# SOME COMBINATORIAL RESULTS ON MATRICES AND POLYNOMIALS

By

**JUSTIN SEMONSEN**

**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, 2020**

## ABSTRACT OF THE DISSERTATION

# Some Combinatorial Results on Matrices and Polynomials

**By JUSTIN SEMONSEN**

**Dissertation Director: Swastik Kopparty**

This thesis studies three problems in combinatorics that concern matrices and polynomials.

The first problem refines a technique by Scheinerman [24] to get an improved upper bound on the determinants of $n \times n$ 0-1 matrices with $k$ ones in each row. Our new bound uses linear programming techniques to analyze a greedy decomposition algorithm, show how it improves on previous methods, and determine its asymptotic behavior as a function of $k$. We also present and analyze an improvement to this method, as well as the limitations and other possibilities of using this technique.

The second problem analyzes the supports of $\mathbb{F}_2$ polynomials on $n$ variables in Hamming balls, and proves an optimal Schwartz-Zippel-like bound. We then use methods based on those of Kasami and Tokura [11], [12] to find and classify all tight polynomials for this bound. Our result is based on studying necessary conditions for "division lemmas" for polynomials.

The third result studies sets of points in $\mathbb{F}_2^n$ for which the sum of any $\mathbb{F}_2$ polynomial of degree $d$ on those points is 0. We prove that for $d \geq 2$, we have that the size of the set must at least twice the affine dimension of the set. For larger $d$, we can show the size of the set must be a constant amount larger than twice the affine dimension, but we conjecture that this can be improved to $\frac{2^{d+1}}{d+2}$ times the affine dimension of the set. We also apply these theorems to prove a bound on the weight distribution of Reed-Muller codes of high dimension.

# Acknowledgements

Firstly I would like to thank my advisor Swastik Kopparty for his guidance in both finding and solving the problems I have worked on during my time at Rutgers. Your help has been invaluable, as well as enjoyable.

I would also like to thank my other committee members, including Bhargav Narayanan, Shubhangi Saraf, and Srikanth Srinivasan on my thesis committee as well as Jeff Kahn, Michael Saks, and Doron Zeilberger on my oral exam committee.

I am also grateful to the numerous other graduate students I have bothered with my research or other subjects, including but not limited to Yonah Biers Ariel, Matt Charnley, Cole Franks, Keith Frankston, Jinyoung Park, Aditya Potukuchi, Abigail Raz, and Daniel Scheinerman. In particular I would like to thank Daniel Scheinerman for pushing me to continue his research, as well as the numerous conversations while working on the first section of this thesis.

I am lucky to have an amazing family that has been there for me in so many (non-mathematical) ways throughout my graduate career, supporting me in more ways than I can count. Thank you to my parents Nora and Kevin, and my sibings Bryan and Lianna.

Lastly I want to acknowledge my wonderful wife Einat Brenner for always being there for me, even when she was far busier than me with her own PhD. I love you, and I do not think I could have done it without you.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Asymptotic Determinant Bounds on 0-1 Matrices with Fixed Row Sums

The results presented in Chapter 2 deal with the determinants of $M_{n\times n}(\{0,1\})$. Specifically, we focus on upper bounding the determinants of the following class of matrices:

$$R(n,k) := \{A \in M_{n\times n}(\{0,1\}) : \|\vec{A}_i\| = \sqrt{k}\,\forall\, 1 \leq i \leq n\}$$

Since each row has norm $\sqrt{k}$, Hadamard's Inequality [10] gives that $M(n,k) := \max_{A\in R(n,k)} \det(A) \leq \sqrt{k}^n$. Ryser [21] was able to improve this to $M(n,k) \leq k(k-\lambda)^{\frac{n-1}{2}}$ for $\lambda = \frac{k(k-1)}{n-1}$. However, he also showed this bound cannot be tight when $\lambda < 1$, meaning that in the regime where $k < \sqrt{n}$, this improvement is small. In fact, for any fixed $k$ both Hadamard and Ryser only showed the following asymptotic bound: $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \leq \sqrt{k}$.

Bruhn and Rautenbach [5] were able to show that $M(n,2) \leq (\sqrt[3]{2})^n$, meaning that $\limsup_{n\to\infty} M(n,2)^{\frac{1}{n}} \leq \sqrt[3]{2} < \sqrt{2}$. Although this improves the asymptotic bound for $k = 2$, they also conjectured that this bound could be improved for $k > 2$. Scheinerman [24] developed a technique based on decomposing the matrices into blocks of rows to show that $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \leq c_{q,k}$ for some $c_{q,k} < \sqrt{k}$.

Scheinerman [24] also presented a greedy algorithm to optimize his bounds, but could only conjecture that this greedy algorithm would generate a bound better than his optimal $c_{q_k,k} = \sqrt{k} - \frac{.096}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$. The results in Chapter 2 use tools derived from linear programming to analyze this greedy algorithm and prove that $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \leq c_k$ for a $c_k < \sqrt{k}$. We also use duality to not only prove that $c_k < c_{q,k}$, but also find the optimal solution and show the asymptotic bound is of the form:

$$c_k = \sqrt{k} - \frac{1 - \frac{\pi^2}{12}}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$$

We also develop and analyze some minor improvements upon the greedy algorithm to get another asymptotic bound $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \leq c'_k < c_k$, although we conjecture that the asymptotic behavior of $c'_k$ is the same.

## 1.2 Supports of Binary Polynomials

The Schwartz-Zippel Lemma [7],[30],[25] says that the probability that a nonzero degree $d$ polynomial is zero over a finite set $S$ is no more than $\frac{d}{|S|}$. For $\mathbb{F}_2$ polynomials on $\mathbb{F}_2^n$, this can be generalized to prove that every nonzero polynomial of degree $d$ has at least $2^{n-d}$ non-zeros.

Kasami and Tokura [11],[12] took this a step further by proving that every polynomial for which this bound is tight is the product of $d$ linearly independent linear functions. They also extended their techniques to categorize the functions that have close to minimal support.

