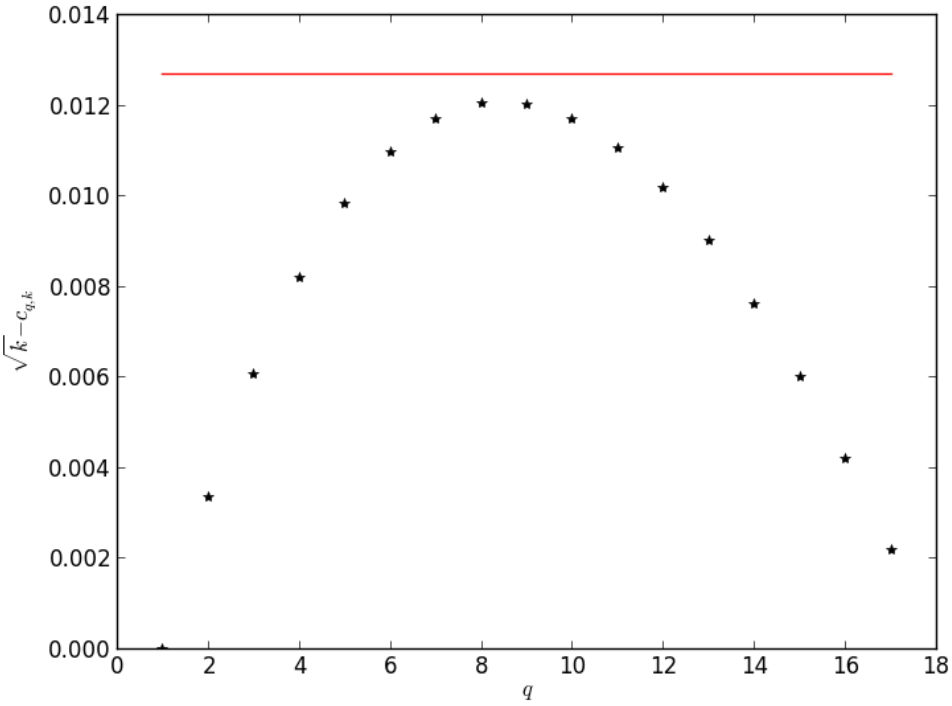


Figure 2.3: q versus $\sqrt{k} - c_{q,k}$ for $k = 17$. We draw a red line at height $\sqrt{k} - \beta_k$ to show that, for $k = 17$, the greedy approach gives a better bound.

3. Using Theorem 2.17 we have $\limsup_{n\to\infty} M_R(n, 17) \leq \beta_{17} \approx 4.1104$.

We note that for general k our bound for α_k is quite large. Due to the uncertainty of the a_i, the product computed in equation (2.9) multiplies this uncertainty k times. Our goal was to minimize β_k and as we were interested in the case where k is constant. However, for any given n we can compute a practical bound. For example, if $k = 17$ as above and $n = 1000$ then we have $M_R(1000, 17) \leq c_{8,17}^{1000} \approx 9.0074 \times 10^{613}$. If we were to just use the bound $M_R(1000, 17) \leq \alpha_{17}\beta_{17}^{1000}$ we would obtain $M_R(1000, 17) \leq 3.7674 \times 10^{707}$ which is a worse bound. However, we can in this case exactly compute the a_i. These counts can be found in Table 2.1. They give the improved bound $M_R(1000, 17) \leq 9.3551 \times 10^{612}$.

q	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
a_q	4	4	3	5	4	5	5	6	7	7	8	10	12	14	20	29	57

Table 2.1: Counts for greedy row removal for $k = 17$ and $n = 1000$.

2.5.1 Dynamic programming

If $A \in R(m,n,k)$ then as observed above we can find up to $q_{\max} = \lceil mk/n \rceil$ rows to bound and remove. This suggests a recursive approach where the optimal $q \in [q_{\max}]$ is chosen and the program recurs on $M_R(m-q,n,k)$. Using memoization this is easily implemented. To avoid overflow we minimize a bound for $\log(M_R(m,n,k))$. Our intuition is that we should choose $q_{\max}$ at each iteration. We implemented this algorithm using Julia [BKSE12] code. For all $3 \le k \le m \le n \le 100$ the dynamic algorithm selected rows greedily. For $m = n = 1000$ and $k = 17$ the dynamic algorithm gave the same counts as in Table 2.1.

Memoized dynamic algorithm to bound $\log(M_R(m,n,k))$

```
1:  procedure BOUND(m,n,k)
2:      if m = 0 then
3:          return 0
4:      bestbound = ∞
5:      q_max = ⌈mk/n⌉
6:      for q = 1, ..., q_max do
7:          b = (1/2) log(q + k − a) + (1/2)(q − a) log(k − 1) + BOUND(m − q, n, k)
8:          if b < bestbound then
9:              bestbound = b
10:     return bestbound
```

2.6 A generalization of Ryser's theorem for matrices in $R(m,n,k)$

In this section we state and establish some facts about the determinants of positive definite matrices. We will use these to prove a generalization of Ryser's theorem for matrices in $R(m,n,k)$. We begin with a little background on the notion of majorization which will prove useful. These facts can be found in [MOA11].

If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ we let $x_{[i]}$ be the i-th largest component. Thus if we write x in non-increasing order we have $x_\downarrow = (x_{[1]}, \ldots, x_{[n]})$. Then we have the following definition.

Definition 2.19. *For $x, y \in \mathbb{R}^n$, we say that x is majorized by y and write $x \prec y$ if*

$$\sum_{i=1}^{k} x_{[i]} \le \sum_{i=1}^{k} y_{[i]}, \quad k = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}.$$

Intuitively, we say $x \prec y$ means that x is "less spread out" than y.

Definition 2.20. *A real-valued function ϕ on a set $\mathcal{A} \subset \mathbb{R}^n$ is said to be Schur-convex if whenever $x, y \in \mathcal{A}$ and $a \prec y$ we have $\phi(x) \leq \phi(y)$. We say ϕ is Schur-concave if $x \prec y$ implies $\phi(x) \geq \phi(y)$.*

Let $\mathcal{A} = \mathbb{R}^n_{\geq 0}$. That is $\mathcal{A}$ is the set of vectors in $\mathbb{R}^n$ with non-negative coordinates. If $f : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ is defined by $f(x) = \prod_{i=1}^n x_i$ then f is Schur-concave. This formalizes the observation that if the elements of x are "less spread out" than those of y, then their product is larger.

In [Olk14] the author proves the following:

Lemma 2.21. *Let A be an $n \times n$, positive definite matrix with diagonal elements $a_{i,i} = 1$. Let $\bar{a} = \frac{1}{n(n-1)} \sum_{i \neq j} a_{i,j}$ be the average of the off-diagonal elements. Let $\tilde{A}$ be the $n \times n$ matrix such that $\tilde{a}_{i,i} = 1$ and $\tilde{a}_{i,j} = \bar{a}$ for $i \neq j$. Then $\lambda(\tilde{A}) \prec \lambda(A)$ and thus* $\det(A) \leq \det(\tilde{A})$. □

Notice that, via a rescaling argument, the requirement $a_{i,i} = 1$ can be replaced by any constant on the diagonal. Recall that $S_{n,a,k}$ is the $n \times n$ matrix with diagonal elements k and off-diagonal elements a. We now restate and prove Lemma 2.15.

Lemma 2.15. *We have* $\det(S_{n,a,k}) = (a(n-1)+k)(k-a)^{n-1}$ *and $S_{n,a,k}$ is positive definite if $a < k$. Further, for any positive definite $n \times n$ matrix A such that A has diagonal elements k and $A \geq S_{n,a,k}$ we have* $\det(A) \leq \det(S_{n,a,k})$.

Proof. To see $\det(S_{n,a,k}) = (a(n-1)+k)(k-a)^{n-1}$ we find the eigenvalues. If u is the all ones vector, then $S_{n,a,k}u = (an + k - a)u$ thus $S_{n,a,k}$ has the eigenvalue $an + k - a = a(n-1) + k$. Further if v is in the codimension one subspace of vectors whose coordinates sum to zero then $S_{n,a,k}v = (k-a)v$ and thus $S_{n,a,k}$ has the eigenvalue $(k-a)$ with multiplicity $n-1$. Thus $\det(S_{n,a,k}) = (a(n-1)+k)(k-a)^{n-1}$. If $a < k$ all eigenvalues are positive.

Next, fix n, k and let $f(x) = \det(S_{n,x,k}) = (x(n-1)+k)(k-x)^{n-1}$. Then

$$\begin{aligned}\frac{d}{dx}f(x) &= (n-1)(k-x)^{n-1} - (x(n-1)+k)(n-1)(k-x)^{n-2} \\ &= (n-1)(k-x)^{n-2}\left[(k-x) - (x(n-1)+k)\right] \\ &= (n-1)(k-x)^{n-2}(-xn) \\ &< 0\end{aligned}$$

for all $x < k$. Thus $f(x)$ is a decreasing function for $x < k$. If $\bar{a}$ is the average of the off-diagonal elements of A then we have $\tilde{A} = S_{n,\bar{a},k}$ and $a \leq \bar{a}$. From Lemma 2.21 we have $\det(A) \leq \det(\tilde{A})$. Since $\det(S_{n,x,k})$ is decreasing we have $\det(\tilde{A}) \leq \det(S_{n,a,k})$. Combining these two inequalities gives the result. □

We use the above lemmas to prove, for $A \in R(m,n,k)$, the following generalization of Ryser's theorem (Theorem 2.4).

Theorem 2.22. *Let $A \in R(m,n,k)$. Let $\mu = \frac{k}{m-1}\left(\frac{mk}{n} - 1\right)$. Then*

$$\mathrm{Vol}(A) \leq k\sqrt{\frac{m}{n}}(k-\mu)^{\frac{m-1}{2}}. \tag{2.14}$$

Notice that if $m = n$ then $\mu = k(k-1)/(n-1) = \lambda$ and we recover Theorem 2.4.

Proof. Let $A \in R(m,n,k)$ and consider the Gram matrix, AA^T. We have $\mathrm{Vol}(A) = \sqrt{\det(AA^T)}$. The diagonal elements of AA^T are all k. Let b_j be the number of ones in column j of A. We have

$$\sum_{j=1}^{n} b_j = mk$$

If there are b_j ones in column j then the number of ordered pairs of distinct rows (r,s) that overlap in these ones is $2\binom{b_j}{2}$. So we have

$$\begin{aligned}\sum_{\substack{r,s\in\mathrm{rows}(A)\\ r\neq s}} \langle r,s\rangle &= \sum_{j=1}^{n} 2\binom{b_j}{2} \\ &= \sum_{j=1}^{n} b_j^2 - \sum_{j=1}^{n} b_j \\ &= \sum_{j=1}^{n} b_j^2 - mk\end{aligned}$$

The sum of the squares of the b_j is minimized when they are all equal. So we get the lower bound

$$\sum_{\substack{r,s\in\text{rows}(A)\\ r\neq s}} \langle r,s\rangle \geq n\left(\frac{mk}{n}\right)^2 - mk = mk\left(\frac{mk}{n}-1\right)$$

The average off-diagonal entry of AA^T can then be bounded.

$$\frac{1}{m(m-1)}\sum_{\substack{r,s\in\text{rows}(A)\\ r\neq s}} \langle r,s\rangle \geq \frac{1}{m(m-1)}\left(mk\left(\frac{mk}{n}-1\right)\right) = \frac{k}{m-1}\left(\frac{mk}{n}-1\right) = \mu.$$

Notice that if $m = n$ then $\mu = k(k-1)/(n-1)$ and thus $\mu = \lambda$ as in Theorem 2.4. Also, notice that this only gives useful information if $\mu > 0$ and thus $m > n/k$. This is not surprising as otherwise $mk < n$ and then we can arrange the rows orthogonally. Thus, Lemma 2.15 gives us

$$\begin{aligned}\det(A) &\leq \det(S_{m,\mu,k})\\ &= (\mu(m-1)+k)(k-\mu)^{m-1}\\ &= k^2\frac{m}{n}(k-\mu)^{m-1}\end{aligned}$$

Taking the square root gives equation (2.14). □

2.6.1 Counterexample to a conjecture of Li, Lin and Rodman

Conjecture 4.8 of [LLR99] states that if $\lambda = k(k-1)/(n-1)$ and $A \in S(n,k)$ is non-singular and the off-diagonal entries, x, of AA^T and A^TA satisfy $|x-\lambda| < 1$ then $|\det(A)| = M(n,k)$. We give the following counterexample. Let $n = 10$ and $k = 3$. In

this case $\lambda = 3 \cdot 2/9 = 2/3$. First observe that $M(10,3) \geq 48$ since if

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

then $B \in S(10,3)$ and $\det(B) = 48$. Next, let

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We see $A \in S(10,3)$ and $\det(A) = 15 < M(10,3)$. Further, we can check that the off-diagonal entries of AA^T and A^TA are exclusively 0 and 1 which of course satisfy $|x - 2/3| < 1$.

2.7 Matrices with kn ones

We generalize the class $T(n,k)$ to rectangular matrices.

Definition 2.23. *Let $T(m,n,k)$ be the set of $m \times n$ zero-one matrices containing km ones.*

Definition 2.24. *Let $M_T(m,n,k) = \max_{A \in T(m,n,k)} \mathrm{Vol}(A)$.*

Let $A \in T(m,n,k)$. We can bound $\mathrm{Vol}(A)$ by applying Theorem 2.12 to the submatrix formed by the rows summing to k and Hadamard's inequality to the rows not summing to k. To bound this latter product we develop a few lemmas.

Lemma 2.25. *Let s be an integer and set $r = 2s$. Let $S = \mathbb{Z} \setminus \{0\}$ be the non-zero integers. Let $\mathcal{A} = \{x \in S^r : \sum_{i=1}^r x_i = 0\}$. Let $x = (1, \ldots, 1, -1, \ldots, -1) \in \mathcal{A}$ be such that 1 and -1 each appears as a coordinate s times. Then for all $y \in \mathcal{A}$, we have $x \prec y$.*

Proof. Suppose for the sake of contradiction there is $y \in \mathcal{A}$ such that x is not majorized by y. Note that the cumulative sums of x are $(1, 2, \ldots, s, s-1, s-2, \ldots, 1, 0)$. Without loss of generality the coordinates of y are in descending order. Let $j < r$ be such that $\sum_{i=1}^j x_i > \sum_{i=1}^j y_i$. We have two cases:

1. $j \leq s$
2. $j > s$

In the first case we have that $\sum_{i=1}^j y_i < j$. Since the y_i are non-increasing non-zero integers, we must have $y_j < 0$ which implies $y_k \leq -1$ for all $k > j$. However, then the sum of the remaining coordinates is less than $-s$ and so $\sum_{i=1}^r y_i < 0$ which is a contradiction.

In the second case let $j' = r - j$. Then $\sum_{i=1}^j y_i < j'$ which implies that $y_j \leq -1$. But the sum of the remaining j' coordinates is at most $-j'$ and thus we obtain the same contradiction that $\sum_{i=1}^r y_i < 0$. □

It follows from Lemma 2.25 that for r even if

$$\mathcal{A} = \left\{ x \in \mathbb{Z}_+^r \ : \ \sum_{i=1}^r x_i = rk \text{ and } x_i \neq k, i \in [r] \right\} \tag{2.15}$$

then $x = (k+1, \ldots, k+1, k-1, \ldots, k-1)$ with $k+1$ and $k-1$ each appearing $s = r/2$ times is majorized by all y in $\mathcal{A}$ and thus $\prod_i x_i = (x+1)^{r/2}(x-1)^{r/2}$ is maximal on $\mathcal{A}$. Note that if r is odd and $\mathcal{A}$ is given by equation (2.15) then there need not be

a vector $x \in \mathcal{A}$ that is majorized by all $y \in \mathcal{A}$. For example, let $r = 5$ and $k = 3$. One can check that among the 15 non-negative integer vectors in $(\mathbb{Z} \setminus \{3\})^5$ summing to 15, the maximum product is achieved uniquely by $x = (5, 4, 2, 2, 2)$. However, if $y = (4, 4, 4, 2, 1)$ we see that x is not majorized by y. Indeed x has cumulative sums $(5, 9, 11, 13, 15)$ whereas y has cumulative sums $(4, 8, 12, 14, 15)$. However, we do have the following.

Lemma 2.26. *Let $r = 2s - 1$ be a positive odd integer and let k be a positive integer. Let $\mathcal{A}$ be as given in equation* (2.15)*. Then*

$$\begin{aligned}\max_{x \in \mathcal{A}} \left(\prod_{i=1}^{r} x_i \right) &\leq (k-1)^s (k+1)^{s-2} (k+2) \\ &= (k-1)^{r/2+1/2} (k+1)^{r/2-3/2} (k+2)\end{aligned}$$

Proof. Let $y \in \mathcal{A}$. Note that if $y_i < k - 1$ and $y_j > k + 1$ then, in the notation of [MOA11], let $T_{i,j}$ be the T-transformation such that if $y' = T_{i,j}(y)$, $y'_\ell = y_\ell$ for $\ell \notin \{i, j\}$ and $y'_i = y_i + 1$ and $y'_j = y_j - 1$. We have $y' \prec y$ and thus $\prod y'_i \geq \prod y_i$. Iterating these T-transformations we can assume that y has coordinates lying in $\{k - 2, k - 1, k + 1, k + 2\}$ with at most one of $k - 2$ and $k + 2$ appearing (else we would apply a T-transformation to reduce them). Suppose some $y_i = k - 2$. Then we do not have $y_j = k + 2$ for any j and thus there must exist j, ℓ such that $y_j = y_\ell = k + 1$. Perform the transformation $(k + 1, k + 1, k - 2) \to (k + 2, k - 1, k - 1)$. Note that this transformation need not be majorizing, however it does increase the product as $(k + 2)(k - 1)^2 - (k + 1)^2(k - 2) = 4$. If the coordinate $k - 2$ remains then apply a T-transformation to $(k + 2, k - 2) \to (k + 1, k - 1)$. Iterating this procedure we can assume that the coordinate $k-2$ does not appear and that the coordinate $k+2$ appears precisely once. The product of the coordinates is then $(k - 1)^s (k + 1)^{s-2} (k + 2)$. □

Now we show that the bound for $M_R(m, n, k)$ given in Theorem 2.12 holds for $M_T(m, n, k)$ when k is an integer.

Theorem 2.27. *Let $k \geq 2$ be an integer. Let*

$$B(m, n, k) = \left(\sqrt{k^2 - 1} \right)^{\frac{m}{2} - \frac{n}{2k}} k^{\frac{n}{2k}} \tag{2.16}$$

as in Theorem 2.12. Then $M_T(n,k) \leq B(m,n,k)$. *In particular, let*

$$c_k = \left(\sqrt{k^2-1}\right)^{\frac{1}{2}\left(1-\frac{1}{k}\right)} k^{\frac{1}{2k}}.$$

Then $M_T(n,k) \leq c_k^n$.

Proof. Let $A \in T(m,n,k)$. We assume $\mathrm{Vol}(A) > 0$ and so the row sums of A are positive integers. Let r be the number of rows not summing to k. Denote their sums by $a_1, \ldots, a_r$. If we apply Hadamard's inequality to these r rows not summing to k and Theorem 2.12 to the rows summing to k we have

$$\mathrm{Vol}(A) \leq \left(\prod_{i=1}^{r} \sqrt{a_i}\right) B(m-r,n,k). \tag{2.17}$$

If r is even then by Lemma 2.25 we have

$$\prod_{i=1}^{r} a_i \leq (k-1)^{r/2}(k+1)^{r/2}.$$

If r is odd then by Lemma 2.26 we have

$$\prod_{i=1}^{r} a_i \leq (k-1)^{r/2+1/2}(k+1)^{r/2-3/2}(k+2).$$

Note that

$$(k-1)^{r/2}(k+1)^{r/2} \geq (k-1)^{r/2+1/2}(k+1)^{r/2-3/2}(k+2)$$

holds provided

$$(k+1)^3 > (k-1)(k+2)^2$$

which holds for all $k > 0$. Thus we have,

$$\begin{aligned}
\mathrm{Vol}(A) &\leq \left(\prod_{i=1}^{r} \sqrt{a_i}\right) B(m-r,n,k) \\
&\leq \sqrt{(k-1)^{r/2}(k+1)^{r/2}}B(m-r,n,k) \\
&= \sqrt{k^2-1}^{r/2} \left(\sqrt{k^2-1}\right)^{\frac{m-r}{2}-\frac{n}{2k}} k^{\frac{n}{2k}} \\
&= \left(\sqrt{k^2-1}\right)^{\frac{m}{2}-\frac{n}{2k}} k^{\frac{n}{2k}} \\
&= B(m,n,k)
\end{aligned}$$

as desired. □

The intuition behind this result is that if $A \in T(n,k)$ has many rows not summing to k then its determinant must be small. Thus A is "nearly" in $R(n,k)$ and since we can apply the same argument to the columns it is "nearly" in $S(n,k)$. This suggests the following conjecture.