In Chapter 3 we take this a different direction, restricting the domain to only inputs with Hamming weight below a given threshold $r$. We then prove a theorem similar to the Schwartz-Zippel Lemma, saying that the support on inputs with Hamming weight less than $r$ of a degree $d$ polynomial on $n$ variables must be at least $\sum_{i=0}^{r-d}\binom{n-d}{i}$.

We then use some of the techniques from Kasami and Tokura [11],[12] to fully characterize the polynomials that achieve the bound. For $r > d$, we show these are products of certain independent linear factors, and along the way we produce alternate conditions for finding linear factors of polynomials in these restricted domains.

## 1.3 Weight Bounds in Dual Reed-Muller Codes

One of many useful applications of $\mathbb{F}_2$ polynomials is in Reed-Muller codes. Reed-Muller codes use the message to be encoded as coefficients of a degree $d$ polynomial on $n$ variables and then encode it by evaluating that polynomial on every point of $\mathbb{F}_2^n$. The

Schwartz-Zippel Lemma and other polynomial results can be used to find the minimum distance and prove other properties of Reed-Muller codes.

We examined the dual codes of Reed-Muller codes, which can be represented as a set of points on which any polynomial of degree $d$ sums (modulo 2) to 0. Finding small size sets in this model is easy, but only if the points sit inside an affine subspace of small dimension. We show that when the smallest affine subpace spanned by the points is dimension $m$, then the size of the set must be at least twice $m$.

In fact, when $S$ is a dual codeword for polynomials of degree $d$ and $S$ is of affine dimension $m$, we conjecture that $|S| \geq \frac{2^{d+1}}{d+2}(m+1)$. This comes from summing independent affine subspaces of dimension $d+1$, and then counting the resulting affine dimension.

Showing the optimality of this construction is more difficult, but we can prove this for $d = 2$ using a rank-based approach. This fails in the $d = 3$ case, but we can show $|S| \geq 2(m+1) + 2^d - d - 2$ using a clever application of results from Chapter 2.

We then demonstrate why this bound is likely not tight and pose various new avenues for proving the full conjecture as well as a weakening. We also apply our results to counting the low weight codewords of Reed-Muller codes, as well as showing how the conjectures would improve these bounds.

# Chapter 2

# Asymptotic Determinant Bounds on 0-1 Matrices with Fixed Row Sums

## 2.1 The Maximum Determinant Problem

Hadamard's maximum determinant problem [10] asks for the largest determinant among all $n \times n$ zero-one matrices. This problem has been well studied [29],[4], [8], [16], [19],[18], but many questions still remain unanswered.

For the remainder of this paper, let $\vec{A_i}$ be the $i$th row of the matrix $A$. Also let $\|\cdot\|$ represent the standard $l_2$ norm on vectors.

We will look at the maximum determinant of the restricted class of zero-one matrices defined below:

**Definition 2.1.** $R(n,k) = \{A \in M_{n\times n}(\{0,1\}) : \|\vec{A_i}\| = \sqrt{k}\,\forall\, 1 \leq i \leq n\}$ *for* $1 \leq k \leq n$.

We can also characterize the matrices in $R(n,k)$ as the matrices in $M_{n\times n}(\{0,1\})$ whose rows each sum to $k$. This means the vector of all ones is a left eigenvector with eigenvalue $k$.

The question originally posed to me by Scheinerman [24] is what is the largest determinant that can be attained in $R(n,k)$? This lets us define the following quantity:

**Definition 2.2.** $M(n,k) = \max_{A\in R(n,k)} |\det(A)|$

To get a lower bound on $M(n,k)$, it suffices to give a single matrix with large determinant. This lets us show that $M(n,k)$ is supermultiplicative in $n$, as if we have two matrices with large determinants, a block diagonal matrix with those two on the diagonal forms a matrix with determinant equal to the product of the original two determinants. This means that if we exhibit an $m \times m$ matrix $A$ for which $\det(A) = c^m$, then $\det(A \otimes I_t) = c^{mt}$ for any $t$, and thus $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \geq c$.

A projective plane of order $k-1$ has an $(k^2-k+1)\times(k^2-k+1)$ incidence matrix $A$ with $k$ ones in each row, and thus is in $R(k^2-k+1,k)$. Since $AA^\top = J+(k-1)I$, we have that $\det(A)^2 = \det(J_{k^2-k+1}+(k-1)I_{k^2-k+1}) = k^2(k-1)^{k^2-k}$, so $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} \geq k^{\frac{1}{k^2-k+1}}(k-1)^{\frac{k^2-k}{2k^2-2k+2}} = \sqrt{k}-\frac{1}{2\sqrt{k}}+O(k^{\frac{-3}{2}})$ [24]. Computer searches have been unable to yield any improvements on this lower bound.

The first upper bound on $M(n,k)$ that was found is due to Hadamard [10]. Hadamard's result says that $|\det(A)| \leq \prod_{i=1}^{n}\|\vec{A}_i\|$. Since $\|\vec{A}_i\| = \sqrt{k}$ for every $i$, this means $|\det(A)| \leq \sqrt{k}^n$.

This exponential upper bound when $k$ is fixed allows us to use Fekete's Lemma to show that $\limsup_{n\to\infty} M(n,k)^{\frac{1}{n}} = \lim_{n\to\infty} M(n,k)^{\frac{1}{n}}$ exists and is finite, in fact no larger than $\sqrt{k}$.

Let $\rho_k = \lim_{n\to\infty} M(n,k)^{\frac{1}{n}}$ be this limit for every $k$. By the construction above we have that $\rho_k \geq \sqrt{k}-\frac{1}{2\sqrt{k}}+O(k^{\frac{-3}{2}})$, while Hadamard says that $\rho_k \leq \sqrt{k}$.