Conjecture 2.28. *For any integer $k \geq 2$,*

$$\limsup_{n\to\infty} M(n,k)^{1/n} = \limsup_{n\to\infty} M_R(n,k)^{1/n} = \limsup_{n\to\infty} M_T(n,k)^{1/n}.$$

Recall from Section 2.2 that $\limsup_{n\to\infty} M(n,2)^{1/n} = \limsup_{n\to\infty} M_R(n,2)^{1/n} = 2^{1/3}$, and $\limsup_{n\to\infty} M_T(n,k)^{1/n} \leq 6^{1/6}$. So this conjecture is open even for $k = 2$.

Example, $k = 3$

Recalling that $c_3 = 24^{1/6} \approx 1.6984$ we have $M_T(n,3) \leq 1.6984^n$. Recall from Section 2.2 that a construction based on the Fano plane gives the lower bound $M_T(n,3) \geq (24^{1/7})^n \approx 1.5746^n$ for infinitely many n. So $\limsup_{n\to\infty} M_T(n,3)^{1/n} \in [24^{1/7}, 24^{1/6}]$. The authors of [BR18] conjecture that $24^{1/7}$ is the true value. We echo this sentiment. At the very least we do not believe our upper bound is tight. Recall that in Section 2.5 we showed a smaller upper bound for $\limsup_{n\to\infty} M_R(n,3)^{1/n}$ and we have conjectured that these two values are the same.

In our proof of Theorem 2.27 we argued that a matrix in $T(n,k)$ that has many rows not summing to k must have determinant smaller than c_k^n. If we consider $T(n,\tilde{k})$ for non-integer $\tilde{k} \in (k, k+1)$ we can extend this reasoning to argue that if the rows of a matrix in $T(n,\tilde{k})$ are not mostly of weight k and $k+1$ in the appropriate ratio then the determinant will be small. This is Theorem 2.29 below.

Theorem 2.29. *Let $\tilde{k} > 1$ be a rational number. Let $k = \lfloor \tilde{k} \rfloor$. Let $\gamma = \tilde{k} - k$. Let $m_1 = (1-\gamma)m$ and $m_2 = \gamma m$. Let $B(m,n,k)$ be as given in equation (2.16). Then*

$$M_T(m,n,\tilde{k}) \leq B(m_1,n,k)^{1-\gamma} B(m_2,n,k+1)^{\gamma}.$$

Consequently, let $c_{\tilde{k}} = c_k^{1-\gamma} c_{k+1}^{\gamma} < \sqrt{\tilde{k}}$. Then $M_T(n,\tilde{k}) \leq (c_{\tilde{k}})^n$.

Proof. Let $A \in T(m,n,\tilde{k})$. Let m_k be the number of rows of A summing to k and let m_{k+1} be the number of rows of A summing to $k+1$. Let A' be the largest $m' \times n$

submatrix of A such that the rows of A' have sums lying in $\{k, k+1\}$ and the number of ones in A' is within one of $\tilde{k}m'$. Let $r = m - m'$ be the number of remaining rows. We argue via strong induction on r. If $r = 0$ then $A' = A$ and A consists of only rows summing to k and $k+1$ and must be in the proportion $1-\gamma : \gamma$ and the result follows from Theorem 2.27. Now suppose $r > 0$ and inductively that the result holds for smaller values of r. Let $a_1 \ldots, a_r$ be the sums of the rows not in A'. We bound $\mathrm{Vol}(A) \leq \mathrm{Vol}(A') \prod_{i=1}^{r} \sqrt{a_i}$. Note that either k or $k+1$ is excluded from the a_i by the maximality of A'. Further there is some $a_i \notin \{k, k+1\}$ otherwise we would have $r = 0$. So we have at least one of two cases

1. If the number of ones of B is at least $\tilde{k}n$ then there exists $a_i \leq k-1$ and $a_j \geq k+1$.
2. If the number of ones of B is at most $\tilde{k}n$ then there exists $a_i \leq k$ and $a_j \geq k+2$.

Now we apply the reductions in the proof of Theorem 2.27 we can assume that if we are in the first case that there is some $a_i = k-1$ and some $a_j = k+1$. We have $\sqrt{a_i a_j} = \sqrt{k^2-1}$. Noting that $B(m_1, n, k)^{1/m_1} > \sqrt{k^2-1}$ for any m_1 so we see that gives a smaller bound using induction on $r-2$. Similarly, in case two we have $\sqrt{k(k+2)} = \sqrt{(k+1)^2-1} < B(m_2, n, k+1)^{1/m_2}$ and so adjoining these gives a smaller bound using induction on $r-2$. □

Example, $\tilde{k} = 2.25$

Let $\tilde{k} = 2.25 = 1.5^2$. We hope to show that $M_T(n, \tilde{k})$ is exponentially smaller than 1.5^n. Indeed this is the case. We have $c_2 = 12^{1/8}$, $c_3 = 24^{1/6}$ and $\gamma = 1/4$. Then we have $c_{\tilde{k}} = c_2^{1-\gamma} c_3^{\gamma} = 12^{3/32} 24^{1/24} \approx 1.4411$. So $M_T(n, 2.25) < 1.4411^n$.

2.8 Perturbations

The techniques in this chapter can be applied to perturbations of combinatorial matrices. There are many different generalizations one might make. In this section we give a small illustration.

Definition 2.30. *For $\delta \in [0, 1)$, let $R_\delta(n, k)$ be the set of $n \times n$ matrices where each row has exactly k non-zero elements each lying in the interval $[1-\delta, 1+\delta]$.*

We can think of a matrix in $R_\delta(n,k)$ as a perturbation of a matrix in $R(n,k)$. If $A \in R_\delta(n,k)$ then the rows have norms at most $\sqrt{k}(1+\delta)$. So Hadamard's inequality tells us that $\det(A) \le k^{n/2}(1+\delta)^{n/2}$. The techniques in this chapter can be used to improve this bound. We illustrate this with the following generalization of Theorem 2.12.

Theorem 2.31. *If $A \in R_\delta(n,k)$, then $\det(A) \le d_\delta(k)^n$ where*

$$d_\delta(k) = \sqrt{k^2(1+\delta)^2 - (1-\delta)^2}^{\frac{1}{2}(1-1/k)} (k(1+\delta)^2)^{\frac{1}{2k}}.$$

Proof. The proof is nearly identical to that of Theorem 2.12. If two rows have overlapping nonzero entries their volume is at most

$$\det \begin{pmatrix} k(1+\delta) & 1-\delta \\ 1-\delta & k(1+\delta) \end{pmatrix} = \sqrt{k^2(1+\delta)^2 - (1-\delta)^2}$$

which is analogous to $\sqrt{k^2-1}$ in the unperturbed case. Once we can no longer guarantee an overlapping pair of rows we apply Hadamard's inequality which uses the max row norm of $k(1+\delta)$. □

If $\delta = o(1/k^2)$ then we will show in Theorem 2.36 in Section 2.10 that for k sufficiently large, $d_\delta(k) < \sqrt{k}$ so an inequality stronger than Hadamard applied to the unperturbed matrix still holds. One can of course consider perturbations of the zero elements as well. In each of these cases the techniques of Sections 2.4 and 2.5 can be applied.

Example, $k = 4$, $\delta = 0.01$

Let $k = 4$ and $\delta = 0.01$ and suppose $A \in R_\delta(n,k)$.

1. We have $\sqrt{k}(1+\delta) = 2.02$. Thus Hadamard's inequality implies $\det(A) \le 2.02^n$.
2. Using Theorem 2.31, we have $\det(A) \le d_\delta(k)^n \approx 1.9892^n$.

2.9 Conclusion and open questions

We summarize some of our results for various k in Table 2.2.

k	$c_{1,k} = \sqrt{k}$	$c_{2,k}$	q_*	$c_{q_*,k}$	α_k	β_k
3.0	1.7321	1.6984	2	1.6984	21.91	1.6977
4.0	2.0	1.9759	3	1.9719	782.53	1.9702
5.0	2.2361	2.2179	3	2.2116	1.2591×10^5	2.2097
6.0	2.4495	2.4352	4	2.4279	1.0075×10^8	2.4257
7.0	2.6458	2.6341	4	2.6258	4.3557×10^{11}	2.6240
8.0	2.8284	2.8187	5	2.8103	1.0925×10^{16}	2.8083
9.0	3.0	2.9917	5	2.9828	1.6920×10^{21}	2.9812
10.0	3.1623	3.1551	5	3.1462	1.7105×10^{27}	3.1447

Table 2.2: A summary of bounds for $k = 3, \ldots, 10$, q_* is the optimal value of q that minimizes $c_{q,k}$ for $q = 1, \ldots, k$.

Since $M(n,k) \leq M_R(n,k)$, Theorem 2.11 shows for k one more than a prime power that

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \geq \sqrt{k} - \frac{1}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2-\varepsilon}}\right).$$

From Theorem 2.33 we have that for a real number $t \approx 0.096$,

$$\limsup_{n\to\infty} M_R(n,k)^{1/n} \leq \sqrt{k} - \frac{t}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2}}\right).$$

Our main open question is resolving this gap.

Resolving the value of $\limsup_{n\to\infty} M_R(n,k)^{1/n}$ for small k is also an interesting question. We do not claim that the constants we have found are the best possible. For example, we showed that $\limsup_{n\to\infty} M_R(n,3)^{1/n} \leq \beta_3 = (40/11)^{11/36} 2^{7/36} \approx 1.6977$. The Fano plane construction gives the lower bound $\limsup_{n\to\infty} M_R(n,3)^{1/n} \geq 24^{1/7} \approx 1.5746$.

We can ask similar questions for $M(n,k)$ and $M_T(n,k)$. In Conjecture 2.28 we conjectured that

$$\limsup_{n\to\infty} M(n,k)^{1/n} = \limsup_{n\to\infty} M_R(n,k)^{1/n} = \limsup_{n\to\infty} M_T(n,k)^{1/n}$$

for all integers k. Recall that we know that for $k = 2$ the first equality holds, but the second is open. Thus, resolving this conjecture even for small k would be quite interesting.

We note one avenue through which our approach may be improved. For $A \in R(m,n,k)$ let $q_{\max}$ be the maximal column sum of A. Then we can take the appropriate $q_{\max}$ rows and bound their volume. In our approach we use the fact that the

matrix resulting after the deletion of these rows lies in $R(m - q_{\max}, n, k)$. However, we know the resulting matrix has a zero column since we have removed all the ones in a maximal sum column. Thus we could recursively use an inequality for the volume of $R(m - q_{\max}, n - 1, k)$. This smaller matrix has a larger density of ones and gives a better bound. This is of course harder to analyze since the maximum column sum depends on A.

For what values of n, k are any of $M(n, k)$, $M_R(n, k)$ and $M_T(n, k)$ equal? For small values of n and k we of course know the answer and if $\lambda = k(k-1)/(n-1)$ and there is an (n, k, λ) combinatorial design these are all equal. We observed in Section 2.2 that $M(7, 2) \neq M_R(7, 2)$. Are there certain values of k for which equality always holds? The same questions apply to $M_R(n, k)$ and $M_T(n, k)$. Finally, we wonder for $\Theta(n^{1/3}) \leq k < \sqrt{n}$, a domain on which no (n, k, λ) combinatorial design exists how much can Ryser's bound be improved?

2.10 Calculations

Theorem 2.13. *Let c_k be defined as in Theorem 2.12 and $\lambda = k(k-1)/(n-1)$ as in Theorem 2.4. If $k = o(n^{1/3})$ then for n sufficiently large, $c_k^n < k(k-\lambda)^{(n-1)/2}$.*

Proof. We want to show that

$$\left(\left(\sqrt{k^2-1}\right)^{\frac{1}{2}\left(1-\frac{1}{k}\right)} k^{\frac{1}{2k}}\right)^n < k(k-\lambda)^{\frac{1}{2}(n-1)} = \frac{k}{k-\lambda}(k-\lambda)^{n/2}. \tag{2.18}$$

Raising both sides to the power $2k/n$ we obtain

$$\left(\sqrt{k^2-1}\right)^{k-1} k < \left(\frac{k}{k-\lambda}\right)^{2k/n} (k-\lambda)^k.$$

So it suffices to show

$$\left(\sqrt{k^2-1}\right)^{k-1} k < (k-\lambda)^k.$$

Since $\sqrt{k^2-1} < k - \frac{1}{2k}$, it suffices to show

$$\left(k - \frac{1}{2k}\right)^{k-1} k < (k-\lambda)^k$$

which simplifies to

$$\left(1 - \frac{1}{2k^2}\right)^{k-1} < (1-\lambda/k)^k.$$

Taking logs,

$$(k-1)\log\left(1-\frac{1}{2k^2}\right) < k\log(1-\lambda/k),$$

thus

$$(k-1)\log\left(-\frac{1}{2k^2}+O\left(\frac{1}{k^4}\right)\right) < k\left(\frac{-\lambda}{k}+O\left((\lambda/k)^2\right)\right) = -\lambda+O\left(\frac{\lambda^2}{k}\right).$$

Thus it suffices to show that

$$\frac{1}{2k^2} \gg \frac{\lambda}{k-1} = \frac{k}{n-1}$$

which holds provided $k = o(n^{1/3})$. □

Next we show that for large k, $c_{q,k}$ is minimized when $q \approx 0.44k$.

Theorem 2.32. *Let*

$$c_{q,k} = (q+k-1)^{\frac{1}{2q}\left(1-\frac{q-1}{k}\right)}(k-1)^{\frac{1}{2}\frac{q-1}{q}\left(1-\frac{q-1}{k}\right)}k^{\frac{(q-1)}{2k}}$$

as in Theorem 2.16. Let

$$q_* = \operatorname{argmin}_{q=1,\ldots,k} c_{k,q}.$$

Let $s \approx 0.4395$ be the positive root of

$$s^3 + s - \log(1+s)(s+1) = 0. \tag{2.19}$$

Then $\lim_{k\to\infty} \frac{q_*}{k} = s$.

Proof. We have

$$c_{q,k}^2 = (q+k-1)^{\frac{1}{q}\left(1-\frac{q-1}{k}\right)}(k-1)^{\frac{q-1}{q}\left(1-\frac{q-1}{k}\right)}k^{\frac{(q-1)}{k}}. \tag{2.20}$$

Noting that the exponents in equation (2.20) sum to one we have

$$\frac{c_{q,k}^2}{k} = \left(1+\frac{q-1}{k}\right)^{\frac{1}{q}\left(1-\frac{q-1}{k}\right)}\left(1-\frac{1}{k}\right)^{\frac{q-1}{q}\left(1-\frac{q-1}{k}\right)}$$

Let $s = (q-1)/k$. Since $c_{2,k} < c_{1,k}$ we can assume $q > 1$ and thus $s \in (0,1)$. We have

$$\frac{c_{q,k}^2}{k} = (1+s)^{\frac{1-s}{sk+1}}\left(1-\frac{1}{k}\right)^{\frac{1-s}{sk+1}sk}.$$

Thus

$$G(s,k) := \log\left(\frac{c_{q,k}^2}{k}\right) = \frac{1-s}{sk+1}\left(\log(1+s) + sk\log\left(1-\frac{1}{k}\right)\right)$$
$$= \frac{1-s}{sk+1}\left(\log(1+s) - s + O\left(\frac{1}{k}\right)\right)$$
$$= \frac{1-s}{sk+1}\left(\log(1+s) - s\right) + O\left(\frac{1}{k^2}\right)$$

Then,

$$\frac{d}{ds}G(s,k) = \frac{ks^3 + ks + 2s^2 - \log(1+s)(ks+k+s+1)}{(ks+1)^2(s+1)} + O\left(\frac{1}{k^3}\right)$$

and thus

$$\frac{(ks+1)^2(s+1)}{k}\frac{d}{ds}G(s,k) = (s^3 + s - \log(1+s)(s+1)) + O\left(\frac{1}{k}\right).$$

So the value of s that, asymptotically, minimizes $G(s,k)$ is the positive root of equation (2.19). □

Theorem 2.33. *Let $s \approx 0.4395$ be the positive root of $s^3 + s - \log(1+s)(s+1) = 0$ as in Theorem 2.32. Let*

$$t = -\frac{(1-s)(\log(1+s) - s)}{s} \approx 0.09591. \tag{2.21}$$

Then

$$c_{sk,k} = \sqrt{k} - \frac{t}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2}}\right) \tag{2.22}$$

as stated in equation (2.4) *in Section 2.4.*

Proof. In the proof of Theorem 2.32 we showed that

$$\log\left(\frac{c_{q,k}^2}{k}\right) = \frac{1-s}{sk+1}\left(\log(1+s) - s\right) + O\left(\frac{1}{k^2}\right).$$

Thus

$$\log\left(\frac{c_{q,k}^2}{k}\right) = \frac{(1-s)\left(\log(1+s) - s\right)}{s}\frac{1}{k} + O\left(\frac{1}{k^2}\right)$$
$$= -\frac{t}{k} + O\left(\frac{1}{k^2}\right).$$

Thus

$$\frac{c_{q,k}^2}{k} = e^{-t/k}\left(1 + O\left(\frac{1}{k^2}\right)\right)$$

which simplifies to

$$\begin{aligned} c_{q,k}^2 &= ke^{-t/k} + O\left(\frac{1}{k}\right) \\ &= k - t + O\left(\frac{1}{k}\right). \end{aligned}$$

Noting that

$$\sqrt{k-t} = \sqrt{k}\sqrt{1-t/k} = \sqrt{k}\left(1 - \frac{t}{2k} + O\left(\frac{1}{k^2}\right)\right),$$

we have

$$c_{q,k} = \sqrt{k} - \frac{t}{2\sqrt{k}} + O\left(\frac{1}{k^{3/2}}\right)$$

as desired. □

Next we show that for n sufficiently large, Theorem 2.16 gives an improved upper bound for $M_R(n,k)$ compared to Ryser's theorem for $k < \sqrt{n/10}$.

Theorem 2.34. *Let $c_{q,k}$ be as given in Theorem 2.16. Let $s \approx 0.4395$ be the positive root of $s^3 + s - \log(1+s)(s+1) = 0$ as in Theorem 2.32. For n sufficiently large and $k < \sqrt{n/10}$ we have*

$$c_{sk,k}^n < k(k-\lambda)^{(n-1)/2}.$$

Proof. Let $H(n,k) = k(k-\lambda)^{(n-1)/2}$ be Ryser's bound. Then,

We can write Ryser's bound as

$$H(n,k) = k(k-\lambda)^{(n-1)/2} = \frac{k}{k-\lambda}(k-\lambda)^{n/2}$$

and thus

$$\begin{aligned} \frac{H(n,k)^{2/n}}{k} &= \left(\frac{k}{k-\lambda}\right)^{2/n}\frac{k-\lambda}{k} \\ &> 1 - \frac{\lambda}{k} \\ &= 1 - \frac{k-1}{n-1}. \end{aligned}$$

On the other hand, we showed in the proof of Theorem 2.33 that

$$\frac{c_(sk,k)^{2/n}}{k} = 1 - \frac{t}{k} + O\left(\frac{1}{k^2}\right).$$

So we want

$$1 - \frac{t}{k} + O\left(\frac{1}{k^2}\right) < 1 - \frac{k-1}{n-1}. \tag{2.23}$$

Equation (2.23) holds trivially if $k = o(n^{1/2})$ and if $k = \Theta(\sqrt{n})$ then equation (2.23) holds provided

$$\frac{t}{k} > \frac{k-1}{n-1}$$

which noting that $t < 1/10$ holds for $k < \sqrt{n/10}$ and n sufficiently large. □

Next we show that for constant k, Theorem 2.17 gives a better asymptotic than Theorem 2.12.

Theorem 2.35. *Let*

$$c_k = \left(\sqrt{k^2-1}\right)^{\frac{1}{2}\left(1-\frac{1}{k}\right)} k^{\frac{1}{2k}}$$

as in Theorem 2.12 and

$$\beta_k = \left(k + \frac{k}{H_k} - 1\right)^{\frac{1}{2}(H_k/k)} (k-1)^{\frac{1}{2}(1-H_k/k)}$$

as in Theorem 2.17. Then $\beta_k < c_k$.

Proof. If we raise β_k and c_k to the power $2k$ and compare we want to show that

$$\left(k + \frac{k}{H_k} - 1\right)^{H_k} (k-1)^{k-H_k} < \sqrt{k^2-1}^{k-1} k.$$

Rearranging, this is equivalent to

$$\left(1 + \frac{1}{H_k} - \frac{1}{k}\right)^{H_k} < \frac{\sqrt{k^2-1}^{k-1}}{(k-1)^{k-H_k}}. \tag{2.24}$$

We see that for all k the left hand side of equation (2.24) is less than e. We use the

inequality $\sqrt{k^2-1} > k - 1/k$ to bound the right hand side.

$$\begin{aligned}\frac{\sqrt{k^2-1}^{k-1}}{(k-1)^{k-H_k}} &> \frac{\left(k-\frac{1}{k}\right)^{k-1}}{(k-1)^{k-H_k}} \\ &= \left(\frac{k-\frac{1}{k}}{k-1}\right)^{k-H_k}\left(k-\frac{1}{k}\right)^{H_k-1} \\ &= \left(1+\frac{1}{k}\right)^{k-H_k}\left(k-\frac{1}{k}\right)^{H_k-1} \\ &> 1\cdot k = k\end{aligned}$$

for $k \geq 4$. Since $4 > e$ the result holds for $k \geq 4$ and one can easily check that it holds for $k < 4$. □

Theorem 2.36. *Let* $d_\delta(k) = \sqrt{k^2(1+\delta)^2-(1-\delta)^2}^{\frac{1}{2}(1-1/k)}(k(1+\delta)^2)^{\frac{1}{2k}}$ *as in Theorem 2.31. Then for* $\delta = o(1/k^2)$, $d_\delta(k) < \sqrt{k}$.

Proof. Raising both sides of the inequality $d_\delta(k) < \sqrt{k}$ to the power $2k$ we find

$$\sqrt{k^2(1+\delta)^2-(1-\delta)^2}^{k-1}k(1+\delta)^2 < k^k$$

which we can simplify to

$$\left(\frac{\sqrt{k^2(1+\delta)^2-(1-\delta)^2}}{k}\right)^{k-1} < \frac{1}{(1+\delta)^2}. \tag{2.25}$$

We simplify and apply the inequality $\sqrt{a^2-b^2} < a - \frac{b^2}{2a}$ in the left hand side of equation (2.25) to obtain

$$\begin{aligned}\left(\frac{\sqrt{k^2(1+\delta)^2-(1-\delta)^2}}{k}\right)^{k-1} &= \sqrt{(1+\delta)^2-(1-\delta)^2/k^2}^{k-1} \\ &\leq \left(1+\delta-\frac{(1-\delta)^2}{2k^2(1+\delta)}\right)^{k-1} \\ &= \left(1-\frac{1}{2k^2}+o\left(\frac{1}{k^2}\right)\right)^{k-1} \\ &= 1-\frac{1}{2k}+o\left(\frac{1}{k^2}\right).\end{aligned}$$

The right hand side of equation (2.25) is

$$\frac{1}{(1+\delta)^2} = 1-\delta^2+o(\delta^2) = 1-o\left(\frac{1}{k^4}\right).$$

So we see the inequality holds. □

Chapter 3

Unique sum free sets

3.1 Introduction

Let p be a prime number. Let $A \subseteq \mathbb{F}_p$ be nonempty. We say that A is *unique sum free* (USF) if every element of the sumset $A + A$ can be written as a sum of two elements from A in at least two different ways. That is for any $s \in A + A$ there exist a, b, c, d with $\{a, b\} \neq \{c, d\}$ such that $s = a + b = c + d$. For example, if $p = 13$ then $A = \{0, 1, 2, 3, 6, 8, 10\}$ is a USF set in $\mathbb{F}_p$. In this case $A + A = \mathbb{F}_p$. We observe, for example, that the sum 9 appears as $1+8 = 3+6$ and the sum 0 appears as $0+0 = 3+10$.

For $p > 2$ it is easy to find large USF sets, for example the entirety of $\mathbb{F}_p$ is USF for $p > 2$, but the problem of finding small USF sets is more challenging and was posed by Kopparty in [Kop17].

As motivation we give the following geometrical reformulation [Fra]. Consider p points on a circle spaced uniformly (for example the p-th roots of unity). Color some subset A of size n of the points orange. For any two orange points draw the line connecting them. If the two points are the same, draw the tangent line to the circle at that point. So we have drawn $\binom{n}{2} + n$ distinct lines. Then A being USF is equivalent to the statement that every line has a parallel. So we may interpret USF as "unique slope free." We would like to find as small a set of points as possible with this property.

Definition 3.1. *Let G be an abelian group. Let $A \subseteq G$ be nonempty. We say that A is unique sum free (USF) if for all $s \in A + A$ there exist $a, b, c, d \in A$ with $\{a, b\} \neq \{c, d\}$ such that $s = a + b = c + d$.*

As noted above we are mainly interested in the case $G = \mathbb{Z}/p\mathbb{Z}$. Since we will often find it useful to scale our set multiplicatively, we will use the field structure of

$\mathbb{Z}/p\mathbb{Z}$. Unless specified otherwise if we state a set A is USF we mean in a field $\mathbb{F}_p$. We caution the reader that a number of papers, e.g. [HS86, Jan07, NQ08], concern the similar question of when a set $A \subseteq \mathbb{F}_p$ is such that for all $s \in A + A$, $|\{(a, b) \in A^2 : a + b = s\}| \geq 2$. In this case they count $a + b = b + a$ as distinct representations of $a + b$ provided $a \neq b$. Thus in this case the question is if $2a$ has an alternative representation as a sum of two elements from A for each $a \in A$. Such sets will be fundamental to our results and will be discussed later.

Definition 3.2. *For $A \subseteq \mathbb{F}_p$ and $s \in A + A$ let $\nu_A(s) = |\{(a, b) \in A^2 : a + b = s\}|$ be the number of ordered representations of s as a sum of elements from A. Let $\nu(A) = \min_{s \in A+A} \nu_A(s)$.*

Thus A is USF means that $\nu(A) \geq 3$. Using the nomenclature of [NQ08] we have the following definition.

Definition 3.3. *Let $A \subseteq \mathbb{F}_p$ be nonempty. We say that A is balanced if for all $a \in A$ there exist $b, c \in A$ with $b \neq c$ such that $2a = b + c$.*

Thus A is balanced means that $\nu(A) \geq 2$. To avoid confusion we will maintain this use of the word "representation."

Definition 3.4. *For $s \in A + A$ we say that $\{\{a, b\} : a, b \in A,\ a + b = s\}$ are the set-representations of s.*

Definition 3.5. *Given a set A we say $s \in A + A$ is unique if s has exactly one set-representation.*

For example if $A = \{1, 2, 3, 4, 5\} \subseteq \mathbb{Z}$ then 4 has two set-representations and 3 is a unique sum. Thus definition 3.1 is consistent with the name unique sum free.

Definition 3.6. *For any prime p let $\mu(p) = min\{|A| : A$ is USF in $\mathbb{F}_p\}$. That is, $\mu(p)$ is the minimal size of a USF subset of $\mathbb{F}_p$. We adopt the convention $\mu(2) = \infty$.*

In this introduction we will give some simple lower and upper bounds for $\mu(p)$. These will be improved in subsequent sections. Observe that if A is USF then so is $uA + b$ where $u, b \in \mathbb{F}_p$ and u is a unit. We formalize this observation in the following definition.

Definition 3.7. *We say a property, $\mathcal{P}$, of subsets of $\mathbb{F}_p$ is affine invariant if $\mathcal{P}$ holds for A if and only if $\mathcal{P}$ holds for $uA + b$ for all $u \in \mathbb{F}_p^*$ and $b \in \mathbb{F}_p$.*

So USF is an affine invariant property. In studying USF sets it is useful to note that although arithmetic progressions (APs) have lots of additive structure they do not yield small examples of USF sets. Notice that if $A \subseteq \mathbb{Z}$ is a finite set of integers then A cannot be USF as if $a = \min(A)$ then $2a$ has no set-representations other than $a + a$. If $A \subseteq \mathbb{F}_p$ is an arithmetic progression of length k then there exist $u, b \in \mathbb{F}_p$ so that $uA + b = \{0, 1, \ldots, k-1\}$. If $k < p/2$ then although $uA + b \subseteq \mathbb{F}_p$ since it lies in $[0, p/2]$ addition behaves as in $\mathbb{Z}$ and thus we can think of it as a set of integers and thus is not USF. So if A is an arithmetic progression and USF we must have $|A| > p/2$. We mention arithmetic progressions in part because they demonstrate the delicacy of this problem. If $A = \{1, \ldots, k\}$ then notice that $A+A = \{2, 3, \ldots, 2k\}$ with $|A+A| = 2k-1$ and the only elements of $A + A$ that are unique sums are $\{2, 3, 2k-1, 2k\}$. Thus as k grows the proportion of unique sums goes to zero. Nonetheless, such a set does not seem to be in any way "close" to USF. In fact we see next that we cannot hope to construct a constant sized USF set.

Theorem 3.8. *Let $A \subseteq \mathbb{F}_p$ be a set of size n with $n < \log_4 p$. Then there exists $u \in \mathbb{F}_p^*$ and $b \in \mathbb{F}_p$ so that $uA + b \subseteq [0, \lfloor p/2 \rfloor - 1]$.*

Proof. Let $A = \{a_1, a_2, \ldots, a_n\}$. Apply an arbitrary ordering to A and thus associate to A a vector $v = (a_1, \ldots, a_n) \in \mathbb{F}_p^n$. Partition $\mathbb{F}_p^n$ into a grid with 4^n boxes. For example the box containing the origin is $\{0, 1, 2, \ldots, \lfloor p/4 \rfloor\}^n$. The width of a box in any dimension is one of $\lfloor p/4 \rfloor$ and $\lceil p/4 \rceil$. Consider the multiples of v: $v, 2v, 3v, \ldots (p-1)v, pv = \vec{0}$. Since $p > 4^n$ there is a box containing two vectors iv, jv. Then if we take $(j-i) \cdot v$ this lies in a box whose width is at most $2(\lceil p/4 \rceil - 1) + 1 = 2\lceil p/4 \rceil - 1 \leq \lfloor p/2 \rfloor$. Thus we set $u = j - i$ and uA lies in an interval of width at most $\lfloor p/2 \rfloor$. Translating by an appropriate b gives the result. □

Corollary 3.9. *If A is USF then $|A| > \log_4 p$. Thus, $\mu(p) > \log_4 p$.*

This result was unknown to us when we initially studied this problem, however it

appears in [Str76] who was studying a related problem that we will discuss subsequently. This result inspires the study of the affine diameter of a set. If $A \subseteq \mathbb{F}_p$, we can identify it as a set of integers by taking representatives in $[0, p-1]$. Then $\operatorname{diam}(A) = \max(A) - \min(A)$. Then let

$$m_n = \max_{|A|=n} \min_{\alpha \neq 0, \beta} \operatorname{diam}(\alpha A - \beta) \tag{3.1}$$

So we have shown in Theorem 3.8 that if $n = \log_4 p$ then $m_n < n/2$. Bounding m_n is studied in detail in [Lev00].

We next want to give an upper bound for $\mu(p)$. It is an elementary exercise to show that for any $\varepsilon > 0$, a random subset of size $p^{1/2+\varepsilon}$ will be USF with high probability. However, this can be easily improved to $O(\sqrt{p})$.

Theorem 3.10. *For any prime $p > 2$, $\mu(p) \leq 2\sqrt{2}\sqrt{p} + O(1)$.*

Proof. For some positive integer $a > 1$, let $A = \{0, 1, 2, \ldots, 2a-1, 2a, 3a, 4a, \ldots, ka\}$ where k is such that $ka < p < (k+1)a$. We can think of the elements of A as residues in $\mathbb{F}_p$. Then $A + A = \mathbb{F}_p$ and if $s = ca + r$ with $0 \leq r < a$ then $s = ca + r = (c-1)a + (a+r)$. So A is USF. We have $|A| \leq 2a + p/a$. Thus we let $a = \left\lfloor \sqrt{p/2} \right\rfloor$ and have $|A| = 2\sqrt{2}\sqrt{p} + O(1)$. □

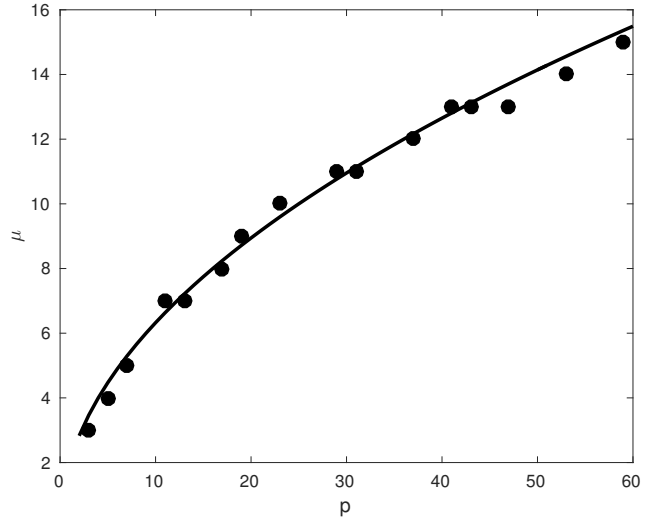
This construction finds a USF set for which every element of $\mathbb{F}_p$ is set-represented at least twice. Since a set of size n has $\binom{n}{2} + n = n^2/2 + O(n)$ candidate sums, the smallest set such that every element of $\mathbb{F}_p$ is set-represented at least twice has size at least $2\sqrt{p}$. This inspires the following question. What is the smallest value c such that for p sufficiently larger there exists a USF set $A \subseteq \mathbb{F}_p$ with $|A| < c\sqrt{p} + o(\sqrt{p})$ and $A + A = \mathbb{F}_p$? We have shown that $c \leq 2\sqrt{2}$.

Thus far we have shown $\log_4(p) < \mu(p) < O(\sqrt{p})$. Resolving this large gap is our goal. Kopparty [Kop17] conjectured that $\mu(p) = \Theta(\sqrt{p})$. For the first few odd primes the values of $\mu(p)$ can be found in Table 3.1. A plot of $\mu(p)$ for the first few primes looks much like a plot of $2\sqrt{p}$. See Figure 3.1. These data and other experiments seemed to bolster the conjecture that $\mu(p) = \Theta(\sqrt{p})$. However, as we will see this conjecture is false. Our main result is that $\mu(p) = O(\log^2 p)$.

p	$\mu(p)$	minimal USF set
3	3	$\{0,1,2\}$
5	4	$\{0,1,2,3\}$
7	5	$\{0,1,2,3,4\}$
11	7	$\{0,1,2,3,4,5,6\}$
13	7	$\{0,1,2,3,6,8,10\}$
17	8	$\{0,1,2,4,10,12,13,14\}$
19	9	$\{0,1,2,3,4,5,9,12,15\}$
23	10	$\{0,1,2,3,4,5,6,9,18,19\}$
29	11	$\{0,1,2,3,4,5,6,12,13,17,20\}$
31	11	$\{0,1,2,3,4,6,8,11,13,21,23\}$
37	12	$\{0,1,2,3,4,5,6,10,20,21,30,32\}$
41	13	$\{0,1,2,3,4,5,7,9,21,23,31,32,36\}$
43	13	$\{0,1,2,3,4,5,6,9,14,28,29,30,33\}$
47	13	$\{0,1,2,3,4,17,27,28,34,37,39,44,45\}$
53	14	$\{0,1,2,3,4,8,9,13,14,25,33,35,41,45\}$
59	15	$\{0,1,2,3,4,5,9,10,16,25,27,32,42,44,48\}$

Table 3.1: $\mu(p)$ and an example minimal USF set for primes $p = 3, \ldots, 59$.

Figure 3.1: A plot of $\mu(p)$ for primes $p = 3, \ldots, 59$. The curve drawn is $2\sqrt{p}$.



3.2 Constructing small USF sets

Our construction has two important ingredients. One main idea is that if A is an arbitrary subset of $\mathbb{F}_p$ then $B = A + A$ has many non-unique sums. This follows since if $a, b, c, d \in A$ are distinct then $s = a + b + c + d \in B + B$ and we can write s as a sum of elements from B in two distinct ways $s = (a+b)+(c+d)$ and $s = (a+c)+(b+d)$. This of course does not yield a USF set for arbitrary A. For example, the sum $a + a + a + a$ can not, in general, be obtained in a manner distinct from $2a+2a$. However, the second ingredient we need is the notion of balanced sets given in Definition 3.3. We will show in Lemma 3.11 that if A is balanced then $A + A$ is USF.

Lemma 3.11. *Let $p > 2$ be a prime. Let $A \subseteq \mathbb{F}_p$ be balanced. Let $B = A + A$. Then B is a USF set.*

Proof. Consider an element $s \in B + B$. We can write s as a 4-sum of elements of A. There are five cases. Below a, b, c, d are unique elements of $\mathbb{F}_p$.

1. $s = a + b + c + d$. In this case $s = (a + b) + (c + d) = (a + c) + (b + d)$ are two distinct set-representations of s. Notice that we cannot have $a+b = a+c$ as then $b = c$, nor can we have $a + b = b + d$ as then $a = d$.
2. $s = a + b + c + c$. In this case $s = (a + b) + (c + c) = (a + c) + (b + c)$ are two distinct set-representations of s.
3. $s = a + a + b + b$. In this case $s = (a + a) + (b + b) = (a + b) + (a + b)$ are two distinct set-representations of s.

   Note that to this point we have not yet used the fact that A is balanced.

4. $s = a + a + a + b$. In this case we note that $2a = a_1 + a_2$ with $a_1 \neq a_2$. Then we can write $s = (a + a) + (a + b) = (a + a_1) + (a_2 + b)$. We note that we cannot have $a + a = a + a_1$ so this demonstrates two distinct set-representations unless $a + b = a + a_1$ in which case $a_1 = b$. In this case, we have the alternative set-representation $s = (a+a_2)+(b+b)$. Notice that $a_2+b \neq a+a_2$ and $a_2+b \neq b+b$.

5. $s = a + a + a + a$. In this case we note that $2a = a_1 + a_2$ with $a_1 \neq a_2$. Then we have $s = (a+a)+(a+a) = (a+a_1)+(a+a_2)$ as two distinct set-representations.

□

Note that we needed $p > 2$ above so that if $2a = a_1+a_2$ are distinct set-representations we cannot have $a_1 = a_2$. Lemma 3.11 tells us that if A is a balanced set then there exists a USF set of size at most $|A + A|$. Note that for an arbitrary set A of size n we have $|A + A| \leq \binom{n}{2} + n$. However for a balanced set we know that the doubles have an alternative representation and thus if A is a balanced set of size n, $|A + A| \leq \binom{n}{2}$.

Definition 3.12. *For a prime p let $\alpha(p)$ be the minimal size of a balanced subset of $\mathbb{F}_p$.*

Thus Lemma 3.11 tells us that for $p > 2$, $\mu(p) = O(\alpha(p)^2)$. More precisely, we have $\mu(p) \leq \binom{\alpha(p)}{2} \leq \alpha(p)^2/2$. Note that there do not exist any USF subsets of $\mathbb{F}_2$ and thus $\mu(2) = \infty$. However, we have $\alpha(2) = 2$ as $\{0, 1\}$ is balanced. In the proof of Lemma 3.11 we used the fact that if $a + a = b + c$ and $\{a\} \neq \{b, c\}$ then b and c are distinct. This is of course true once $p > 2$. Notice that balanced is an affine invariant property. Furthermore, much like USF sets there are no finite balanced sets of integers. So Theorem 3.8 applies to balanced sets. That is

$$\alpha(p) > \log_4 p.$$

Table 3.2 gives $\alpha(p)$ and an example of a balanced set of size $\alpha(p)$ for primes $p \leq 61$.

To establish a simple example where $\alpha(p)$ is small let $p = 2^q - 1$ be a Mersenne prime. Consider the polynomial $-2+x+x^2 \in \mathbb{F}_p[x]$. We can factor $-2+x+x^2 = (x-1)(x+2)$ so it has roots $1, -2 \in \mathbb{F}_p$. We note that 2 has order q in $\mathbb{F}_p^*$. Thus -2 has order $2q$ in $\mathbb{F}_p^*$. Let $A = \langle -2 \rangle = \{(-2)^i \ : \ i \in [2q]\}$ be the multiplicative subgroup of $\mathbb{F}_p^*$ generated by -2. Note that A is balanced: for any i, we have $2a^i = a^{i+1} + a^{i+2}$ with exponents taken modulo $2q$. So for p a Mersenne prime we have $\alpha(p) < 2\log_2 p + 1$ and indeed there is a small balanced set which is a multiplicative subgroup. Thus there is a USF set of size $O(q^2) = O(\log^2 p)$.

p	$\alpha(p)$	minimal balanced set
2	2	$\{0,1\}$
3	3	$\{0,1,2\}$
5	4	$\{0,1,2,3\}$
7	5	$\{0,1,2,3,4\}$
11	5	$\{0,1,2,4,7\}$
13	6	$\{0,1,2,3,5,8\}$
17	6	$\{0,1,2,3,6,11\}$
19	6	$\{0,1,2,4,7,12\}$
23	7	$\{0,1,2,3,4,8,15\}$
29	7	$\{0,1,2,3,6,10,19\}$
31	7	$\{0,1,2,3,6,11,20\}$
37	7	$\{0,1,2,4,7,13,24\}$
41	8	$\{0,1,2,3,4,8,15,26\}$
43	7	$\{0,1,2,4,12,23,39\}$
47	8	$\{0,1,2,3,5,9,27,38\}$
53	8	$\{0,1,2,3,5,10,30,43\}$
59	8	$\{0,1,2,3,6,27,44,53\}$
61	8	$\{0,1,2,3,6,11,33,50\}$

Table 3.2: $\alpha(p)$ and an example minimal balanced set for primes $p = 2, \ldots, 61$.