A result of Ryser [21] gives that $M(n,k) \leq k(k-\lambda)^{\frac{n-1}{2}}$ where $\lambda = \frac{k(k-1)}{n-1}$. When $k$ is large relative to $n$, this is an exponential improvement on Hadamard, but Ryser showed this bound can't be tight unless $\lambda$ is integral.

In the asymptotic regime where $k$ is fixed and $n\to\infty$, this means $\lambda\to 0$, so Ryser's bound is not only not tight, but gives the same bound on $\rho_k$ as Hadamard's bound.

This question has been studied in these and other regimes by many others including [29],[4], [8], [16], [19], and [18]. However, none of these results show any improvement in this exponential factor as $n$ approaches infinity.

Only recently did Bruhn and Rautenbach [5] give an exponentially better bound when $k=2$: $M(n,2) \leq \left(\sqrt[3]{2}\right)^n$. However, their methods left them only able to conjecture that a similar exponential improvement was possible for $k=3$ and larger.

Scheinerman [24] proved such a smaller exponential bound exists by decomposing the matrix into blocks of rows, then analyzing each block separately.

**Theorem 2.3** (Scheinerman [24]). *Let $q$ be an integer with $1\leq q\leq k$. Then $M(n,k)\leq c_{q,k}^n$ for*

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

For $q = 1$, this matches the Hadamard bound, but when $q > 1$, we have that $c_{q,k} < \sqrt{k}$. Scheinerman also gave a particular sequence $q_k$ for which $c_{q_k,k} = \sqrt{k} - \frac{t_1}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$ a constant $t_1 \approx 0.096$ and thus $\rho_k \leq \sqrt{k} - \frac{0.096}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$

Scheinerman also proposed a greedy algorithm for combining his methods to provide an even tighter bound, but his analysis only provided an improvement for $k \leq 27$. In this work, we provide a tighter analysis of this greedy algorithm based on linear programming, show that it improves on Scheinerman's other techniques for bounding the determinant, and analyze the asymptotic behavior of this new bound in $k$. In particular, we show that $\rho_k \leq \sqrt{k} - \frac{0.178}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$.

We also use similar techniques to slightly improve upon Scheinerman's algorithm, and show that the resulting bound is an improvement upon the previous algorithm.

## 2.2 Decomposition Bounds

Since these techniques depend heavily on the decomposition by Scheinerman in [24], here we will give a simple overview of the needed definitions and methods as they appear in this paper.

For any $m \times n$ matrix $B$ we define $Vol(B) = \sqrt{|\det(BB^\top)|}$. Since the inner matrix product is the Gram matrix of the rows, this quantity is essentially the $m$-dimensional volume of the box with sides given by the rows of $B$.

This measure has two useful properties: First, if $A$ is an $n \times n$ square matrix, then $Vol(A) = \sqrt{|\det(AA^\top)|} = \sqrt{\det(A)^2} = |\det(A)|$. Secondly if $B_1$ and $B_2$ are $m_1 \times n$ and $m_2 \times n$ matrices respectively, then let $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$, the $(m_1 + m_2) \times n$ block matrix with $B_1$ above $B_2$. We can see that $BB^\top = \begin{bmatrix} B_1B_1^\top & B_1B_2^\top \\ B_2B_1^\top & B_2B_2^\top \end{bmatrix}$, so by Fischer's Inequality:

$$\begin{aligned} Vol(B) &= \sqrt{\left|\det\left(\begin{bmatrix} B_1B_1^\top & B_1B_2^\top \\ B_2B_1^\top & B_2B_2^\top \end{bmatrix}\right)\right|} \\ &\leq \sqrt{\left|\det\left(\begin{bmatrix} B_1B_1^\top & 0 \\ 0 & B_2B_2^\top \end{bmatrix}\right)\right|} = Vol(B_1)Vol(B_2) \end{aligned} \tag{2.1}$$

For convenience of notation we will sometimes use $M(n,k)^2$ instead of $M(n,k)$ in calculations, as $M(n,k)^2 = \max_{A \in R_{n,k}} Vol(A)^2 = \max_{A \in R_{n,k}} \det(AA^\top)$.

Given a matrix $A \in R(n,k)$, we can decompose $A$ into its rows $\vec{A_i}$. Then by repeatedly using the submultiplicativity, we can see that $Vol(A)^2 \le \prod_{i=1}^n Vol(\vec{A_i})^2 = k^n$. This was already given by Hadamard, but other decompositions will yield better results.

Scheinerman noticed that if we decompose $A$ into blocks $A_i$ of rows such that each block has a column of all ones, then if $A_i$ is an $m_i \times n$ block, then $A_i A_i^\top$ is an $m_i$ by $m_i$ matrix with $k$ on the main diagonal and strictly positive integers off of it.

Using a result by Olkin [17], Scheinerman proved that $Vol(A_i)^2 \le \det(J_{m_i} + (k-1)I_{m_i}) = (m_i + k - 1)(k-1)^{m_i - 1}$. Since the $A_i$ are a partition of the $n$ rows $\sum_i m_i = n$ and thus $Vol(A)^2 \le \prod_i (m_i + k - 1)(k-1)^{m_i - 1} = (k-1)^n \prod_i \left(1 + \frac{m_i}{k-1}\right)$.

We can see that we have at most $n$ such partitions, so we can assume that there are exactly $n$, with some allowed to be of size 0. By Jensen's inequality balancing these sizes maximizes the product, but that is the same as making each row its own partition. We are hoping to find a smaller bound, and thus go for as large of blocks as we can.

With this in mind, Scheinerman proposed the following algorithm: At each step $i$, choose the column with the most 1s in it. Let the rows that contain those ones be the block $A_i$, remove $A_i$, and then repeat the same steps on the remaining rows to partition the entire matrix. This greedy algorithm gives us a method for finding a more useful decomposition of a matrix.

In order derive a bound on this, Scheinerman noted that the average number of ones per column in an $m \times n$ matrix is $\frac{km}{n}$, so there is always a column with at least $\lceil \frac{km}{n} \rceil$ ones. While it might be possible to find a column with more ones, this is always guaranteed, and this will allow us to bound the determinant for all $A \in R(n,k)$.

## 2.3 Analysis of the Greedy Algorithm Bound

In this section, we use a new analysis technique to get the following result:

**Theorem 2.4.** *For every $k$, there is a $c_k$ such that $M(n,k) \le c_k^n$ where $c_k = \sqrt{k} -$*

$\frac{t_2}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$ *for* $t_2 = 1 - \frac{\pi^2}{12}$.

This theorem gives $\rho_k \leq \sqrt{k} - \frac{t_2}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$, slightly closing the gap towards the lower bound given by the projective plane.

To get this bound from the greedy algorithm, we will set up a linear programming problem to help us analyze it: Fix a matrix $A$ in $R(n,k)$. Let $x_j$ be the number of blocks of size $j$ in the decomposition given by Scheinerman's greedy algorithm. This gives us $n$ variables, as the blocks can be of size 1 up to $n$.

Since we always take the column with the most ones, we know that after we remove all the blocks of size $i$ or larger, the $m$ remaining rows can't have any columns with $i$ ones in them. This means the average number of ones $\frac{km}{n}$ must be no bigger than $i-1$, where we calculate the number of rows remaining by $m = n - \sum_{j=i}^{n} jx_j$. This means that the greedy decomposition of any matrix $A$ must satisfy the constraints $\sum_{j=i}^{n} kjx_j \geq (k-i+1)n$ for every $i \in [n]$.

Since the $x_i$ give us a partition of the $n$ rows, we also have that $\sum_{j=1}^{n} jx_j = n$, and that $x_i \geq 0$ for every $i$.

**Lemma 2.5.** *Every greedy algorithm decomposition of a matrix $A \in R(n,k)$ has $x_j$ blocks of size $j$, where $x_j$ is an integral feasible solution to the following LP:*

$$\begin{aligned} &\sum_{j=i}^{n} kjx_j \geq (k-i+1)n, && i = 1,...,n \\ &\sum_{j=1}^{n} jx_j = n && \\ &x_j \geq 0, && j = 1,...,n \end{aligned} \tag{2.2}$$

A given set of $x_j$ generates the determinant bound $Vol(A)^2 \leq (k-1)^n \prod_{j=1}^{n} \left(1 + \frac{j}{k-1}\right)^{x_j}$. In order to find a bound for all matrices, we want to find the largest this greedy algorithm bound could be for any matrix in $R(n,k)$.

While this bound doesn't fit the LP framework, we can remove the $(k-1)^n$ term and take the logarithm, giving that the objective function to be maximized is $\sum_{j=1}^{n} x_j \ln\left(1 + \frac{j}{k-1}\right)$.

Since the right hand side of each bound is a multiple of $n$, we can simply scale down the variables by $a_j = \frac{x_j}{n}$. In addition, since each $a_j$ is non-negative, we can see that all the constraints with $i > k$ are trivially satisfied, and can thus be dropped. Similarly, the equality constraint ensures that the $i = 1$ constraint is also trivially satisfied.

Lastly, we note that relaxing the integrality constraint can only weaken the bound, so we are left with the following lemma:

**Lemma 2.6.** *For any fixed $k$, let $\alpha$ be the optimal solution of the following LP:*

$$\begin{array}{lll} \textit{maximize} & \alpha = \sum\limits_{j=1}^{n} \ln\left(1 + \frac{j}{k-1}\right) a_j & \\ \textit{subject to} & \sum\limits_{j=i}^{n} kja_j \geq k - i + 1, & i = 2, ..., k \\ & \sum\limits_{j=1}^{n} ja_j = 1 & \\ & a_j \geq 0, & j = 1, ..., n \end{array} \tag{2.3}$$

*Then $M(n,k)^2 \leq \gamma_k^n$ where $\gamma_k = (k-1)e^{\alpha}$.*

Since there are $k$ constraints and $n$ variables, we can define the dual linear program as follows:

$$\begin{array}{lll} \text{minimize} & \beta = \sum\limits_{i=1}^{k} (k-i+1)b_i & \\ \text{subject to} & \sum\limits_{i=1}^{j} kjb_i \geq \ln\left(1 + \frac{j}{k-1}\right), & j = 1, ..., n \\ & b_i \leq 0, & i = 2, ..., k \end{array} \tag{2.4}$$

This formulation lets us give an explicit formula for the optimal $\alpha$:

**Lemma 2.7.** *The optimal solution to equation 2.3 is given by $a_j = \frac{1}{kj}$ for $1 \leq j \leq k$ and $a_j = 0$ otherwise. The optimal solution to equation 2.4 is given by $b_1 = \frac{\ln\left(1+\frac{1}{k-1}\right)}{k}$ and $b_i = \frac{\ln\left(1+\frac{i}{k-1}\right)}{ik} - \frac{\ln\left(1+\frac{i-1}{k-1}\right)}{(i-1)k}$ for $2 \leq i \leq k$.*

*Proof.* Since $a_j = 0$ for all $j > k$, we simply need to show that $\sum\limits_{j=i}^{k} kja_j \geq k - i + 1$ for each $2 \leq i \leq k$ and $\sum\limits_{j=1}^{k} ja_j = 1$. However, since $kja_j = 1$, it is easy to see that $a_j$ are a feasible solution to equation 2.3.

In fact, these $a_j$ are the solution given by making all constraints tight, and thus if we let $\vec{a} \in \mathbb{R}^k$ be the vector of the non-zero $a_j$, we see that $M\vec{a} = \vec{v}$ where $\vec{v}$ is the vector given by $v_i = k - i - 1$ and $M$ is the matrix with $M_{ij} = kj$ when $j \geq i$ and 0 otherwise.

This means that $\vec{a} = M^{-1}\vec{v}$ and so if we let $\vec{c}$ be given by $c_i = \ln\left(1 + \frac{i}{k-1}\right)$, we have that $\alpha = \vec{c}^\top \vec{a} = \vec{c}^\top M^{-1}\vec{v}$.