The following construction of a balanced set is due to Straus [Str76]. Straus was trying to construct sets $A \subseteq \mathbb{F}_p$ so that there were no unique differences. He noted that if he constructed a symmetrical set, i.e., $A = -A$, then if $a \neq b$ we have $a-b = (-b)-(-a)$ and thus the remaining challenge is to find another representation for $a - (-a) = 2a$. If this exists, i.e. $2a = b - c$ then since A is symmetrical, $-c \in A$ and thus $b + (-c)$ is a distinct set representative for $2a$. So we see that a symmetrical set has no unique differences if and only if it is balanced. Straus gave the construction

$$A = \{\pm[p/2], \pm[p/4], \ldots, \pm[p/2^i], \ldots, \pm 1, 0\}.$$

Observe that $|A| < 2\log_2 p + 1$. The set A is clearly symmetrical. For $i > 1$, $2[p/2^i] = [p/2^{i-1}] + \varepsilon$ where $\varepsilon \in \{0, -1\}$. Next $2[p/2] = p - 1 = 0 + (-1)$. By symmetry the doubles of the negatives are twice set-represented as well. Thus Straus established that

$$\alpha(p) < 2\log_2 p + 1$$

which is the same bound achieved by the subgroup construction for Mersenne primes. Thus, using Lemma 3.11 and the construction of Straus we have established that $\mu(p)$

is well short of $\Theta(\sqrt{p})$.

Theorem 3.13. *For prime $p \geq 2$, $\mu(p) \leq O(\log^2 p)$.* □

In Section 3.1 we gave a lower bound $\mu(p) > \log_4 p$ based on an argument that a set smaller than size $\log_4 p$ had an affine mapping to lie in $[0, p/2]$. Since no balanced set can lie in $[0, p/2]$ the argument shows that $\alpha(p) > \log_4 p$. Thus we have $\log_4 p < \alpha(p) < 2\log_2 p + 1$. We will discuss improved lower and upper bounds for balanced sets in Section 3.4. For the moment, observe that the constructions we have highlighted of balanced sets grow quadratically under addition. For example, if $p = 2^q - 1$ is a Mersenne prime and $A = \langle -2 \rangle$ then $B := \{1, 2, 4, \ldots, 2^{q-1}\} \subseteq A$ and we see that $|B + B| = \Omega(q^2) = \Omega(|A|^2)$. The same is easily seen to hold for the more general construction of Straus.

Next, we relax the notion of balanced to obtain a larger class of USF sets.

Definition 3.14. *We say a nonempty set $A \subseteq \mathbb{F}_p$ is NUT (no unique triples) if for each $a \in A$ there exists $b, c, d \in A$ with b, c, d not all equal such that $3a = b + c + d$.*

We note that a balanced set is NUT as if $2a = b + c$ with $b \neq c$ then $3a = a + b + c$ and a, b, c are not all equal. Note that the property NUT (like balanced and USF) is an affine invariant property.

Definition 3.15. *For any prime p let $\beta(p) = \min\{|A| \ : \ A \text{ is NUT}\}$.*

Since all balanced sets are NUT, we have $\beta(p) \leq \alpha(p)$. Table 3.3 gives $\beta(p)$ and an example of a NUT set of size $\beta(p)$ for primes $p \leq 61$.

Lemma 3.16. *Let $p > 2$ be a prime and $A \subseteq \mathbb{F}_p$ be a NUT set. Let $B = A + A$. Then B is a USF set.*

Proof. Notice that the result is trivial if $p = 3$, so we can assume $p > 3$ and thus for $x, y \in \mathbb{F}_p$, $3x = 3y$ implies $x = y$. As in the proof of Lemma 3.11 we consider an element element $s \in B + B$. We write s as a 4-sum of elements of A. The first three cases are identical to before and do not require any special structure on A. The final two NUT cases require discussion.

p	$\beta(p)$	minimal NUT set
2	2	$\{0,1\}$
3	2	$\{0,1\}$
5	3	$\{0,1,2\}$
7	3	$\{0,1,3\}$
11	4	$\{0,1,2,5\}$
13	4	$\{0,1,2,6\}$
17	4	$\{0,1,3,7\}$
19	4	$\{0,1,3,8\}$
23	5	$\{0,1,2,4,11\}$
29	5	$\{0,1,2,5,12\}$
31	5	$\{0,1,2,5,13\}$
37	5	$\{0,1,3,7,15\}$
41	5	$\{0,1,2,6,29\}$
43	5	$\{0,1,3,7,18\}$
47	5	$\{0,1,3,7,20\}$
53	5	$\{0,1,3,8,37\}$
59	6	$\{0,1,2,3,20,28\}$
61	5	$\{0,1,3,21,55\}$

Table 3.3: $\beta(p)$ and an example minimal NUT set for primes $p = 2, \ldots, 61$.

4. $s = a + a + a + b$. In this case we note that $3a = a_1 + a_2 + a_3$ with a_1, a_2, a_3 not all equal. If a_1, a_2 and a_3 are distinct then s is a sum of at least three distinct elements, so we are in either case (1) or case (2) of Lemma 3.11. If the a_i are not distinct then, without loss of generality, $a_2 = a_3$. If $b \neq a_1$ and $b \neq a_2$ then we have three distinct elements and are fine. If not, we have two subcases:
   (a) If $b = a_1$ then $s = a_1 + a_1 + a_2 + a_2$ which is case (3) of Lemma 3.11.
   (b) If $b = a_2$ then we show that $s = (a + a) + (a + a_2) = (a_1 + a_2) + (a_2 + a_2)$ gives two distinct set-representations. We first note that $a + a_2 \neq a_2 + a_2$ since $a \neq a_2$. Next we note that $a + a \neq a_2 + a_2$ again since $a \neq a_2$.
5. $s = a + a + a + a$. As before, $3a = a_1 + a_2 + a_3$ with a_1, a_2, a_3 not all equal. If the a_i are distinct then s is the sum of at least three distinct elements and are fine. If two of the a_i are equal, we have without loss of generality, $a_2 = a_3$. Again $a \neq a_1$ and $a \neq a_2$. Thus we still have at least three distinct elements.

□

In analogy with the construction of Straus we have the following construction of a

NUT set. Let p be a prime. Then let

$$A = \{\pm[p/3], \pm[p/9], \ldots, \pm[p/3^i], \ldots, \pm 1, 0\}.$$

It is straightforward to see that A is NUT. This construction shows that $\beta(p) < 2\log_3 p + 1$.

3.3 Lower bounds for balanced and NUT sets

In Section 3.1 we gave Theorem 3.8 which showed that a set smaller than size $\log_4 p$ had an affine mapping to lie in $[0, p/2]$. Thus we have shown $\mu(p)$ and $\alpha(p)$ must be at least $\log_4 p$. Notice that a similar argument shows that if $|A| < \log_6 p$ then we can find an affine mapping to the interval $[0, p/3]$ and thus $\beta(p) > \log_6 p$. In [BDS76] the authors show that $\alpha(p) > \log_2 p$ and adapting their approach we can show that $\beta(p) > \log_3 p$. Our attempts to further adapt these arguments to give an improved lower bound for USF sets inspired Chapter 2.

To begin let $A = \{a_1, \ldots, a_n\}$ be a balanced set. We can associate to A a digraph, G, whose vertices are $[n]$ and for each i we consider some equation of the form $2a_i = a_j + a_k$, $j \neq k$, and add directed edges $i \to j$ and $i \to k$. We note that at least one equation of this form must exist since A is balanced. If more than one exists then we have a choice and thus there may be more than one digraph we can associate to a balanced set, A. Similarly, if A is NUT then we construct a (possibly multi) digraph G as follows. If $2a_i = a_j + a_k$ we add edges as before and if no such representation exists then we have $3a_i = a_j + a_k + a_\ell$ where $i \notin \{j, k, \ell\}$, and we add (possibly multi) edges $i \to j$, $i \to k$ and $i \to \ell$. Note that we have avoided loops since if $3a_i = a_i + a_j + a_k$ this simplifies to $2a_i = a_j + a_k$. We independently developed this approach and made several of the following conclusions, but it seems to have originated in [BDS76] and was rediscovered in [NQ08].

The first key observation is that if A is a minimal, under inclusion, balanced (respectively NUT) set then an associated digraph, G, is strongly connected. This is true regardless of the choice of G. This follows as if there existed a proper subset of vertices, U, with no outgoing edges then $B = \{a_i \ : \ i \in U\}$ would be balanced (respectively

NUT) and a proper subset of A. Next, let M be the adjacency matrix of G, i.e. $m_{i,j}$ is the number of directed edges from i to j. Let D be the degree matrix of G. That is, D is a diagonal matrix with $d_{i,i}$ equal to the out-degree of vertex i. Then, in analogy with the case for simple graphs, let $L = D - M$ be the Laplacian matrix of G. If A is balanced then L has diagonal elements equal to 2. If A is NUT then the diagonal elements are either 2 or 3.

Recalling that $n = |A|$, we next want to show that $\text{rank}_{\mathbb{Q}}(L) = n - 1$. Suppose $x \in \mathbb{R}^n$ is not a multiple of the all ones vector and $Lx = 0$. Then $U = \{i : x_i = \min(x)\}$ is a proper subset of the vertices of G. Since G is strongly connected there exists an edge $i \to j$ with $i \in U$ and $j \notin U$. Then we have either $2x_i = x_j + x_k$ for some k or $3x_i = x_j + x_k + x_\ell$ for some k, ℓ but in either case as $x_i < x_j$ we have a contradiction to the minimality of x_i. Thus $\text{rank}_{\mathbb{R}}(L) = n - 1$ and consequently $\text{rank}_{\mathbb{Q}}(L) = n - 1$. Next if $x \in \mathbb{F}_p^n$ is such that $x_i = a_i$ we see that over $\mathbb{F}_p$ we have $Lx = 0$. Since x is not a multiple of the all ones vector we have $\text{rank}_{\mathbb{F}_p}(L) \leq n-2$. Thus L has a $(n-1) \times (n-1)$ submatrix with non-zero determinant and p divides this determinant. For $1 \leq i, j \leq n$ let $L_{i,j}$ be the $(n-1) \times (n-1)$ submatrix formed by deleting row i and column j. By the directed matrix tree theorem [Cha82], $\det(L_{i,j})$ is the number of oriented spanning trees rooted at i and is independent of j. We have established there exists some i so that $p | \det(L_{i,i})$ and thus

$$p \leq \det(L_{i,i}). \tag{3.2}$$

Now if A is balanced note that $L_{i,i}$ has 2 on the diagonal and each row has at most two -1's off the diagonal. We begin by noting that if we apply the Gershgorin circle theorem we get the bound $\det(L_{i,i}) \leq 4^{n-1}$ and obtain $n > \log_4 p + 1$. If we apply Hadamard's inequality [Had93] then as $\sqrt{2^2 + (-1)^2 + (-1)^2} = \sqrt{6}$ we have $n > \log_{\sqrt{6}} p + 1$. However, in [Sch78], Schinzel proved the following upper bound for the determinant of real matrices:

$$\det(A) \leq \prod_{i=1}^{n} \max \left\{ \sum_{\substack{1 \leq j \leq n \\ a_{i,j} > 0}} a_{i,j}, \sum_{\substack{1 \leq j \leq n \\ a_{i,j} < 0}} a_{i,j} \right\}. \tag{3.3}$$

Applying equation (3.3) to $L_{i,i}$ we have $p < 2^{n-1}$ and thus $n > \log_2 p + 1$. So $\alpha(p) > \log_2 p + 1$.

If A is NUT then the above argument goes through except that $L_{i,i}$ has diagonal entries at most 3 and applying equation (3.3) shows that $\beta(p) > \log_3 p + 1$. We summarize this in Theorem 3.17 below.

Theorem 3.17. *For all primes p,*

$$\alpha(p) > \log_2 p + 1$$

and

$$\beta(p) > \log_3 p + 1.$$

□

3.4 Tight bounds for balanced sets and improved bounds for NUT sets

In Section 3.1, we described a construction due to Straus [Str76] that showed $\alpha(p) < 2\log_2 p + 1$. In the same paper, Straus gave a more complicated construction of a symmetrical balanced set that showed for any $\varepsilon > 0$ and p sufficiently large,

$$\alpha(p) < (2 + \varepsilon)\log_3 p + 1.$$

In [BDS76], this upper bound was improved to $\alpha(p) \leq (2 + o(1))\log_3 p$. Both these constructions were symmetrical since their goal was to find sets with no unique differences. In [Ned09] the authors show via algorithmic construction that

$$\alpha(p) < (1 + o(1))\log_2 p. \tag{3.4}$$

The same authors in [Ned12] use an improved form of Schinzel's inequality found in [JN80] to modestly improve the lower bound for balanced sets to $\alpha(p) > \log_2 p + 1 + \log_2(4/3) \approx \log_2 p + 1.41$. Together with equation (3.4) this establishes pretty tight bounds for balanced sets. Using equation (3.4) and the fact that for any balanced set A, $|A + A| \leq \frac{1}{2}|A|^2$ we have

Theorem 3.18. *For prime $p > 2$, $\mu(p) < (1/2 + o(1)) \log_2^2 p$.* □

In [Jan07] the authors study a generalized notion of balanced. If $v = (v_1, \ldots, v_k) \in \mathbb{F}_p^k$ is a vector of coefficients and $A \subseteq \mathbb{F}_p$ one can ask when for all $(a_1, \ldots, a_k) \in A^k$ whether $\sum_{i=1}^k v_i a_i$ has multiple representations as a v-weighted sum. For example if $v = (1, 1)$ then this describes balanced sets. Again, we stress that they consider $a + b = b + a$ as distinct representations. They define

Definition 3.19. *For $v \in \mathbb{F}_p^k$ Let $f(v, p)$ be the size of the minimal, nonempty, set $A \subseteq \mathbb{F}_p$ such that for all $(a_1, \ldots, a_k) \in A^k$, the sum $\sum_{i=1}^k v_i a_i$ has at least two representations.*

They prove the following theorem by generalizing the construction in [Str76].

Theorem 3.20. *Let $v \in \mathbb{F}_p^k$. For all $\varepsilon > 0$, $k \geq 2$ and every prime $p > p_\varepsilon$ we have*

$$f(v, p) < \left(\frac{1 + \varepsilon}{\log(2k - 1)} \right) \log p + 3. \tag{3.5}$$

□

Equation (3.20) applied to $v = (1, 1, 1)$ shows the following:

Corollary 3.21. *For all $\varepsilon > 0$ and p sufficiently large, $\beta(p) < \left(\frac{2+\varepsilon}{\log 5} \right) \log p + 3$.*

Note that $2/\log 5 < 1/\log 2$, thus for p sufficiently large we have $\beta(p) < \alpha(p)$. Thus we can improve Theorem 3.18 to the following.

Theorem 3.22. *For all $\varepsilon > 0$ and p sufficiently large, $\mu(p) < \left(\frac{2+\varepsilon}{\log^2 5} \right) \log^2 p + O(\log p)$.*

□

We suspect that, as with balanced sets, the lower bound for NUT sets should be close to tight and thus we suspect Theorem 3.22 can be improved to $\mu(p) < (1+\varepsilon) \log_3^2 p$. We will give some experimental evidence for this and discuss the notion of *regular* balanced and NUT sets in Section 3.5.

3.5 Regular balanced and NUT sets

As noted in Section 3.3, to a balanced or NUT set A of size n we can associate a digraph G. In the case of balanced sets, such digraphs are 2-out regular. For a NUT set we can associated a digraph G for which each vertex has out-degree 2 or 3. We contrast the following two constructions. In Section 3.2 we gave the construction of balanced set $A = \{\pm[p/2], \pm[p/4], \ldots, \pm[p/2^i], \ldots, \pm 1, 0\}$. We noted that $2[p/2^i] = [p/2^{i-1}] + \varepsilon$ where $\varepsilon \in \{0, -1\}$. We can associate a digraph to this construction. Observe that the vertices corresponding to $-1, 0, 1 \in \mathbb{F}_p$ together have larger average in-degree than the other vertices. Certainly, we do not expect that we can associate a digraph to this set where all the in-degrees are the same. On the other hand, in Section 3.2, we gave for a Mersenne prime $p = 2^q - 1$ the construction $A = \langle -2 \rangle$. If we write $A = \{a_1, \ldots, a_n\}$ where $a_i = (-2)^i$ and $n = 2q$ then we observed that $2a_i = a_{i+1} + a_{i+2}$ with indices modulo n. Thus we can associate a digraph G on vertices $[n]$ where $i \to i + 1$ and $i \to i + 2$ again taking vertices modulo n. This digraph is 2-regular. This inspires the following definition.

Definition 3.23. *We say a balanced set, $A \subseteq \mathbb{F}_p$ is regular if there is an associated digraph, G, which is 2-regular.*

Definition 3.24. *We say a NUT set, $A \subseteq \mathbb{F}_p$ is regular if there is an associated digraph, G, which is 3-regular.*

We of course have the same lower bounds for regular balanced and regular NUT sets as established previously. The goal of this section is to provide experimental evidence that these lower bounds are close to tight even for this constrained class of sets.

3.5.1 Circulant matrices and balanced and NUT subgroups

Definition 3.25. *We say a digraph, G, on vertices $[n]$ is circulant if there exists a set $S \subseteq [n-1]$ such that $i \to i + s$ if and only if $s \in S$. The sum is taken modulo n. We denote this graph $C_n(S)$.*

It is straightforward to see that $C_n(S)$ is strongly connected if and only if $\gcd(S)$ and

n are relatively prime. Note that the construction of a balanced set from a Mersenne prime, $p = 2^q - 1$, can be associated to the circulant digraph $C_{2q}(\{1, 2\})$. Furthermore, note that the balanced set was a multiplicative subgroup.

Lemma 3.26. *Let A be a subgroup of $\mathbb{F}_p^*$ of size n. Then A is balanced if and only if A has a generator g which is the root of $-2 + x^i + x^d$ with $1 \leq i < d < n$. In this case we can associate to A the digraph $C_n(\{i, d\})$.*

Proof. If g is a root of $-2 + x^i + x^d$ with order $n > d$ then $A = \langle g \rangle$ has order n and if $a = g^j \in A$ then $2a = g^{j+i} + g^{j+d}$ with exponents taken modulo n. Now if A is a subgroup of $\mathbb{F}_p^*$ of size n which is balanced, then $1 \in A$ is twice set-represented. This means $2 = g^i + g^d$ for some $1 \leq i < d < k$. This of course means g is a root of $-2 + x^i + x^d$. □

In order to allow for NUT sets with sums of the form $3a = 2b + c$ we extend our notions to multi-digraphs.

Definition 3.27. *Let n be a positive integer. Let S be a multiset of elements from $[n]$. We say a multi-digraph, G, on vertices $[n]$ is multi-circulant if for all $i \in [n]$, the edge $i \to i + s$ has multiplicity equal to the multiplicity of s in S.*

When it is clear that S is a multiset, we will expand our notation and denote this graph $C_n(S)$ as in the case of simple digraphs. The proof of the following lemma is nearly identical to the proof of Lemma 3.26.

Lemma 3.28. *Let A be a subgroup of $\mathbb{F}_p^*$ of size n. Then A is NUT if and only if A has a generator g which is the root of $-3 + x^i + x^j + x^d$ with $1 \leq i \leq j \leq d < n$ and i, j, d not all equal. We can associate to A the digraph $C_n(\{i, j, d\})$.* □

Lemmas 3.26 and Theorem 3.17 combine to imply the following. Let $f(x) = -2 + x^i + x^d$ with $1 \leq i < d$. Let $g(x) = x^n - 1$ with $n < \log_2 p$. Then $h(x) = \gcd(f(x), g(x))$ has no roots in $\mathbb{F}_p$ except for 1. This follows for if $h(x)$ had a non-trivial root in $\mathbb{F}_p$ it would generate a balanced subgroup of size less than $\log_2 p$ contradicting Theorem 3.17. Similarly we cannot have $f(x) = -3 + x^i + x^j + x^d$ with $1 \leq i \leq j \leq d < n$ not all equal and $g(x) = x^n - 1$ with $n < \log_3 p$ have any non-trivial roots in common.

Let A be a minimal balanced (respectively NUT) set in $\mathbb{F}_p$. Let G be an associated digraph (possibly a multi-graph if A is NUT). We observed in Section 3.3 that if G is strongly connected and L is the Laplacian matrix of G then $\text{rank}_{\mathbb{Q}}(L) = n - 1$ and $\text{rank}_{\mathbb{F}_p}(L) \leq n - 2$. In the case that A is a balanced subgroup of $\mathbb{F}_p^*$ we can associate a circulant digraph, G. The Laplacian matrix, L, is a circulant matrix. That is each row is a cyclic rotation of the previous row. The number of oriented spanning trees rooted at any vertex is independent of the choice of vertex by symmetry and equal to $\det(L_{1,1})$. The number of oriented spanning trees in circulant digraphs is well studied. See for example [LPW01, McK83]. It is shown in [LPW01] that amongst all 2-regular circulant digraphs, the digraph with the maximal number of oriented spanning trees is $C_n(\{1, 2\})$ which has rooted at any given vertex $\left\lfloor \frac{2^n+1}{3} \right\rfloor$ oriented spanning trees. See [Slob]. We can similarly find regular balanced NUT sets.