As for the dual solution, we can use Jensen's Inequality to verify that $b_i \leq 0$ for every $2 \leq i \leq k$. We can also see that the first $k$ constraints are actually tight. The remaining constraints hold because $\frac{\ln\left(1+\frac{j}{k-1}\right)}{j}$ is a decreasing sequence.

This means that these $b_j$ form a feasible solution, and furthermore the one given by making the first $k$ constraints tight. Thus if we let $\vec{b} \in \mathbb{R}^k$ be the vector of the $b_i$, we have that $M^\top \vec{b} = \vec{c}$. This means that $\vec{b} = (M^\top)^{-1}\vec{c}$, and thus $\beta = \vec{v}^\top \vec{b} = \vec{v}^\top (M^{-1})^\top \vec{c}$.

This means that $\alpha = \beta$, and thus by duality these sets of $a_j$ and $b_i$ are optimal solutions to their respective problems. □

This allows us to finally prove Theorem 2.4 by combining Lemma 2.6 with Lemma 2.7:

*Proof.* By Lemma 2.6, we have that $M(n,k)^2 \leq \gamma_k^n$, so $M(n,k) \leq c_k^n$ where $c_k = \sqrt{\gamma}$. By Lemma 2.7, we can write $\gamma$ and $c_k$ as follows:

$$\gamma = (k-1)e^{\left(\sum\limits_{j=1}^{k} \frac{\ln\left(1+\frac{j}{k-1}\right)}{jk}\right)}$$

$$c_k = \sqrt{k-1}e^{\frac{\sum\limits_{j=1}^{k} \frac{\ln\left(1+\frac{j}{k-1}\right)}{j}}{2k}}$$

Using $x = \frac{j}{k-1}$, we can manipulate a part of this equation to resemble a Riemann approximation of the integral $\int\limits_0^1 \frac{\ln(1+x)}{x} = \frac{\pi^2}{12}$. This gives that:

$$\begin{aligned}\sum_{j=1}^{k} \frac{\ln\left(1+\frac{j}{k-1}\right)}{j} &= \frac{\ln\left(1+\frac{k}{k-1}\right)}{k} + \frac{\sum\limits_{j=1}^{k-1} \frac{\ln\left(1+\frac{j}{k-1}\right)}{\frac{j}{k-1}}}{k-1} \\ &= \int\limits_0^1 \frac{\ln(1+x)}{x} + O(k^{-1}) \\ &= \frac{\pi^2}{12} + O(k^{-1})\end{aligned}$$

Since $e^{\frac{\pi^2}{12k}+O(k^{-2})} = 1 + \frac{\pi^2}{12k} + O(k^{-2})$, we have that $\frac{\gamma_k}{k} = 1 - \frac{1-\frac{\pi^2}{12}}{k} + O(k^{-2})$. Since $\gamma_k = c_k^2$, we can use the Taylor series decomposition of $\sqrt{1+x}$ to see that $\frac{c_k}{\sqrt{k}} = \sqrt{1 - \frac{t_2}{k} + O(k^{-2})} = 1 - \frac{t_2}{2k} + O(k^{-2})$. □

The fact that $t_2 > t_1$ implies that $c_k \leq c_{q,k}$ for large enough $k$, but we can actually prove more:

**Theorem 2.8.** *Any decomposition bound that decomposes using this greedy approach with restricted block sizes gives a bound that is no better than this greedy algorithm. In particular, $c_{q,k} \geq c_k$ for any $q \leq k$.*

*Proof.* If block size $i$ is not allowed, then the LP corresponding to Equation 2.3 for that algorithm has $a_i = 0$ and omits the inequality corresponding to $i$. Taking the dual of this new LP gives an LP whose feasible region is contained in the feasible region of Equation 2.4, because $b_i = 0$.

This means that the optimal solution of the modified primal problem is the same value as a feasible dual solution, and thus is at least as large as $\alpha$. This means that the induced bound on $M(n,k)^2$ is larger with the modified algorithm.

Scheinerman's methods [24] for $c_{q,k}$ can be framed as modified greedy algorithms, where only blocks of size 1 and $q$ are used (only $a_1$ and $a_q$ are non-zero). Therefore the unmodified greedy algorithm bound $c_k \leq c_{q,k}$ for any $k$ and $q$.

□

## 2.4 Improving the Greedy Algorithm

To attempt to do better, we try to utilize the idea that we take all the rows with a one in the chosen column, meaning that column cannot be chosen in future iterations. In fact, since the remainder of the chosen column is guaranteed to have no ones, we can remove that column from all future iterations of our algorithm without affecting the determinant.

This means that after we have removed all blocks of size at least $i$, there are still $m = n - \sum_{j=i}^{n} jx_j$ rows, but now only $n - \sum_{j=i}^{n} x_j$ columns, meaning that we can find more large blocks. This means that the average number of ones per row is $k\frac{n-\sum_{j=i}^{n} jx_j}{n-\sum_{j=i}^{n} x_j}$ which again must be no larger than $i-1$. Rearranging this similarly to Theorem 2.6 gives us the constraints in the following linear program:

**Lemma 2.9.** *For any fixed $k$, let $\alpha$ be the optimal solution of the following LP:*

$$\begin{aligned} &\textit{maximize} && \alpha' = \sum_{j=1}^{n} \ln\left(1+\tfrac{j}{k-1}\right) a'_j && \\ &\textit{subject to} && \sum_{j=i}^{n} (kj-i+1)a'_j \geq k-i+1, && i=2,...,k \\ & && \sum_{j=1}^{n} ja'_j = 1 && \\ & && a_j \geq 0, && j=1,...,n \end{aligned} \tag{2.5}$$

*Then $M(n,k)^2 \leq (\gamma'_k)^n$ where $\gamma'_k = (k-1)e^{\alpha'}$.*

We can also find the dual in the same manner as before:

$$\begin{aligned} &\text{minimize} && \beta' = \sum_{i=1}^{k} (k-i+1)b'_i && \\ &\text{subject to} && \sum_{i=1}^{j} (kj-i+1)b'_i \geq \ln\left(1+\tfrac{j}{k-1}\right), && j=1,...,n \\ & && b'_i \leq 0, && i=2,...,k \end{aligned} \tag{2.6}$$

While the analog of Lemma 2.7 holds again, the proof is a little more involved:

**Lemma 2.10.** *The optimal solution to equation 2.5 is given by $a'_j = 0$ for $j > k$ and the other $a'_j$ determined by making all $k$ non-trivial constraints tight. The optimal solution to equation 2.6 is given by the $b'_i$ that make the first $k$ constraints tight.*

*Proof.* To show feasibility of the primal solution, we need to show that each $a_j$ is positive, as the other constraints are already tight. This is trivial for each $j > k$, so we focus on the case where $j \leq k$.

To do this, we note that $\sum_{j=i}^{k}(kj-i+1)a'_j = k-i+1$ and $\sum_{j=i+1}^{k}(kj-i)a'_j = k-i$. When solving for $a'_i$, we get the recursive definition: $a'_i = \frac{1-\sum_{j=i+1}^{k} a'_j}{(k-1)i+1}$. The exact values can be computed from this, but this is unnecessary to show that $a'_i \geq 0$.

We can simply note that this means that each $a'_i$ is a small fraction of the distance between $\sum_{j=i+1}^{k} a'_j$ and 1. By induction down from $k$, this means that every $\sum_{j=i+1}^{k} a'_j \leq n$, and thus $a'_j \geq 0$ for every $j$.

In addition, we can write the non-zero $a'_j$ in the same manner as in Lemma 2.7, so $\vec{a'} = (M')^{-1}\vec{v}$ and $\alpha = \vec{c}^\top (M')^{-1}\vec{v}$, where all quantities are the same as in Lemma 2.7 except that $M'_{ij} = kj - i + 1$ for $j \geq i$ instead.

To show feasibility of the dual solution, we need to show that $b'_i \leq 0$ for each $i \geq 2$ and also that $\sum\limits_{i=1}^{j}(kj - i + 1)b'_i \geq \ln\left(1 + \frac{j}{k-1}\right)$ for each $j > k$, as the other constraints are already tight.

To show both of these things, we rely heavily on the following equation:

$$(j-1)\sum_{i=1}^{j}(kj-i+1)b'_i - j\sum_{i=1}^{j-1}(k(j-1)-i+1)b'_i = (j-1)[(k-1)(j-2)+1]b'_j + \sum_{i=1}^{j-1}(i-1)b'_i \tag{2.7}$$

To show that $b'_j \leq 0$ for $2 \leq j \leq k$, we simply notice that $\sum_{i=1}^{j}(kj - i + 1)b'_i = \ln\left(1 + \frac{j}{k-1}\right)$ and $\sum_{i=1}^{j-1}(k(j-1)-i+1)b'_i = \ln\left(1 + \frac{j-1}{k-1}\right)$ because $j \leq k$. This means that the left hand side is equal to $(j-1)\ln\left(1 + \frac{j}{k-1}\right) - j\ln\left(1 + \frac{j-1}{k-1}\right)$, so in order to show $b'_j$ is negative, we need to show that $\sum_{i=1}^{j-1}(i-1)b'_i \geq (j-1)\ln\left(1 + \frac{j}{k-1}\right) - j\ln\left(1 + \frac{j-1}{k-1}\right)$ for each $j \geq 2$.

We show this by induction on $j$. The base case is when $j = 2$, where the left hand side is 0. A simple application of Jensen's inequality gives that the right hand side is negative, and thus $b'_2 \leq 0$.

Now inductively assume that $\sum_{i=1}^{j-1}(i-1)b'_i \geq (j-1)\ln\left(1 + \frac{j}{k-1}\right) - j\ln\left(1 + \frac{j-1}{k-1}\right)$, and thus that $b'_j \leq 0$. Using Jensen, we have that:

$$\begin{aligned}&\left[j\ln\left(1+\frac{j+1}{k-1}\right) - (j+1)\ln\left(1+\frac{j}{k-1}\right)\right]\\ &\qquad - \left[(j+1)\ln\left(1+\frac{j+2}{k-1}\right) - (j+2)\ln\left(1+\frac{j+1}{k-1}\right)\right]\\ &\qquad = (j+1)[2\ln\left(1+\frac{j+1}{k-1}\right) - \ln\left(1+\frac{j}{k-1}\right) - \ln\left(1+\frac{j+2}{k-1}\right)] > 0\end{aligned} \tag{2.8}$$

This means $(j-1)[(k-1)(j-2)+1]b'_j \leq (j-1)b'_j$, so using our constraints we get that $\sum_{i=1}^{j}(i-1)b'_i \geq (j-1)[(k-1)(j-2)+1]b'_j + \sum_{i=1}^{j-1}(i-1)b'_i$. By equation 2.8, we have that $j\ln\left(1 + \frac{j+1}{k-1}\right) - (j+1)\ln\left(1 + \frac{j}{k-1}\right) \geq (j-1)\ln\left(1 + \frac{j}{k-1}\right) - j\ln\left(1 + \frac{j-1}{k-1}\right)$, so we have our induction.

This gives that $b'_j \leq 0$ for $j \geq 2$. To show the other constraints are satisfied, we simply leverage equations 2.7 and 2.8 differently.