Example

Let $f(x) = -3 + 2x + x^2 = (x - 1)(x + 3)$. A root of f in $\mathbb{F}_p$ generates a NUT subgroup $A = \{a_1, \ldots, a_n\}$ where $3a_i = 2a_{i+1} + a_{i+2}$. So we seek a prime for which the element -3 has small multiplicative order. For example, $3^{71} - 1 = 2p$ where p is the 112-bit prime 3754733257489862401973357979128773. Then $A = \langle -3 \rangle$ is a 142 element regular NUT subgroup. Thus $\mu(p) \leq \binom{142}{2} = 10011$.

3.5.2 Experimentally finding small regular balanced and NUT sets

Ideally we would like for a given prime p to be able to show that there exists small regular balanced and NUT subsets. In this case small means close to $\log_2 p$ and $\log_3 p$ respectively. Short of this goal we will give experimental evidence that for a given n there exist regular balanced, respectively NUT sets of size n for primes close to 2^n respectively, 3^n. We define the following.

Definition 3.29. *We say a permutation $\pi \in S_n$ is a ménage permutation if for all $i \in [n]$, $\pi(i) \notin \{i, i + 1\}$ with indices taken modulo n.*

Thus a ménage permutation is a derangement with the added requirement that

$\pi(i) \neq i+1$. These arise in studying the famous ménage problem [Li15, KR46]. In [KR46], it is shown that the number of such permutations is asymptotic to $n!/e^2$.

Definition 3.30. *We say an $n \times n$ matrix, M, is a ménage matrix if there exists a ménage permutation $\pi \in S_n$ such that for all i, $m_{i,i} = 2$, $m_{i,i+1} = m_{i,\pi(i)} = -1$. All other entries in M are zero.*

Thus if M is an order n ménage matrix it is the Laplacian matrix of the digraph, G, on vertices $[n]$ where $i \to i+1$ and $i \to \pi(i)$ for each i. The fact that the digraph is 2-regular follows from the definition of ménage permutations. Further, since $i \to i+1$ for all i, G contains a directed Hamiltonian cycle and thus is strongly connected. By the regularity of G the number of oriented spanning trees at any vertex is the same and thus the determinant of any $(n-1)\times(n-1)$ submatrix is the same. So let $t = \det(M_{1,1})$. So if a prime p divides t then a nontrivial solution to $Mx = 0$ gives a candidate regular balanced subset in $\mathbb{F}_p$. Note that we do not immediately have $A = \{x_i : i \in [n]\}$ as a regular balanced set since we may have $x_i = x_j$. However, if every coordinate appears at most twice A will be balanced.

Definition 3.31. *For $x \in \mathbb{F}_p^k$ let $U(x) = \{x_i \ : \ i \in n\}$ be the set containing the coordinates of x.*

We use "U" because we have "uniqued" the coordinates of x.

Theorem 3.32. *Let L be the Laplacian matrix of a digraph for which all vertices have out-degree two. For a prime p, let $x \in \mathbb{F}_p^n$ be such that $Lx = 0$. Suppose that each $a \in \mathbb{F}_p$ appears at most twice as a coordinate of x. Then $U(x)$ is balanced.*

Proof. Suppose some element, a, appears as a coordinate twice in x. It suffices to show that we can remove that coordinate and doubles are still covered. Without loss of generality, $a = x_1 = x_2$. For $i > 2$, $x_i \neq a$. We need to show that we can remove x_2 and still have non unique doubles. We have two cases. Case one: If we remove x_2 can we still represent $2x_1$ as a sum in a different way? If $2x_1 = x_2 + x_d$ with $d > 2$ then $x_d = x_1$, but then x_1 appears at least three times. So $2x_1 = x_i + x_d$ with $i, d > 2$. So removing x_2 does not prevent covering $2a$. Case two: If we remove x_2 can we still represent $2x_i$

for $i \geq 3$ as a sum in a different way? It suffices to check that $2x_3$ is covered. We cannot have $2x_3 = x_1 + x_2$ as then $x_3 = x_1 = x_2$. So suppose $2x_3 = x_2 + x_d$. But $x_1 = x_2$. So $2x_3 = x_1 + x_d$. So removing x_2 does not prevent covering $2x_3$. □

A similar statement holds for NUT sets.

Theorem 3.33. *Let L be the Laplacian matrix of a digraph for which all vertices have out-degree 2 or 3. For a prime $p > 3$, let $x \in \mathbb{F}_p^n$ be such that $Lx = 0$. Suppose that each $a \in \mathbb{F}_p$ appears at most twice as a coordinate of x. Then $U(x)$ is NUT.*

Proof. Suppose without loss of generality that $a = x_1 = x_2$. We showed in the proof of Theorem 3.32 that removing x_2 does not cause equations of the form $2x_i = x_j + x_k$ to fail. If we have $3x_i = x_j + x_k + x_\ell$ and $x_j = x_k = a$ then $3x_i = 2x_1 + x_\ell$ satisfies the requirement in Definition 3.24. If we have $3x_i = 2x_j + x_k$ with $j \neq k$ and $x_j = x_k$ then $3x_i = 3x_j$ which, as $p > 3$, implies $x_i = x_j = x_k$ so the element appears three times. □

Note that if M is a ménage permutation and $Mx = 0$ in $\mathbb{F}_p$ it is not necessarily the case that $U(x)$ is a regular balanced set since uniquing may have destroyed regularity. However, as noted if the coordinates of x are distinct then they do form a regular balanced set.

In [Bal10] generalizations of ménage permutations are studied. We can extend the relationship described above.

Definition 3.34. *We say $\pi \in S_n$ is a ménage-3 permutation if for all $i \in [n]$, $\pi(i) \notin \{i, i+1, i+2\}$ with indices taken modulo n. We say an $n \times n$ matrix, M, is a ménage-3 matrix if there exists a ménage-3 permutation $\pi \in S_n$ such that for all i, $m_{i,i} = 2$ and $m_{i,i+1} = m_{i,i+2} = m_{i,\pi(i)} = -1$. All other entries in M are zero.*

We employed the following experimental approach to finding small regular balanced sets. We chose a value n and for several iterations $i = 1, 2, ..$ we choose M to be a random ménage matrix of order n. We did this by choosing the accompanying ménage permutation uniformly at random. We compute $t_i = \det(M_{1,1})$. This gives us a sequence $t_1, t_2, \ldots$ of integers which from Schinzel's inequality given in equation (3.3) are at most

2^{n-1}. We would like to find the largest prime dividing this list. As n increases, factoring all the t_i is quickly infeasible. However, since we are interested in the largest prime dividing this list this is unnecessary. Under the hopefully mild assumption that these numbers behave like "random" integers if we have $\Omega(n)$ such numbers a large prime should divide one of them. So for each t_i we employed a few iterations of the Pollard rho algorithm [Pol75] to prune off small primes. Once Pollard rho found no more prime factors if a pseudoprimality test said the remaining factor was prime we recorded it. Ultimately, we chose the largest such prime we found and solved the resulting system of equations to test for uniqueness. In so doing we were able to find very large primes with small regular balanced sets. An analogous approach using ménage-3 matrices found large primes with small regular NUT sets. See Table 3.4 and Table 3.5 for some results for $n = 100, 200, 1000, 2000, 4000$ of regular balanced (respectively NUT) sets for which the size of the corresponding field is close to 2^n (respectively 3^n). For example, we found when $n = 4000$ that for a prime $p \approx 2.721 \times 10^{1904}$ there is a regular balanced NUT set in $\mathbb{F}_p$ of size $n = 4000$. In this case $\log_3 p/4000 \approx 0.997878$.

For completeness we give an example of a large balanced NUT set found via computer experimentation. The smallest example given in Table 3.5 is for $n = 100$ and $p = 1223346310053449250273350725461948655697480950 3$. Since A is given as a solution to $Lx = 0$ where L is a ménage-3 matrix it suffices to give the ménage-3 permutation, π, and the interested reader can compute A for themselves and confirm that it is indeed a regular NUT set. We note that for this particular L the rank over $\mathbb{F}_p$ is $n-2$ and thus up to affine transformation the resulting NUT set is unique. Writing π in one-line notation, i.e. $\pi(1), \pi(2), \ldots$ we have $\pi = (14, 55, 46, 7, 15, 23, 17, 72, 99, 1, 25, 73, 18, 71, 41, 93, 29, 7$ $5, 52, 83, 57, 45, 95, 12, 68, 90, 96, 84, 94, 38, 50, 24, 79, 6, 98, 92, 4, 78, 66, 26, 60, 51, 91, 33, 6$ $9, 21, 89, 28, 9, 77, 54, 32, 87, 58, 88, 5, 16, 61, 35, 43, 31, 27, 82, 85, 53, 76, 47, 100, 74, 40, 36,$ $42, 81, 3, 39, 86, 80, 37, 30, 67, 49, 19, 97, 44, 64, 10, 8, 48, 2, 20, 13, 63, 11, 62, 70, 65, 59, 22, 5$ $6, 34)$.

The experimental evidence above suggests the following conjectures.

Conjecture 3.35. *For all $\varepsilon > 0$, there are infinitely many primes, p, such that there exists a regular balanced set $A \subseteq \mathbb{F}_p$ with $|A| < (1 + \varepsilon) \log_2 p$.*

n	p approximation	$\log_2 p/n$
100	1.036×10^{27}	0.897441
200	1.329×10^{57}	0.948807
1000	1.524×10^{296}	0.983899
2000	7.064×10^{596}	0.991345
4000	2.942×10^{1197}	0.994476

Table 3.4: Regular balanced sets

n	p approximation	$\log_3 p/n$
100	1.223×10^{46}	0.96595
200	2.734×10^{93}	0.979174
1000	3.072×10^{474}	0.99448
2000	1.411×10^{951}	0.996759
4000	2.721×10^{1904}	0.997878

Table 3.5: Regular NUT sets

Conjecture 3.36. *For all $\varepsilon > 0$, there are infinitely many primes, p, such that there exists a regular NUT set $A \subseteq \mathbb{F}_p$ with $|A| < (1+\varepsilon)\log_3 p$.*

To give a sense of the distribution of $\det(M_{1,1})$ for random ménage and ménage-3 matrices we performed the following experiment. We set $n = 20$ and for 10^7 random trials we computed $\det(M_{1,1})$ for random ménage and ménage-3 matrices. We normalized these values by dividing by 2^{n-1} and 3^{n-1} for ménage and ménage 3 matrices respectively. A histogram of results can be found in Figures 3.2 and 3.3. Given that the number of ménage matrices is asymptotic to $n!/e^2 \gg 2^n$ and the histograms do not show any artifacts we strengthen Conjectures 3.35 and 3.36 to Conjectures 3.37 and 3.38 below.

Conjecture 3.37. *For all $\varepsilon > 0$, for p sufficiently large, there exists a regular balanced set $A \subseteq \mathbb{F}_p$ with $|A| < (1+\varepsilon)\log_2 p$.*

Conjecture 3.38. *For all $\varepsilon > 0$, for p sufficiently large, there exists a regular NUT set $A \subseteq \mathbb{F}_p$ with $|A| < (1+\varepsilon)\log_3 p$.*

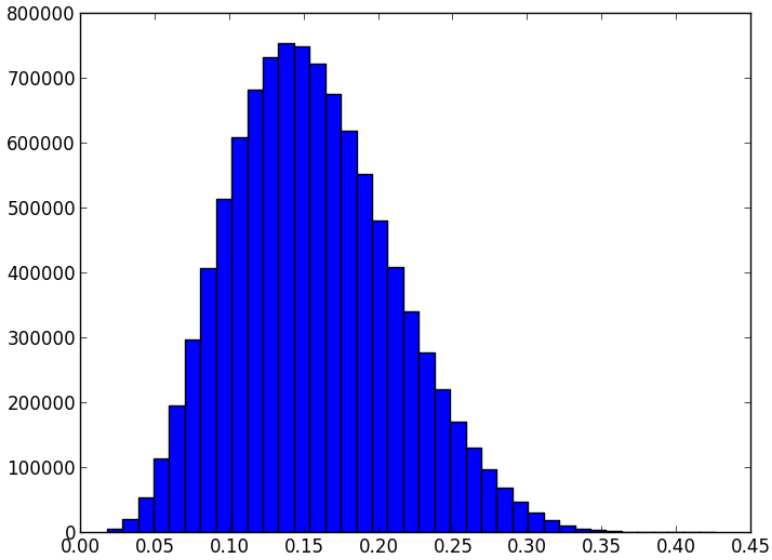


Figure 3.2: A histogram of $\det(M_{1,1})/2^{n-1}$ for $n = 20$ for 10^7 random ménage matrices, M.

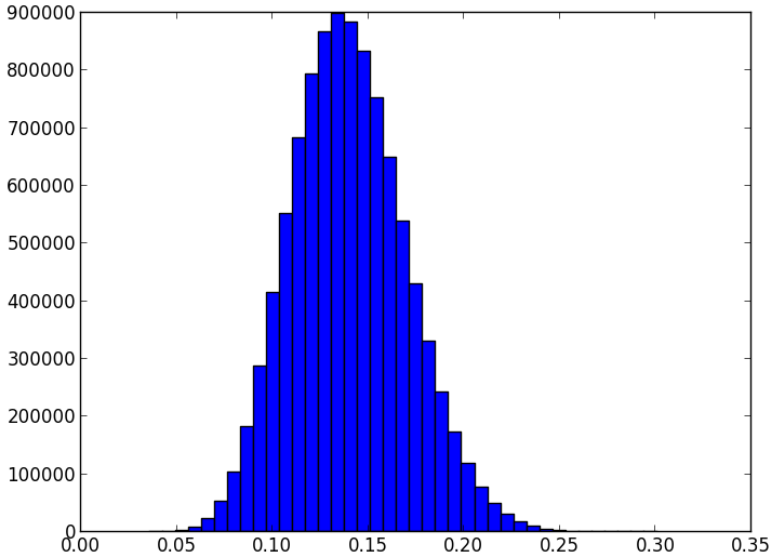


Figure 3.3: A histogram of $\det(M_{1,1})/3^{n-1}$ for $n = 20$ for 10^7 random ménage-3 matrices, M.

3.6 USF sets in $\mathbb{Z}/n\mathbb{Z}$ for composite n

We can extend our definition of μ to composite n.

Definition 3.39. *If $n \geq 2$ is a positive integer let $\mu(n)$ be the minimum size of a USF set in $\mathbb{Z}/n\mathbb{Z}$. We set $\mu(1) = \mu(2) = \infty$.*

The first few values of $\mu(n)$ for $n \geq 3$ can be found in Table 3.6.

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\mu(n)$	3	4	4	3	5	4	3	4	7	3	7	5	3	4	8	3	9	4

Table 3.6: $\mu(n)$ for small n.

As $\mathbb{Z}/n\mathbb{Z}$ is a subgroup of $\mathbb{Z}/mn\mathbb{Z}$, it is clear that for any $m, n \in \mathbb{N}$ that $\mu(mn) \leq \min(\mu(m), \mu(n))$. Experimentally, we seem to have equality in this statement. So we make the following conjecture.

Conjecture 3.40. *For all $n > 1$, $\mu(n) = \min_{p|n} \mu(p)$.*

The analogous statement holds for balanced sets. In [NQ08], the authors extend α to the integers and prove the following theorem.

Theorem 3.41. *Let $n > 1$. We have $\alpha(n) = \min_{p|n} \alpha(p)$.* □

Their main tool is the following lemma.

Lemma 3.42. *Let $m|n$ and let S be a balanced set modulo n that is minimal under inclusion. Then $S \mod m$ is balanced or consists of a single point.* □

We note that the analogous statement is not true for USF sets. Let $n = 22$ and $A = \{0, 1, 2, 6, 9, 11, 12, 13, 17, 20\}$. One can check that A is USF in $\mathbb{Z}/n\mathbb{Z}$ and is a minimal USF set under inclusion. Note that $\mu(22) = \mu(11) = 7$ so A is not minimal. We can write $A = B \cup (B + 11)$ where $B = \{0, 1, 2, 6, 9\}$. So $A \mod 11 = B$ and B is not USF in $\mathbb{Z}/11\mathbb{Z}$.

3.7 Unique Differences, Products and Quotients

We have already remarked, but we reiterate that the question of when a set is unique difference free (UDF) was studied in [Str76]. To establish notation we will let $\mu_-(p)$ be

the size of the smallest UDF set in $\mathbb{F}_p$. We have observed that $\mu_-(p) = O(\log p)$. We motivated our study of USF sets with a geometrical example and sets with no unique differences have a similar geometric motivation. Suppose we have p uniformly spaced points on a circle. Let A be a subset of these points. For every pair of points draw the line segment connecting them. Then A being UDF is equivalent to the statement that for every line segment there is another line segment of the same length. Thus we can interpret UDF as "unique distance free."

Definition 3.43. *We call a nonempty set $A \subseteq \mathbb{F}_p$ unique product free (UPF) if every element of its product set is at least twice covered. That is for all $t \in A \cdot A$ there exist a, b, c, d with $\{a, b\} \neq \{c, d\}$ such that $t = ab = cd$.*

It is straightforward to see that if A is UPF then $A \cup \{0\}$ is UPF and also if $B \cup \{0\}$ is UPF then B is UPF. So we can safely ignore the element 0. In the case of quotients we just require 0 not to be present.

Definition 3.44. *We call a nonempty set $A \subseteq \mathbb{F}_p^*$ unique quotient free (UQF) if every element of its quotient set is at least twice covered. That is for all $t \in A/A$ there exist a, b, c, d with $\{a, b\} \neq \{c, d\}$ such that $t = a/b = c/d$.*

Definition 3.45. *For any prime p, let $\mu^*(n)$ be the minimal size of a UPF subset of $\mathbb{F}_p$ and let $\mu^*_-(n)$ be the minimal size of a UQF subset of $\mathbb{F}_p$.*

Let A be a UPF set. As observed above we can remove the element 0 if present and preserve the UPF property. Thus a minimal UPF set does not contain 0. Since $\mathbb{F}_p^*$ is a cyclic group of order $p-1$, the set A corresponds to a USF set in $\mathbb{Z}/(p-1)\mathbb{Z}$. Similarly, a UQF set in $\mathbb{F}_p$ corresponds to a UDF set in $\mathbb{Z}/(p-1)\mathbb{Z}$. Thus we have

$$\mu^*(p) = \mu(p-1)$$

and

$$\mu^*_-(p) = \mu_-(p-1).$$

3.8 Conclusion and Main Open Questions

We have shown that $\mu(p) = O(\log^2 p)$ disproving the conjecture that $\mu(p) = \Theta(\sqrt{p})$. We have the lower bound $\mu(p) = \Omega(\log p)$. Resolving this gap is an open question. We conjectured in Section 3.5 that regular balanced and NUT sets exist of size close to $\log_2 p$ and $\log_3 p$ respectively. Finally, for composite n our main open question is to resolve Conjecture 3.40. That is, for all positive integers, n, do we have $\mu(n) = \min_{p|n} \mu(p)$?

Chapter 4

Affine fall k-colorings of the Hamming cube

This chapter is joint work with Keith Frankston.

4.1 Introduction and Background

Let G be a simple graph. That is an undirected graph without loops or multiple edges. We denote the vertices of G by $V(G)$ and the edges by $E(G)$.

Definition 4.1. *For a positive integer k, a k-coloring of G is a function $C : V(G) \to S$ where $|S| = k$.*

Clearly, the number of k-colorings for a graph on n vertices is k^n.

Definition 4.2. *A coloring $C : V(G) \to S$ is proper if for all $u, v \in V(G)$ if $u \sim v$ then $C(u) \neq C(v)$.*

The following terminology originated in [DHH$^+$00].

Definition 4.3. *We say a k-coloring $C : V(G) \to S$ is a fall k-coloring if C is a proper k-coloring and for every vertex v and for every color $j \neq C(v)$ there exists a vertex u such that $u \sim v$ and $C(u) = j$.*

That is every vertex is adjacent to a vertex in each of the other $k-1$ color classes. Equivalently, $C : V(G) \to S$ is a fall k-coloring if $|S| = k$ and for all v, $\{C(u) : u \sim v\} = S \setminus \{C(v)\}$. The name is seasonally inspired, as each vertex has a maximally colorful view. Indeed [DHH$^+$00] has the following definition.

Definition 4.4. *We say a vertex, v, in a k-coloring is colorful if it is adjacent to at least one vertex in each of the other color classes.*

In this chapter the graph of interest is the Hamming cube of dimension n (sometimes called n-cube or cube graph) which we denote by Q_n. This is the graph whose vertices are the 2^n bit strings of length n and two vertices are adjacent if and only if they differ in exactly one bit.