We prove that $\sum_{i=1}^{j}(kj-i+1)b'_i \geq \ln\left(1+\frac{j}{k-1}\right)$ inductively, where the base case is the given equality when $j=k$. Assuming truth for $j$, the inductive step is:

$$\begin{aligned}
&\sum_{i=1}^{j+1}(k(j+1)-i+1)b'_i \\
&= \frac{j[(k-1)(j-1)+1]b'_{j+1} + \sum_{i=1}^{j}(i-1)b'_i + (j+1)\sum_{i=1}^{j}(kj-i+1)b'_i}{j} \\
&\geq \frac{j\ln\left(1+\frac{j+1}{k-1}\right) - (j+1)\ln\left(1+\frac{j}{k-1}\right) + (j+1)\ln\left(1+\frac{j}{k-1}\right)}{j} \\
&= \ln\left(1+\frac{j+1}{k-1}\right)
\end{aligned}$$

This means that the $b'_i$ are a feasible solution to the dual, and so like in Lemma 2.7, we have $\vec{b'} = (M'^{\top})^{-1}\vec{c}$ and $\beta' = \vec{v}^{\top}\vec{b'} = \vec{v}^{\top}((M')^{-1})^{\top}\vec{c}$. Since $\alpha' = \beta'$, both are optimal. □

As in Theorem 2.4, we now can say that:

**Theorem 2.11.** *For every $k$, there is a $c'_k$ such that $M(n,k) \leq c'_k$ where $c'_k = \sqrt{k} - \frac{t_3}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$ for $t_3 \approx 0.178$.*

*Proof.* Using Lemma 2.9 and Lemma 2.10, we have that $M(n,k) \leq (c'_k)^n$ for $c'_k = \sqrt{(k-1)e^{\alpha'}}$. □

Asymptotically, this also means that $\rho_k \leq \sqrt{k} - \frac{t_3}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$.

Experimentally $t_3 \approx 0.178$, but the recursive nature of the computations for finding the $a_i$ (shown in the proof of Lemma 2.10) makes explicit calculation of $t_3$ difficult. While this seems identical to Theorem 2.4, we can prove that this new $c'_k$ is a strict improvement on the $c_k$ from the previous analysis.

**Theorem 2.12.** *The optimal solution $\alpha'$ to equation 2.5 is strictly smaller than $\alpha$ in equation 2.3. Therefore, the improved algorithm gives a strictly better bound on $M(n,k)$.*

*Proof.* Since $\alpha = \sum_{j=1}^{k} a_j \ln\left(1+\frac{j}{k-1}\right)$, the difference is $\sum_{j=1}^{k}\ln\left(1+\frac{j}{k-1}\right)\left(a_j - \frac{1}{jk}\right)$. Rearranging the sums gives that this difference is:

$$\ln\left(1+\frac{1}{k-1}\right)\left(\sum_{j=1}^{k} ja'_j-1\right)+\sum_{i=2}^{k}\frac{\ln\left(1+\frac{i}{k-1}\right)}{i}-\frac{\ln\left(1+\frac{i-1}{k-1}\right)}{i-1}\left(\sum_{j=i}^{k} ja'_j-\frac{k-i+1}{k}\right) \tag{2.9}$$

Because $\sum_{j=1}^{k} ja'_j = 1$, the first term is simply 0. Using the other constraints in equation 2.5, we know that $\sum_{j=i}^{k}(kj-i+1)a'_j = k-i+1$. This means that $\sum_{j=i}^{k} ja'_j - \frac{k-i+1}{k} = \frac{i-1}{k}\sum_{j=i}^{k} a'_j > 0$ for each $2 \leq i \leq k$.

Using Jensen's inequality, we can see that $\frac{\ln\left(1+\frac{i}{k-1}\right)}{i} - \frac{\ln\left(1+\frac{i-1}{k-1}\right)}{i-1} < 0$ for every $i \geq 2$, so the other $k-1$ terms are negative. This means that $\alpha' < \alpha$ , and so $c'_k < c_k$ is an improvement on our earlier algorithm. □

Experimentally, this difference appears to be $O(k^{-2})$, and thus has no effect on the asymptotic convergence in Theorem 2.4, so $t_2 = t_3$. A plot of this is shown below in Figure 2.1:

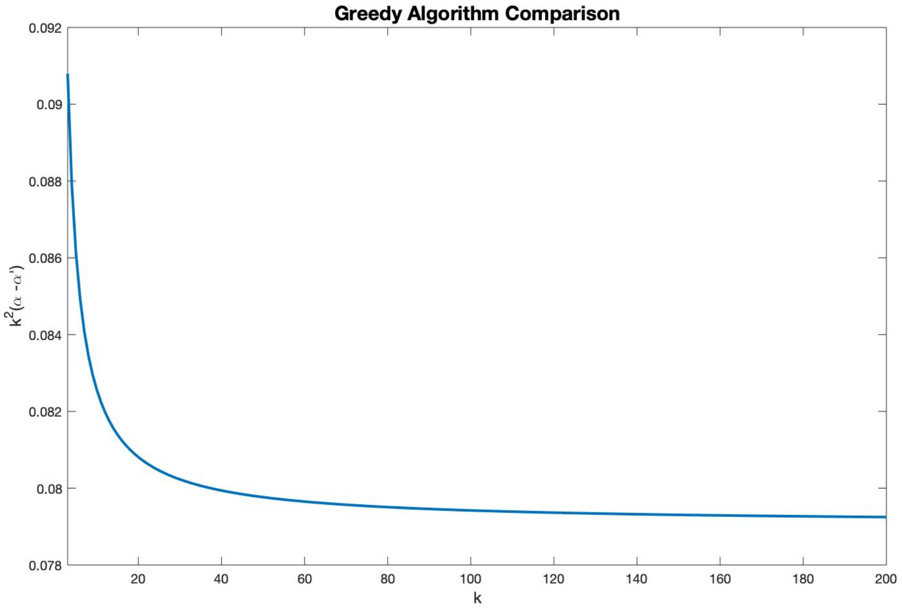

Figure 2.1: The differences between the exponents of the algorithms described in Theorems 2.4 and 2.11. The computed line represents $k^2$ times the difference computed in Theorem 2.12, seemingly showing that $\alpha - \alpha' = O(k^{-2})$ and thus $t_2 = t_3$.