If $C : V(G) \to [k]$ is a k-coloring we denote by $V_i = \{v \in V(G) : C(v) = i\}$ the i-th color class. Notice that for a proper coloring each color class, V_i, is an independent set. If C is a fall k-coloring then notice that V_i is a dominating set because if $u \notin V_i$ then by Definition 4.3 there must be some $v \in V_i$ such that $u \sim v$. Thus the V_i are independent, dominating and partition $V(G)$. Consequently, some authors call fall k-colorings "idomatic partitions" [GH13]. Since our main construction will be a coloring scheme based on trees it seems natural to use the "fall" terminology. Figure 4.1 shows a fall 4-coloring of Q_3.

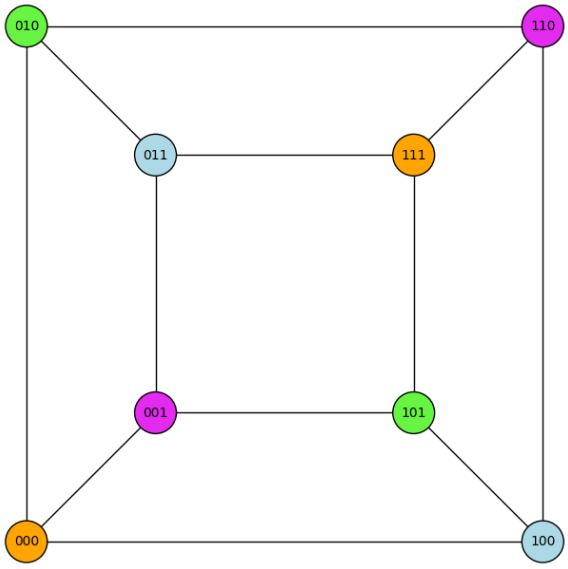


Figure 4.1: An affine fall 4-coloring of Q_3. Note that vertices of the same color are antipodal.

As an aside, we mention that the similar question of finding the domatic number of graphs, in particular the Hamming cube is well studied [Zel82, Zel91, Sloa, HHW88]. Here we set $\tau(n)$ to be the size of the largest domatic partition of Q_n, i.e. the maximum number of sets, V_i, such that the V_i partition $V(Q_n)$ and each V_i is dominating. These sets need not be independent. Zelinka [Zel82] showed that if $n = 2^a$ or $n = 2^a - 1$

then $\tau(Q_n) = 2^a$. Other values are mostly open [Sloa]. For example it is known that $\tau(Q_{10}) \in \{8, 9\}$, but which of these values is correct is an open question.

In [DHH+00] it was asked for which k do there exist Q_n with fall k-colorings? This question was answered by Laskar and Lyle in [LL09] who proved the following theorem.

Theorem 4.5. *If $k = 3$ then for any n, Q_n is not fall k-colorable. For $k \neq 3$, let a be the smallest integer such that $k \leq 2^a$. Then for $n \geq 2^a - 1$ there is a fall k-coloring of Q_n.* □

There is a natural identification of Q_n with the vector space $\mathbb{F}_2^n$. Let $e_1, \ldots, e_n$ be the standard basis vectors for $\mathbb{F}_2^n$. If $u, v \in \mathbb{F}_2^n$ then we say $u \sim v$ if and only if there exists some e_i so that $x + e_i = y$. In this context, we may ask algebraic questions about the color classes. In particular we are interested in colorings where each color class, V_i, is an affine subspace.

Definition 4.6. *We call a coloring $C : \mathbb{F}_2^n \to S$ affine if for each $i \in S$, the color class V_i is an affine subspace of $\mathbb{F}_2^n$.*

We note that if k is a power of 2 then constructing affine fall k-colorings of Q_n is straightforward.

Theorem 4.7. *If $k = 2^a$ and $n \geq k - 1$ then there is an affine fall k-coloring of Q_n.*

Proof. Let e_i be the i-th standard basis vector of $\mathbb{F}_2^n$. Construct a linear function $\phi : \mathbb{F}_2^n \to \mathbb{F}_2^a$ by ensuring that $\phi(\{e_1, \ldots, e_n\}) = \mathbb{F}_2^a \setminus \{0\}$. This is achievable provided $n \geq k - 1 = 2^a - 1$. Extend by linearity to define ϕ. Note that ϕ gives a proper coloring as if $u, v \in \mathbb{F}_2^n$ and $u \sim v$ we have some i such that $u + e_i = v$ and $\phi(v) = \phi(u + e_i) = \phi(u) + \phi(e_i) \neq \phi(u)$ as $\phi(e_i) \neq 0$. Furthermore, for any $v \in \mathbb{F}_2^n$ if $C(v) = x \in \mathbb{F}_2^a$, let $y \in \mathbb{F}_2^a$ be any color such that $y \neq x$. Then $x + y \neq 0$ and thus there is some i such that $\phi(e_i) = x + y$. Then let $u = v + e_i$ and we see that u is a neighbor of v and $\phi(v) = \phi(u + e_i) = \phi(u) + \phi(e_i) = x + (x + y) = y$. Thus, ϕ is an affine fall k-coloring of $\mathbb{F}_2^n$. □

In fact for $k = 2^a$ and $n \geq k - 1$, the linear map $\phi : \mathbb{F}_2^n \to \mathbb{F}_2^a$ given in the proof of Theorem 4.7 shows we can give a affine k-coloring of $\mathbb{F}_2^n$ such that each of the 2^a color

classes are parallel. For example, in Figure 4.1 each of the color classes, V_i, has size two and therefore are trivially dimension one affine subspaces. However, note that they are parallel. For each class $V_i = \{u, v\}$, u and v are antipodal. That is $u + v = (1, 1, 1)$.

When k is not a power of 2, the construction of Laskar and Lyle does not give an affine fall k-coloring. Our main result is that for k even and the same range of values of n as in Theorem 4.5 that Q_n has an affine fall k-coloring.

4.1.1 Recoloring graphs and a note about the case $k = 3$

This line of research began in an attempt to understand the properties of the recoloring graph of Q_n. At the time we were unaware of the result of Laskar and Lyle. For any graph G, and integer k, the k-recoloring graph, $R_k(G)$, is the simple graph whose vertices are proper k-colorings of G and colorings C_1 and C_2 are adjacent if and only if there is exactly one vertex on which they disagree. A fundamental question is whether $R_k(G)$ is connected. If so one can walk from any proper k-coloring, C_1, to any other proper k-coloring, C_2, by changing the color at one vertex at a time preserving a proper coloring at each step. A fall k-coloring of Q_n gives an isolated vertex (coloring) of $R_k(Q_n)$ since no vertex of Q_n can be recolored without colliding with a neighboring color class. Theorem 4.7 shows that if k is a power of 2 and n is sufficiently large, $R_k(Q_n)$ contains an isolated vertex. For $k = 3$, however, not only do fall 3-colorings of Q_n not exist for any n, but for all n, $R_3(Q_n)$ is connected [Gal03]. The result of Laskar and Lyle shows that for any $k \neq 3$ and n sufficiently large, $R_k(Q_n)$ contains an isolated vertex and therefore is not connected.

4.1.2 Integer linear programming and the special case $k = 5$

The results in this chapter were greatly aided by computer experimentation. In particular, the discovery of an affine fall 6-coloring of $\mathbb{F}_2^7$ motivated further investigation and generalizations. Fall k-colorings of Q_n can be modeled as solutions to an integer linear program (ILP). The following ILP was implemented using the Julia [BKSE12] programming language using the package JuMP [DHL17]. Through an academic license we used the solver Gurobi [GO18] to discover several colorings and eliminate others.

Let k and n be integers. We describe how the question of if there exists a fall k-coloring of Q_n can be modeled as a binary ILP. Define $k2^n$ binary variables, $x[v,i]$, where $v \in Q_n$ and $i \in [k]$ to indicate

$$x[v,i] = \begin{cases} 1, \text{ if } v \text{ is colored } i \text{ and} \\ 0, \text{ otherwise.} \end{cases}$$

We need to ensure that each vertex is assigned exactly one color. For each $v \in Q_n$ we have the linear constraint

$$\sum_{i=1}^{k} x[v,i] = 1.$$

To ensure that the coloring is proper we have for each u, v with $u \sim v$ and for each $i \in [k]$,

$$x[u,i] + x[v,i] \leq 1.$$

Finally, we want to ensure each vertex is colorful. Let d be the Hamming distance metric. That is for $u, v \in Q_n$, $d(u,v)$ is the number of bits at which they differ or, equivalently, the length of the shortest path from u to v. The fall k-coloring constraint is equivalent to the constraint that in each Hamming ball of radius one, each color appears at least once. Thus for each $v \in Q_n$ and each $i \in [k]$ we have the constraint

$$\sum_{d(u,v)\leq 1} x[u,i] \geq 1.$$

In this case there is no objective function that the solver seeks to optimize. Rather it just checks if the ILP is feasible.

Theorem 4.5 shows that for $k = 5$ there is a fall k-coloring of Q_n for $n \geq 7$. The fact that there is a fall 5-coloring of Q_6 was, we believe, previously unknown and discovered via our ILP implementation. To give a compact representation, if we order the 64 vertices of Q_6 lexicographically then the corresponding colors are

$$2,1,3,4,4,2,1,5,1,5,5,2,3,4,2,3,1,3,2,5,5,4,3,2,4,2,3,1,2,1,4,5,$$

$$5,4,1,2,1,3,2,4,2,3,4,5,5,2,3,1,3,2,4,3,2,5,5,1,5,1,2,4,4,3,1,2.$$

Note that this coloring is *not* affine. Indeed via ILP methods we know that there is no affine fall 5-coloring of Q_6.

4.2 Machinery

To begin we give a simple lower bound on the size of the color classes of a fall k-coloring.

Lemma 4.8. *For any fall k-coloring of Q_n each color class V_i must satisfy $|V_i| \geq 2^n/(n+1)$.*

Proof. As V_i is a dominating set each of the remaining $2^n - |V_i|$ vertices must be adjacent to at least one vertex in V_i. Since Q_n is n-regular we have $|V_i|n \geq 2^n - |V_i|$ which simplifies to $|V_i| \geq 2^n/(n+1)$. □

If $\delta(G)$ is the minimum degree of a vertex of G it is clear that for G to have a fall k-coloring, we must have $k \leq \delta(G) + 1$. Otherwise the minimal degree vertex cannot be colorful. Thus for Q_n the largest possible k for which Q_n conceivably has a fall k-coloring is $k = n + 1$.

Lemma 4.9. *Let $k \geq 2$ and let $n = k - 1$. Then Q_n has a fall k-coloring if and only if k is a power of 2.*

Proof. Suppose Q_n has a fall k-coloring where $k = n + 1$. Then by Lemma 4.8, $|V_i| \geq 2^n/(n+1) = 2^n/k$ for all i. So each color class is at least as large as the average size of the color classes. Thus $|V_i| = 2^n/k$ for all i. So k divides 2^n and thus k is a power of 2.

Conversely, if $k = 2^a$ we have already observed in Theorem 4.7 that there there is a linear map that gives an fall k-coloring provided $n \geq k - 1$. □

Our main result is that there are affine fall k-colorings for k even and n sufficiently large.

Theorem 4.10. *Let $k \geq 2$ be even. Let 2^a be the smallest power of 2 such that $2^a \geq k$. Then if $n \geq 2^a - 1$ there is an affine fall k-coloring of Q_n.*

To prove Theorem 4.10, we give a construction using an object similar to a parity decision tree [O'D14]. In our case we give a decision tree that uses parity to determine which branch to proceed along, however, each leaf is assigned a different value. Thus if the decision tree has k leaves it will give a k-coloring of Q_n.

Definition 4.11. *A parity coloring tree (PCT) is a tuple (T, k, n, h, C) where:*

- *T is a full binary decision tree. We call the non-leaves "decision nodes." The nodes of T are labelled with binary strings. The root of T is given the null string and for any decision node labelled t, its right child is labeled $t0$ and its left child is labelled $t1$. That is we append a 0 or a 1 depending on if the decision is to progress to the right or the left respectively.*
- *k is the number of leaves of T.*
- *n is the dimension on which the PCT classifies vectors.*
- *h is a function from the decision nodes of T to $\mathbb{F}_2^n$.*
- *C is the resulting coloring function. If L is the set of leaves of T then $C : \mathbb{F}_2^n \to L$.*

Let $\langle \cdot, \cdot \rangle$ be the dot product. If $t \in T$ is a decision node then we associate the decision vector $h(t) \in \mathbb{F}_2^n$. For any $v \in \mathbb{F}_2^n$ to be classified if $\langle h(t), v \rangle = 0$ then the decision node goes to child $t0$ and otherwise to $t1$.

In Figure 4.2 is a PCT that gives the affine fall 4-coloring of Q_3 shown in Figure 4.1. The decision nodes are in white and are labelled with their decision vector. Let (T, k, n, h, C) be this PCT. To illustrate, suppose we wish to color the vector $(1, 1, 1) \in \mathbb{F}_2^3$. Let r be the root of T. It is a decision node and we see that $h(r) = (0, 1, 1)$. We can compute $\langle (0, 1, 1), (1, 1, 1) \rangle = 0$. Thus we proceed to the right child which we denote with the binary string 0. We see that $h(0) = (1, 0, 1)$. Thus we compute $\langle (1, 0, 1), (1, 1, 1) \rangle = 0$. So we proceed to the right child which is the leaf 00 which we depict in orange. Thus $(1, 1, 1)$ is given the color 00 depicted in orange and we see that in Figure 4.1 the vertex $(1, 1, 1)$ is indeed orange. One can check that, for example, $v = (0, 0, 0)$ would also be colored orange. Thus $C((1, 1, 1)) = C((0, 0, 0)) = 00$.

Definition 4.12. *If T is a rooted tree and $v \in T$ is a decision node, let $T|_v$ be the subtree with root v containing v and all its descendants.*

Definition 4.13. *For a PCT (T, k, n, h, C) and any $t \in T$ let S be the sequence of decision nodes along the path from the root down to t. Let m be the height of t. If t is a decision node then the length of S is $m + 1$ and if t is a leaf the length of S is m. Define the decision set, $D(t)$, to be the set of decision vectors $\{h(v) \ : \ v \in S\}$. Define*

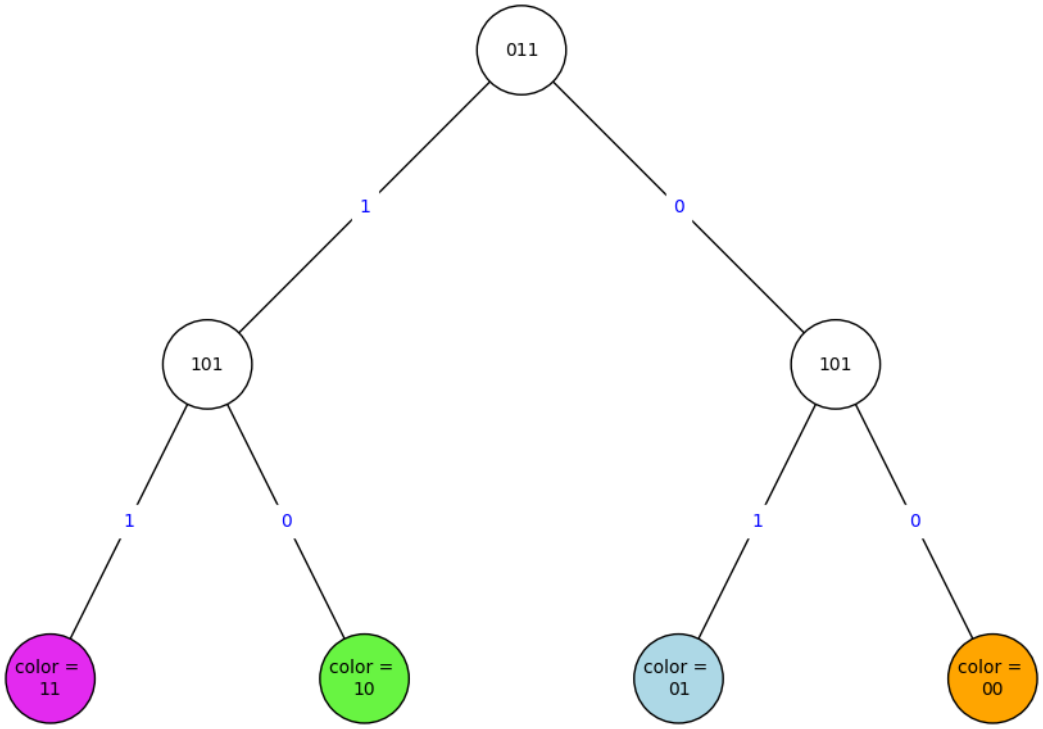


Figure 4.2: A PCT that gives the affine fall 4-coloring of Q_3 shown in Figure 4.1. The decision nodes are in white and are labelled by their decision vector. For example, the decision vector of the root is $(0, 1, 1)$.

the decision matrix, D_t to be the $m \times n$ matrix whose i-th row is the decision vector of the i-th element of S.

Thus the decision set of any leaf $\ell \in T$ are the vectors considered when classifying a vertex in Q_n that reaches ℓ. For example, in the PCT depicted in Figure 4.2, let $\ell = 00$ be the rightmost leaf. We see that $D(\ell) = \{(0, 1, 1), (1, 0, 1)\}$ and

$$D_\ell = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

In fact, for this example, we see that D_t is the same matrix for each of the two decision nodes below the root and and the four leaves.

Definition 4.14. *For a PCT (T, k, n, h, C) and $\ell \in T$ a leaf. We define the color class* $V_\ell = \{v \in \mathbb{F}_2^n \ : \ C(v) = \ell\}$.

For any leaf $\ell \in T$ we can identify ℓ with an element of $\mathbb{F}_2^m$ where m is the height of ℓ. With this identification, we see that $C(v) = \ell$ precisely when $D_\ell v = \ell$. Thus we

have $V_\ell = \{v \in \mathbb{F}_2^n \ : \ D_\ell v = \ell\}$. So it is clear that the color classes of a PCT are affine subspaces.

Definition 4.15. *We say a PCT (T, k, n, h, C) is degenerate if there is some color class that is unobtainable. That is there is some leaf ℓ such that $C(v) \neq \ell$ for all $v \in \mathbb{F}_2^n$. Equivalently, the PCT is degenerate if C is not onto the leaves of T. Otherwise we say the PCT is non-degenerate.*

Lemma 4.16. *For a PCT (T, k, n, h, C), if ℓ is a leaf at a height m and $D(\ell)$ is linearly independent then $|V_\ell| = 2^{n-m}$. Furthermore, if $D(\ell)$ is linearly dependent then C is not onto. Thus a PCT is degenerate if and only if for for some leaf ℓ, $D(\ell)$ is linearly dependent.*

Proof. Suppose ℓ is a leaf at height m and $D(\ell)$ is linearly independent. Then as $V_\ell = \{v \in \mathbb{F}_2^n \ : \ D_\ell v = \ell\}$, we have $|V_\ell| = 2^{n-m}$.

Next, suppose $D(\ell)$ is linearly dependent. Let t be the vertex of minimum height on the path from the root r to ℓ for which $h(t)$ is a linear combination of its ancestors decision vectors. Then for any v that we are classifying that reaches t, $\langle v, h(t)\rangle$ is uniquely determined. So some of t's descendants are unreachable. □

Definition 4.17. *We say a PCT (T, k, n, h, C) contains a zero column if for some leaf $\ell \in T$ the matrix D_ℓ contains a zero column.*

Equivalently, the PCT contains a zero column if for some leaf $\ell \in T$ all vectors in the decision set $D(\ell)$ are zero at a particular coordinate.

Theorem 4.18. *A PCT (T, k, n, h, C) gives a proper k-coloring using all k colors if and only if it is non-degenerate and does not contain a zero column.*

Proof. We showed in Lemma 4.16 that all k colors are used if and only if it is non-degenerate. Next, suppose that the PCT contains a zero column. Then for some leaf ℓ and coordinate i, $h(v)_i = 0$ for all ancestors v of ℓ. Suppose $u \in \mathbb{F}_2^n$ is colored ℓ by the PCT. Then, $C(u+e_i) = C(u)$ and the coloring is not proper. Conversely, if the coloring is improper then there is some u colored ℓ such that for some $i \in [n]$, $C(u+e_i) = C(u)$.

But then $\langle h(v), u\rangle = \langle h(v), u + e_i\rangle$ for all ancestors v which implies $\langle h(v), e_i\rangle = 0$ for all ancestors v and thus there is a a zero column at coordinate i. □

Definition 4.19. *We say a PCT (T, k, n, h, C) is autumnal if C is a fall k-coloring of $\mathbb{F}_2^n$.*

Notice that if the PCT (T, k, n, h, C) is autumnal then by appending a one to each decision vector we see that there is an autumnal PCT for the same value of k and dimension $n + 1$. Thus there exists an autumnal PCT for all higher dimensions. From Theorem 4.18, we see that for a PCT to be autumnal it is necessary that it be non-degenerate and not contain a zero column. However, these are insufficient. It does not follow from these criterion alone that each color class is dominating. The following terminology will be useful.

Definition 4.20. *For a PCT (T, k, n, h, C) and decision node $t \in T$, let $(T, k, n, h, C)|_t$ be a PCT (T', k', n', h', C') as follows:*

- $T' = T|_t$
- *k' is the number of leaves of $T|_t$*
- *$n' = n$, i.e. the new PCT still colors vertices in $\mathbb{F}_2^n$*
- *If U is the set of decision nodes in $T|_t$ then $h' = h|_U$. That is h' is the restriction of h to decision nodes in $T|_t$.*
- *C' is the resulting coloring function. The domain of C' is $\mathbb{F}_2^n$ and the range is a subset (possibly proper) of the leaves of $T|_t$.*

We call $(T, k, n, h, C)|_t$ a sub-PCT of the original.

Definition 4.21. *Given a PCT (T, k, n, h, C) and a set of coordinates $I \subseteq [n]$, we say the PCT is move anywhere (MA) on I if for all $u \in \mathbb{F}_2^n$, $|\{C(u+e_i) \,:\, i \in I\} \setminus \{C(u)\}| = k - 1$. That is from any u colored $C(u)$ in the directions specified by I it has adjacent to it all $k-1$ other colors. If we say a PCT is move anywhere without specifying a set I, we assume that $I = [n]$.*

Thus, a PCT is autumnal if and only if it is proper and MA.

Definition 4.22. *Given a PCT (T, k, n, h, C) and a set of coordinates $I \subseteq [n]$, we say the PCT is reach anywhere (RA) on I if for all $u \in \mathbb{F}_2^n$, $|\{C(u+e_i) \ : \ i \in I\}| = k$. That is from any $u \in \mathbb{F}_2^n$, its neighbors in the directions specified by I exhibit all k colors.*

Clearly if a PCT is RA then it does not give a proper coloring. However, it will be a useful property for sub-PCTs to have. To illustrate the above two definitions consider the PCT depicted in Figure 4.2. Denote this PCT by P. We have already observed that P is autumnal and therefore proper and MA. As an example of the MA property observe as before that $C((0,0,0)) = 00$. If ℓ' is any leaf other than 00 then there is some e_i such that $C(e_i) = \ell'$. For example, if $\ell' = 10$ then note that $C(e_2) = 10$. Further, consider the decision node $t = 0$ which is the right child of the root. Let P_0 be the sub-PCT rooted at t. This sub-PCT contains one decision node, t, and two leaves ℓ and ℓ'. Observe that P is MA on $\{3\}$ and RA on $\{2, 3\}$. Let C_0 be the coloring function of P_0. Then if $C_0(v) = \ell$ we see that $C_0(v + e_3) = \ell'$ and vice versa. So we see that P_0 is MA on $\{3\}$. Next to see that P is *RA* on $\{2, 3\}$, observe that $C_0(v + e_2) = C_0(v)$ for all $v \in \mathbb{F}_2^3$ and as before adding e_3 switches color classes.

Theorem 4.23. *Let (T, k, n, h, C) be a PCT and $I \subseteq [n]$ a set of coordinates. Let r be the root of T and let r_0 and r_1 be its right and left children respectively. Define coordinate sets*

$$I_0 = \{i \ : \ h(r)_i = 0\} \cap I$$

and

$$I_1 = \{i \ : \ h(r)_i = 1\} \cap I.$$

The PCT is MA on I if the following four conditions hold:

(1) The PCT $(T, k, n, h, C)|_{r_0}$ is MA on $I_0 \cap I$.

(2) The PCT $(T, k, n, h, C)|_{r_1}$ is MA on $I_0 \cap I$.

(3) The PCT $(T, k, n, h, C)|_{r_0}$ is RA on $I_1 \cap I$.

(4) The PCT $(T, k, n, h, C)|_{r_1}$ is RA on $I_1 \cap I$.

The PCT is RA on I if the following four conditions hold:

(a) The PCT $(T, k, n, h, C)|_{r_0}$ is RA on $I_0 \cap I$.

(b) The PCT $(T, k, n, h, C)|_{r_1}$ is RA on $I_0 \cap I$.

(c) The PCT $(T, k, n, h, C)|_{r_0}$ is RA on $I_1 \cap I$.

(d) The PCT $(T, k, n, h, C)|_{r_1}$ is RA on $I_1 \cap I$.

Proof. First we show the PCT is MA on I provided conditions (1)-(4) hold. For any $u \in \mathbb{F}_2^n$ we have $C(u) = \ell$ where ℓ is a leaf of T. Let ℓ' be any leaf of T with $\ell' \neq \ell$. We have four cases.

1. $\ell \in T|_{r_0}$ and $\ell' \in T|_{r_0}$. Using condition (1) we see that $(T, k, n, h, C)|_{r_0}$ is MA on $I_0 \cap I$. Let C' be the restricted coloring function. Since the sub-PCT is MA on $I_0 \cap I$ there exists $i \in I_0 \cap I$ such that $C'(u + e_i) = \ell'$. Note that for the root r, $h(r)_i = 0$ since $i \in I_0$. Therefore $\langle u, h(r)\rangle = \langle u + e_i, h(r)\rangle$ and so changing this bit does not affect the behavior at the root. Thus $C(u + e_i) = \ell'$.
2. $\ell \in T|_{r_1}$ and $\ell' \in T|_{r_1}$. Using condition (2) we see that $(T, k, n, h, C)|_{r_1}$ is MA on $I_0 \cap I$. Let C' be the restricted coloring function. As in case one, since the sub-PCT of interest is MA, there exists $i \in I_0 \cap I$ such that $C'(u + e_i) = \ell'$. Since $i \in I_0$ changing this bit does not change the behavior at the root and thus $C(u + e_i) = \ell'$.
3. $\ell \in T|_{r_1}$ and $\ell' \in T|_{r_0}$. Using condition (3) we see that $(T, k, n, h, C)|_{r_0}$ is RA on $I_1 \cap I$. Let C' be the restricted coloring function. Then there exists $i \in I_1 \cap I$ such that $C'(u + e_i) = \ell'$. As ℓ and ℓ' are in different subtrees below the root we need for $\langle h(r), u\rangle \neq \langle h(r), u + e_i\rangle$ which is true since $i \in I_1$. Thus $C(u + e_i) = \ell'$.
4. $\ell \in T|_{r_0}$ and $\ell' \in T|_{r_1}$. Using condition (4) we see that $(T, k, n, h, C)|_{r_1}$ is RA on $I_1 \cap I$. Let C' be the restricted coloring function. As in case three, there exists $i \in I_1 \cap I$ such that $C'(u + e_i) = \ell'$. As ℓ and ℓ' are in different subtrees below the root we need for $\langle h(r), u\rangle \neq \langle h(r), u + e_i\rangle$ which is true since $i \in I_1$. Thus $C(u + e_i) = \ell'$.

Next, we show the PCT is RA on I provided (a)-(d) hold. For any $u \in \mathbb{F}_2^n$ we have $C(u) = \ell$ where ℓ is a leaf of T. Let ℓ' be any leaf of T possibly equal to ℓ. We have four cases.

1. $\ell \in T|_{r_0}$ and $\ell' \in T|_{r_0}$. Using condition (a) we see that $(T, k, n, h, C)|_{r_0}$ is RA on $I_0 \cap I$. Let C' be the restricted coloring function. Since the sub-PCT is RA

on $I_0 \cap I$ there exists $i \in I_0 \cap I$ such that $C'(u + e_i) = \ell'$. Since $i \in I_0$ we have $C(u + e_i) = \ell'$.

2. $\ell \in T|_{r_1}$ and $\ell' \in T|_{r_1}$. Using condition (b) we see that $(T, k, n, h, C)|_{r_1}$ is RA on $I_0 \cap I$. Let C' be the restricted coloring function. There exists $i \in I_0 \cap I$ such that $C'(u + e_i) = \ell'$. Since $i \in I_0$ we have $C(u + e_i) = \ell'$.
3. $\ell \in T|_{r_1}$ and $\ell' \in T|_{r_0}$. Using condition (c) we see that $(T, k, n, h, C)|_{r_0}$ is RA on $I_1 \cap I$. Let C' be the restricted coloring function. As the sub-PCT is RA on $I_1 \cap I$ there exists $i \in I_1 \cap I$ such that $C'(u + e_i) = \ell'$. As ℓ and ℓ' are in different subtrees below the root we need for $\langle h(r), u\rangle \neq \langle h(r), u + e_i\rangle$ which is true since $i \in I_1$. Thus $C(u + e_i) = \ell'$.
4. $\ell \in T|_{r_0}$ and $\ell' \in T|_{r_1}$. Using condition (d) we see that $(T, k, n, h, C)|_{r_1}$ is RA on $I_1 \cap I$. Let C' be the restricted coloring function. As in case three, there exists $i \in I_1 \cap I$ such that $C'(u + e_i) = \ell'$. As ℓ and ℓ' are in different subtrees below the root we need for $\langle h(r), u\rangle \neq \langle h(r), u + e_i\rangle$ which is true since $i \in I_1$. Thus $C(u + e_i) = \ell'$.

□

Let $P = (T, k, n, h, C)$ be the PCT depicted in Figure 4.2. If r is the root of T then as $h(r) = (0, 1, 1)$, we see that $I_0 = \{1\}$ and $I_1 = \{2, 3\}$. Let P_0 be the sub-PCT rooted at the right child of the root and P_1 be the sub-PCT rooted at the left child. From Theorem 4.23 the fact that this PCT is MA on $\{1, 2, 3\}$ follows from the easy to check facts:

1. P_0 is MA on $\{1\}$.
2. P_1 is MA on$\{1\}$.
3. P_0 is RA on $\{2, 3\}$.
4. P_1 is RA on $\{2, 3\}$.

Next we give a lower bound for the dimension at which a PCT can be autumnal.

Theorem 4.24. *Let 2^a be the smallest power of 2 such that $2^a \geq k$. Then if a PCT (T, k, n, h, C) is autumnal we must have $n \geq 2^a - 1$.*

Proof. We have $2^{a-1} < k \leq 2^a$. The height of T must be at least a, for otherwise T

would have at most $2^{a-1} < k$ leaves. Since the PCT is autumnal in particular it is non-degenerate. If a leaf ℓ is a distance at least a from the root it determines a color class of size $|V_\ell| \leq 2^{n-a}$. From Lemma 4.8 we have that $|V_\ell| \geq 2^n/(n+1)$. Thus,

$$\frac{2^n}{2^a} \leq |V_\ell| \leq \frac{2^n}{n+1}$$

which implies $n \geq 2^a - 1$ as desired. □

In Section 4.3 we will show that this lower bound is achievable for even k.

4.3 Construction

Let k be a positive, even integer. Let 2^a be the smallest power of 2 such that $k \leq 2^a$. We begin by noting that there is a canonical labelled binary tree with k leaves of height a. We label the root by the empty string and if k is a power of 2 then the tree is the complete binary tree of height a where the right and left child of node t are labelled $t0$ and $t1$ respectively. We can write $k = 2^{a-1} + b$ where $b \leq 2^{a-1}$. Then below the right child of the root we put the complete binary tree of height $a-1$ prepending the string 0 to all its nodes and, recursively, below the left child of the root we put the canonical binary tree with b leaves prepending the string 1 to all its nodes. For example if $k = 6$ then as $6 = 4 + 2$ we put the complete tree of height two below the right child of the root and the complete tree of height one below the left. See Figure 4.3.

The following notation will be useful.

Definition 4.25. *For integers $0 \leq i < a$ let $\mathrm{alt}(i, a)$ be the vector in $\mathbb{F}_2^{2^a}$ which is the concatenation of 2^i blocks, B, of the form $B = (0, \ldots, 0, 1, \ldots, 1)$ where 0 appears 2^{a-i-1} times as does 1.*

For example we have

$$\mathrm{alt}(0,3) = (0,0,0,0,1,1,1,1)$$

$$\mathrm{alt}(1,3) = (0,0,1,1,0,0,1,1)$$

$$\mathrm{alt}(2,3) = (0,1,0,1,0,1,0,1).$$

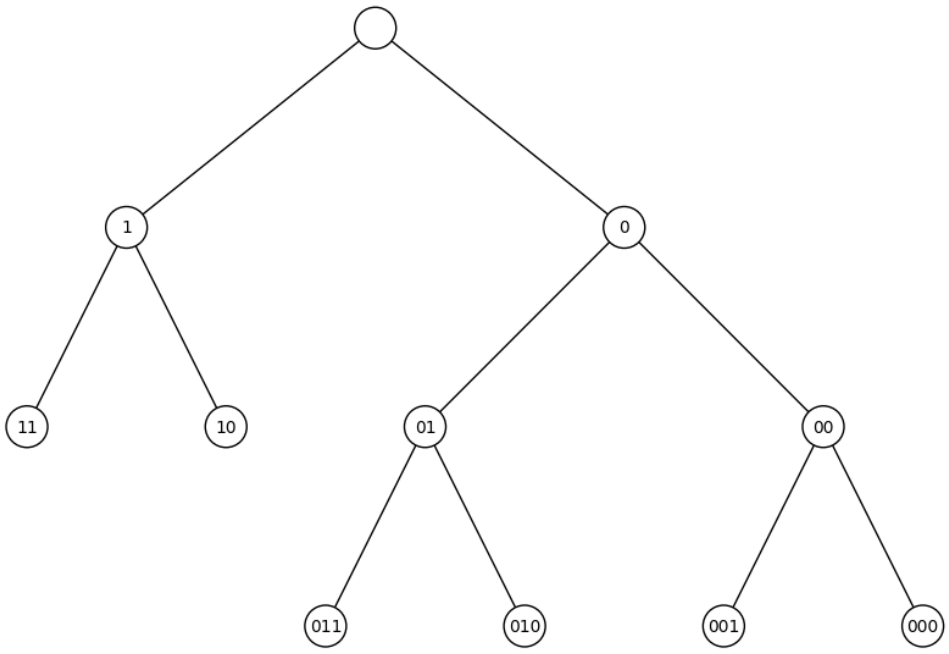


Figure 4.3: The canonical binary tree with $k = 6$ leaves.

Definition 4.26. *For a positive, even integer k, let CONSTRUCTION$_0(k)$ be the PCT (T, k, n, h, C) defined as follows. Let 2^a be the smallest power of 2 so that $k \leq 2^a$. Let $n = 2^a$. Let T be the canonical tree with k leaves. For each decision node, t, let m be the height of t. Then set $h(t) = \text{alt}(m, n)$.*

Since $\text{alt}(m, n)_1 = 0$ for all $m < n$ it is easy to see that CONSTRUCTION$_0(k)$ will have a column of zeros. Although this precludes the coloring being proper it does allow the PCT to be RA as shown in Lemma 4.28. The following notation will be useful.

Definition 4.27. *Given a PCT (T, k, n, h, C) and a nonempty set of coordinates $I \subseteq [n]$ we define $(T, k, n, h, C) \cap I$ to be the PCT (T', k', n', h', C') where*

- $T' = T$
- $k' = k$
- $n = |I|$
- *h' is the projection of h onto the coordinates of I*
- *C' is the resulting coloring function*

Intuitively for a PCT P and set of coordinates I, $P \cap I$ is the PCT given by only considering the coordinates of I.

Lemma 4.28. *For k even and positive, let $\mathit{CONSTRUCTION}_0(k) = (T, k, n, h, C)$. This PCT is RA.*

Proof. Let $2^{a-1} < k \leq 2^a$. Let r be the root of T and $I_0 = \{i \ : \ h(r)_i = 0\}$ and $I_1 = \{i \ : \ h(r)_i = 1\}$. Observe that (T, k, n, h, C) has the following recursive structure

- $(T, k, n, h, C)|_{r_0} \cap I_0 = \text{CONSTRUCTION}_0(2^{a-1})$
- $(T, k, n, h, C)|_{r_0} \cap I_1 = \text{CONSTRUCTION}_0(2^{a-1})$
- $(T, k, n, h, C)|_{r_1} \cap I_0 = \text{CONSTRUCTION}_0(k - 2^{a-1})$
- $(T, k, n, h, C)|_{r_1} \cap I_1 = \text{CONSTRUCTION}_0(k - 2^{a-1})$

If each of these sub-PCTs are RA then, by Theorem 4.23, the original PCT is RA. Thus the fact that $\text{CONSTRUCTION}_0(k)$ is RA follows by strong induction on k provided we demonstrate the base case $k = 2$. Note that $\text{CONSTRUCTION}_0(2)$ consists of a tree with one decision node, the root, r. We have $n = 2$ and $h(r) = (0, 1)$. There are two leaves ℓ and ℓ' and for any $v \in \mathbb{F}_2^2$ we have $C(v + e_1) = C(v)$ and $C(v + e_2) \neq C(v)$. So $\text{CONSTRUCTION}_0(2)$ is indeed RA. □

Definition 4.29. *For a PCT (T, k, n, h, C) we call a node $t \in T$ a twig if either of its children is a leaf.*

Definition 4.30. *For integers $0 \leq i < a$ define* $\text{altfill}(i, a)$ *as in Definition 4.25 except that we set the leftmost block to be all ones.*

For example,

$$\text{altfill}(1, 3) = (1, 1, 1, 1, 0, 0, 1, 1)$$
$$\text{altfill}(2, 3) = (1, 1, 0, 1, 0, 1, 0, 1).$$

We use these notions to tweak $\text{CONSTRUCTION}_0(k)$ so as to remove the zero columns.

Definition 4.31. *For a positive, even integer k, let $\mathit{CONSTRUCTION}_1(k)$ be the PCT (T, k, n, h, C) defined as follows. Let 2^a be the smallest power of 2 so that $k \leq 2^a$. Let $n = 2^a$. Let T be the canonical tree with k leaves. For each decision node, t, let m be the height of t. If t is a twig, then set $h(t) = \text{altfill}(m, n)$. Otherwise set $h(t) = \text{alt}(m, n)$.*

Lemma 4.32. *For k even and positive, let $CONSTRUCTION_1(k) = (T, k, n, h, C)$. Let $2^{a-1} < k \leq 2^a$. Let r be the root of T and $I_0 = \{i \ : \ h(r)_i = 0\}$ and $I_1 = \{i \ : \ h(r)_i = 1\}$. This PCT is MA and has the following recursive structure:*

1. $(T, k, n, h, C)|_{r_0} \cap I_0 = CONSTRUCTION_1(2^{a-1})$
2. $(T, k, n, h, C)|_{r_1} \cap I_0 = CONSTRUCTION_1(k - 2^{a-1})$
3. $(T, k, n, h, C)|_{r_0} \cap I_1 = CONSTRUCTION_0(2^{a-1})$
4. $(T, k, n, h, C)|_{r_1} \cap I_1 = CONSTRUCTION_0(k - 2^{a-1})$.

Furthermore, $CONSTRUCTION_1(k)$ contains no zero column and as such gives a proper coloring. Finally, for any decision node, $t \in T$, the decision matrix, D_t, contains a column of the form $(0, 0, \ldots, 0, 1)$. Thus for any $t \in T$, $D(t)$ is linearly independent. So the PCT is non-degenerate. Thus $CONSTRUCTION_1(k)$ is autumnal.

Proof. First we show that $\text{CONSTRUCTION}_1(k)$ is MA. The recursive structure follows from the definition. As in the proof of Lemma 4.28, the fact that $\text{CONSTRUCTION}_1(k)$ is MA follows from Theorem 4.23 provided that the first two sub-PCTs are MA and the latter two are RA. We have already shown in Lemma 4.28 that the latter two PCTs are RA. The fact that the first two sub-PCTs are are MA follows by by strong induction on k provided we demonstrate the base case $k = 2$. Note that $\text{CONSTRUCTION}_1(2)$ consists of a tree with one decision node, the root, r. We have $n = 2$ and $h(r) = (1, 1)$. There are two leaves ℓ and ℓ' and for any $v \in \mathbb{F}_2^2$ we have $C(v + e_1) \neq C(v)$. Thus $\text{CONSTRUCTION}_1(2)$ is indeed MA.

The fact that $\text{CONSTRUCTION}_1(k)$ contains no zero column follows from the definition. In $\text{CONSTRUCTION}_0(k)$ the only columns of zeros occurred in the leftmost blocks which we have set to one in each twig. Finally, for any decision node, t, of height m observe that the leftmost block B, of $\text{alt}(m, n)$ of its decision vector is either of the form $B = (0, \ldots, 0, 1 \ldots, 1)$ if t is not a twig or all ones if it is a twig. For any ancestor, t', with height $m' < m$ we have blocks at least twice as large and thus all zeros for the coordinates of B. Thus D_t contains a column of the form $(0, \ldots, 0, 1)$. □

Note that $\text{CONSTRUCTION}_1(2)$ illuminates an inefficiency in this construction. It would have sufficed to have $n = 1$ and for the root, r, to have $h(r) = (1)$. In

general, columns one and two of CONSTRUCTION$_1(k)$ are the same. That is, if CONSTRUCTION$_1(k) = (T, k, n, h, C)$ then for all decision nodes, t, $h(t)_1 = h(t)_2$. We can remove this redundancy by deleting the leftmost column which will give us a construction of dimension $n = 2^a - 1$ as desired.

Definition 4.33. *Let CONSTRUCTION(k) be the dimension $n = 2^a - 1$ construction derived from deleting the leftmost column in CONSTRUCTION$_1(k)$.*

See Figure 4.4 for an illustration of CONSTRUCTION(6). Note that removing an identical column does not affect the properties of being MA, RA, proper or non-degenerate. Thus we have the following lemma.

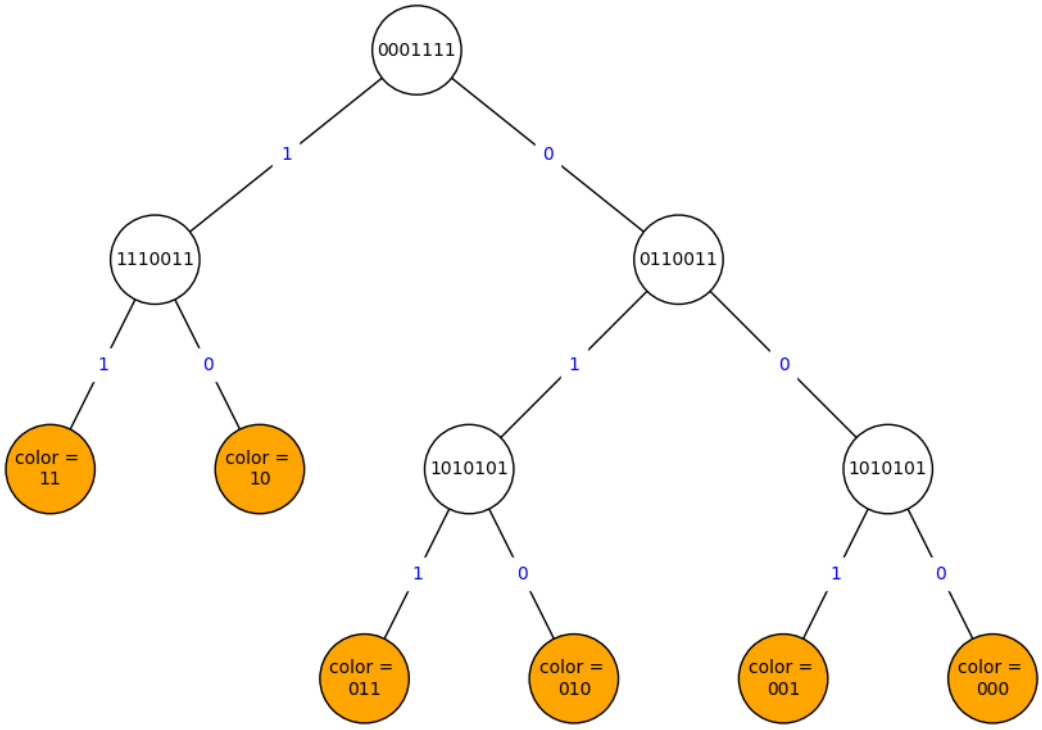


Figure 4.4: CONSTRUCTION(6). The decision nodes are drawn in white. For each decision node, t, we have shown $h(t)$. The leaves are shown in orange.

Lemma 4.34. *Let k be even with $2^{a-1} < k \leq 2^a$. Then we have $n = 2^a - 1$ and CONSTRUCTION$(k) = (T, k, n, h, C)$ is autumnal. Thus C is an affine fall k-coloring of $\mathbb{F}_2^n$.* □

We see that Theorem 4.10 follows immediately from Lemma 4.34. In Figure 4.5 we

give an illustration of CONSTRUCTION(14). Next we see that a PCT can only be autumnal for even k.

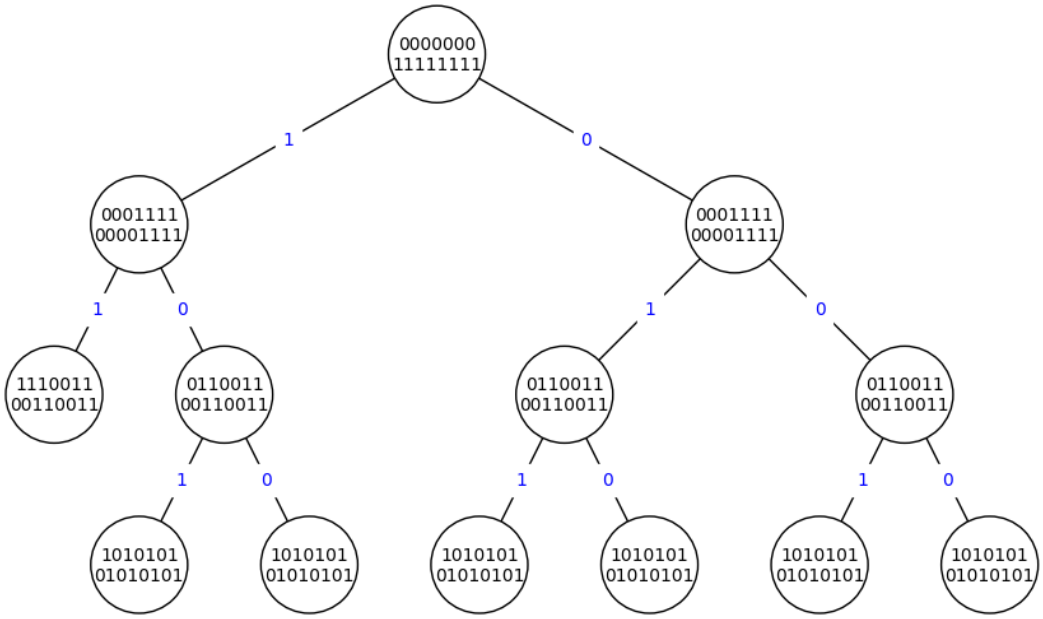


Figure 4.5: The decision nodes of CONSTRUCTION(14) which gives an affine fall 14-coloring of Q_{15}. To avoid clutter, leaves are not drawn. Thus each of the seven apparent leaves in this diagram are in fact twigs and have as children two leaves. In each decision node, t, we have shown $h(t)$.

Theorem 4.35. *For any autumnal PCT each twig must have as children two leaves. In particular, any PCT with k leaves where k is odd is not autumnal.*

Proof. Suppose for the sake of contradiction that the PCT (T, k, n, h, C) is autumnal and there is some twig, t, with children ℓ, a leaf, and t', a decision node. Let ℓ_1 and ℓ_2 be any distinct leaves that are descendants of t'. As an example where ℓ_1 and ℓ_2 are children of t' see Figure 4.6. If the PCT is autumnal then there exists $v \in \mathbb{F}_2^n$ such that $C(v) = \ell_1$ and an index i such that $C(v + e_i) = \ell_2$. Consider the $m \times n$ decision matrix, $D_{t'}$. Column i of $D_{t'}$ must be $(0, \ldots, 0, 1)$. If any of the first $m' - 1$ coordinates is nonzero then $C(v+e_i) \neq \ell_2$ as the decision process will not reach t'. The final coordinate must be one so that $C(v + e_i) \neq C(v)$. However, this means there is a column of zeros in coordinate i for D_t. In particular, for any u such that $C(u) = \ell$ we would have $C(u+e_i) = \ell$ and the coloring would not be proper. This is a contradiction.

Thus if the PCT is autumnal each twig has as children two leaves and the number of leaves must be even. □

Theorem 4.35 has the following immediate corollary.

Corollary 4.36. *For any coloring given by an autumnal PCT, each color class has a parallel.*

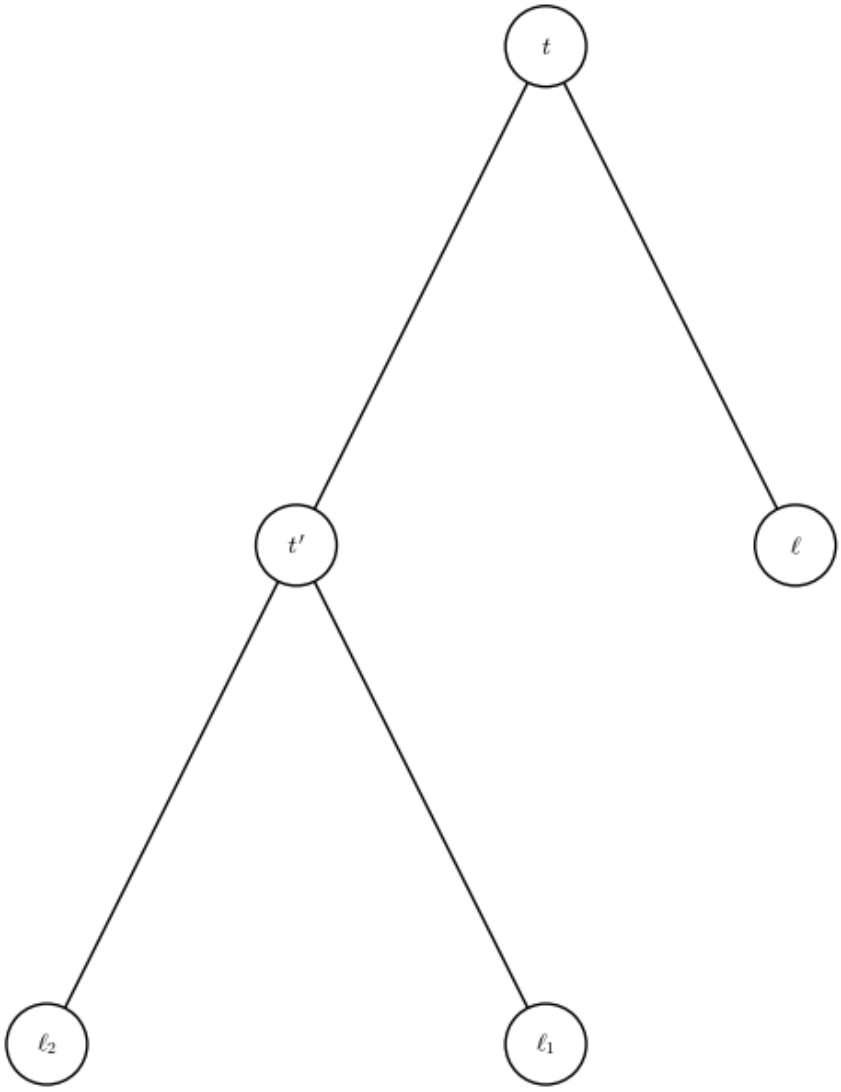


Figure 4.6: A twig, t, with children ℓ and t' that are a leaf and a decision node respectively.

4.4 Conclusion and open questions

We have shown that for any positive, even integer k that there exists an affine fall k-coloring of Q_n for $n \geq 2^a - 1$ where 2^a is the smallest power of 2 such that $k \leq 2^a$. Furthermore, for our method of construction, parity coloring trees, our construction achieves the minimum possible dimension. We have not, however, shown that for $2^{a-1} < k \leq 2^a$ that $n = 2^a - 1$ is the minimum dimension for which an affine fall k-coloring of Q_n exists. We know this statement holds for k a power of 2 and, by ILP solving, for $k = 6$. But for other values of k we leave this as an open question. Furthermore, we have shown for k odd that parity coloring trees cannot give an affine fall k-coloring. Recall that for $k = 3$ there are no fall k-colorings of Q_n for any n. For $k > 3$, we make the following conjecture.

Conjecture 4.37. *Let $k > 3$ be an odd, positive integer. For all n, there is no affine fall k-coloring of Q_n.*

We do not know if Conjecture 4.37 holds even for $k = 5$.

References

[AS00] Noga Alon and Joel H. Spencer, *The probabilistic method*, second ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience [John Wiley & Sons], New York, 2000, With an appendix on the life and work of Paul Erdős. MR 1885388

[Bal10] Vladimir Baltić, *On the number of certain types of strongly restricted permutations*, Appl. Anal. Discrete Math. **4** (2010), no. 1, 119–135. MR 2654934

[BC72] Joel Brenner and Larry Cummings, *The Hadamard maximum determinant problem*, Amer. Math. Monthly **79** (1972), 626–630. MR 0301030

[BDS76] J. Browkin, B. Diviš, and A. Schinzel, *Addition of sequences in general fields*, Monatsh. Math. **82** (1976), no. 4, 261–268. MR 0432581

[BJL85] Thomas Beth, Dieter Jungnickel, and Hanfried Lenz, *Design theory*, Bibliographisches Institut, Mannheim, 1985. MR 779284

[BKSE12] Jeff Bezanzon, Stefan Karpinski, Viral Shah, and Alan Edelman, *Julia: A fast dynamic language for technical computing*, Lang.NEXT, apr 2012.

[BR18] Henning Bruhn and Dieter Rautenbach, *Maximal determinants of combinatorial matrices*, Linear Algebra Appl. **553** (2018), 37–57. MR 3809368

[Bre73] L. M. Bregman, *Certain properties of nonnegative matrices and their permanents*, Dokl. Akad. Nauk SSSR **211** (1973), 27–30. MR 0327788

[Cha82] Seth Chaiken, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Algebraic Discrete Methods **3** (1982), no. 3, 319–329. MR 666857

[DHH$^+$00] J. E. Dunbar, S. M. Hedetniemi, S. T. Hedetniemi, D. P. Jacobs, J. Knisely, R. C. Laskar, and D. F. Rall, *Fall colorings of graphs*, J. Combin. Math. Combin. Comput. **33** (2000), 257–273, Papers in honour of Ernest J. Cockayne. MR 1772767

[DHL17] Iain Dunning, Joey Huchette, and Miles Lubin, *Jump: A modeling language for mathematical optimization*, SIAM Review **59** (2017), no. 2, 295–320.

[FHHJ17] Claus Fieker, William Hart, Tommy Hofmann, and Fredrik Johansson, *Nemo/hecke: Computer algebra and number theory packages for the julia programming language*, Proceedings of the 2017 ACM on International Symposium on Symbolic and Algebraic Computation (New York, NY, USA), ISSAC '17, ACM, 2017, pp. 157–164.

[Fra] Cole Franks, personal communication.

[FvdD97] Shaun Fallat and P. van den Driessche, *Maximum determinant of* $(0,1)$ *matrices with certain constant row and column sums*, Linear and Multilinear Algebra **42** (1997), no. 4, 303–318. MR 1487516

[Gal03] David Galvin, *On homomorphisms from the Hamming cube to* $\mathbf{Z}$, Israel J. Math. **138** (2003), 189–213. MR 2031957

[GH13] Wayne Goddard and Michael A. Henning, *Independent domination in graphs: a survey and recent results*, Discrete Math. **313** (2013), no. 7, 839–854. MR 3017969

[GO18] LLC Gurobi Optimization, *Gurobi optimizer reference manual*, 2018.

[Had93] J. Hadamard, *Resolution d'une question relative aux determinants*, Bull. des Sciences Math. **2** (1893), 240–246.

[HHW88] Frank Harary, John P. Hayes, and Horng-Jyh Wu, *A survey of the theory of hypercube graphs*, Comput. Math. Appl. **15** (1988), no. 4, 277–289. MR 949280

[HJ13] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, second ed., Cambridge University Press, Cambridge, 2013. MR 2978290

[HS86] David Lee Hilliker and E. G. Straus, *Uniqueness of linear combinations* $(\mathrm{mod}\, p)$, J. Number Theory **24** (1986), no. 1, 1–6. MR 852185

[Hun07] J. D. Hunter, *Matplotlib: A 2d graphics environment*, Computing In Science & Engineering **9** (2007), no. 3, 90–95.

[Jan07] Miroslawa Janczak, *A note on a problem of Hilliker and Straus*, Electron. J. Combin. **14** (2007), no. 1, Note 23, 8. MR 2350451

[JN80] Charles R. Johnson and Morris Newman, *A surprising determinantal inequality for real matrices*, Math. Ann. **247** (1980), no. 2, 179–185. MR 568207

[Kop17] Swastik Kopparty, *open problem session*, Harvard CMSA, October 2017.

[KR46] Irving Kaplansky and John Riordan, *The problème des ménages*, Scripta Math. **12** (1946), 113–124. MR 0019074

[Lev00] Vsevolod F. Lev, *Simultaneous approximations and covering by arithmetic progressions in* $\mathbf{F}_p$, J. Combin. Theory Ser. A **92** (2000), no. 2, 103–118. MR 1796468

[Li15] Yiting Li, *Ménage numbers and ménage permutations*, J. Integer Seq. **18** (2015), no. 6, Article, 15.6.8, 23. MR 3360901

[LL09] Renu Laskar and Jeremy Lyle, *Fall colouring of bipartite graphs and Cartesian products of graphs*, Discrete Appl. Math. **157** (2009), no. 2, 330–338. MR 2479807

[LLR99] Chi-Kwong Li, Julia Shih-Jung Lin, and Leiba Rodman, *Determinants of certain classes of zero-one matrices with equal line sums*, Rocky Mountain J. Math. **29** (1999), no. 4, 1363–1385. MR 1743375

[LPW01] Zbigniew Lonc, Krzysztof Parol, and Jacek M. Wojciechowski, *On the number of spanning trees in directed circulant graphs*, Networks **37** (2001), no. 3, 129–133. MR 1826836

[McK83] Brendan D. McKay, *Spanning trees in regular graphs*, European J. Combin. **4** (1983), no. 2, 149–160. MR 705968

[Min63] Henryk Minc, *Upper bounds for permanents of* (0, 1)*-matrices*, Bull. Amer. Math. Soc. **69** (1963), 789–791. MR 0155843

[MOA11] Albert W. Marshall, Ingram Olkin, and Barry C. Arnold, *Inequalities: theory of majorization and its applications*, second ed., Springer Series in Statistics, Springer, New York, 2011. MR 2759813

[Ned09] Zhivko Nedev, *An algorithm for finding a nearly minimal balanced set in* $\mathbb{F}_p$, Math. Comp. **78** (2009), no. 268, 2259–2267. MR 2521288

[Ned12] ———, *Lower bound for balanced sets*, Theoret. Comput. Sci. **460** (2012), 89–93. MR 2980518

[NQ08] Zhivko Nedev and Anthony Quas, *Balanced sets and the vector game*, Int. J. Number Theory **4** (2008), no. 3, 339–347. MR 2424326

[O'D14] Ryan O'Donnell, *Analysis of Boolean functions*, Cambridge University Press, New York, 2014. MR 3443800

[Olk14] Ingram Olkin, *A determinantal inequality for correlation matrices*, Statist. Probab. Lett. **88** (2014), 88–90. MR 3178337

[Orr05] William P. Orrick, *The maximal* $\{-1, 1\}$*-determinant of order 15*, Metrika **62** (2005), no. 2-3, 195–219. MR 2274990

[OS07] William P. Orrick and Bruce Solomon, *Large-determinant sign matrices of order* $4k + 1$, Discrete Math. **307** (2007), no. 2, 226–236. MR 2285193

[Pol75] J. M. Pollard, *A Monte Carlo method for factorization*, Nordisk Tidskr. Informationsbehandling (BIT) **15** (1975), no. 3, 331–334. MR 0392798

[Rys56] H. J. Ryser, *Maximal determinants in combinatorial investigations*, Canad. J. Math. **8** (1956), 245–249. MR 0079555

[S+17] W. A. Stein et al., *Sage Mathematics Software (Version 8.1)*, The Sage Development Team, 2017, `http://www.sagemath.org`.

[Sch78] A. Schrijver, *A short proof of Minc's conjecture*, J. Combinatorial Theory Ser. A **25** (1978), no. 1, 80–83. MR 0491216

[Sch78] A. Schinzel, *An inequality for determinants with real entries*, Colloq. Math. **38** (1977/78), no. 2, 319–321. MR 0485920

[Sloa] N. J. A. Sloane, *The domatic number of the n-cube, online encyclopedia of integer sequences*, `http://oeis.org/ A157887`.

[Slob] ———, *Jacobsthal sequence in the online encyclopedia of integer sequences*, `http://oeis.org/A001045`.

[Str76] E. G. Straus, *Differences of residues* (mod p), J. Number Theory **8** (1976), no. 1, 40–42. MR 0392876

[Syl67] J. J. Sylvester, *Thoughts on inverse orthogonal matrices, simultaneous sign-successions, and tessellated pavements in two or more colors, with applications to newtons rule, ornamental tile-work, and the theory of numbers*, Phil. Mag. **34** (1867), 461.

[Wil46] John Williamson, *Determinants whose elements are 0 and 1*, Amer. Math. Monthly **53** (1946), 427–434. MR 0017261

[Zel82] Bondan Zelinka, *Domatic numbers of cube graphs*, Math. Slovaca **32** (1982), no. 2, 117–119. MR 658244

[Zel91] Bohdan Zelinka, *Domination in cubes*, Math. Slovaca **41** (1991), no. 1, 17–19. MR 1094979