## 2.5 Summary of Results

While it is impossible to improve on Ryser's bound when $\lambda = \frac{k(k-1)}{n-1}$ is $\Omega(1)$ ($k = \Omega(\sqrt{n})$), for fixed $k$ we have can see the improvement on the previously known bounds as shown in Table 2.1 below for $k = 3$ and various $n$:

Table 2.1: Generated Bounds on $M(n, 3)$

| Values of $n$ | 7 | 14 | 21 | 28 |
|---|---|---|---|---|
| Hadamard [10] | 46.7654 | $2.1870 \times 10^3$ | $1.0228 \times 10^5$ | $4.7830 \times 10^6$ |
| Ryser [21] | 24 | $1.2789 \times 10^3$ | $0.6177 \times 10^5$ | $2.9311 \times 10^6$ |
| Scheinerman [24] | 40.7612 | $1.6615 \times 10^3$ | $0.6772 \times 10^5$ | $2.7605 \times 10^6$ |
| Theorem 2.4 | 38.8508 | $1.5094 \times 10^3$ | $0.5864 \times 10^5$ | $2.2782 \times 10^6$ |
| Theorem 2.11 | 37.5030 | $1.4065 \times 10^3$ | $0.5275 \times 10^5$ | $1.9782 \times 10^6$ |

Note that Ryser's bound [21] outperforms the others at small $n$, but rapidly approaches the Hadamard bound [10] as $n$ gets larger.

By plotting the exponential growth factors as $k$ gets larger, we can see how each iteration improves upon the earlier bounds on $\rho_k$, as shown in Figure 2.2. The actual asymptotic bounds are laid out in Table 2.2:

Table 2.2: Known Asymptotic Bounds on $\rho_k$

| | $M(n, k)$ bound | Asymptotic bound on $\rho_k$ |
|---|---|---|
| Hadamard [10] | $k^{\frac{n}{2}}$ | $\sqrt{k}$ |
| Ryser [21] | $k(k-\lambda)^{\frac{n-1}{2}}$ | $\sqrt{k}$ |
| Scheinerman [24] | $(c_{2,k})^n$ | $\sqrt{k} - \frac{1}{4k^{\frac{3}{2}}} + O(k^{-3})$ |
| Scheinerman [24] | $(c_{q_k,k})^n$ | $\sqrt{k} - \frac{.1}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$ |
| Theorem 2.4 | $(c_k)^n$ | $\sqrt{k} - \frac{.18}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$ |
| Theorem 2.11 | $(c'_k)^n$ | $\sqrt{k} - \frac{.18}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$[1] |

None of these approaches achieve the known lower bound on $\rho_k$ given by a block diagonal matrix of projective planes ($\rho_k \geq \sqrt{k} - \frac{1}{2\sqrt{k}} + O(k^{\frac{-3}{2}})$), but the gap is closing.

[1] Constant is conjectured

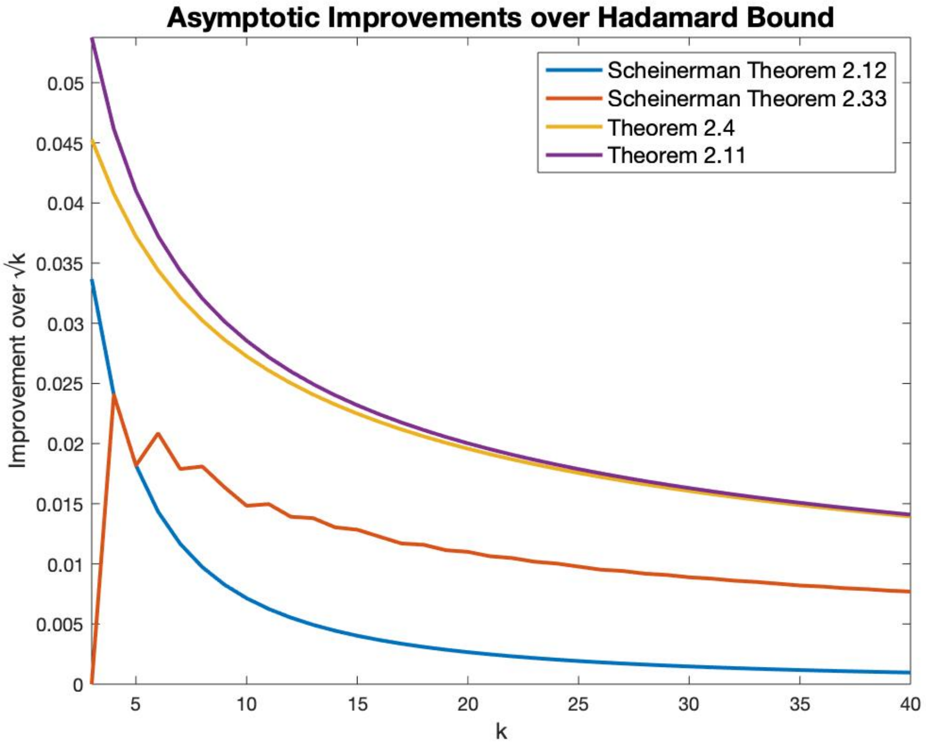

Figure 2.2: The improvements over $\sqrt{k}$ of various theorems' bounds on $\rho_k$ from $k = 3$ to $k = 40$. Ryser's bound is not included as it is asymptotically no better than the Hadamard bound (which is the $x$ axis).

### 2.5.1 Limitations and avenues for improvement

Note that these bounds have three places where there could be loss:

1. In Fischer's inequality for bounding the volume when subdividing into blocks.
2. In the determinant estimation within a single block.
3. In the LP relaxation of the integer program.

Since all constraints are integral in Equations 2.3 and 2.5, the optimal $a_j$ are always rational. This means that if we choose an $n$ that is a multiple of the least common denominator, the LP solution is then an integral solution. For instance, if we consider Equation 2.5 with $k = 3$, using $n = 35$ means we can decompose the rows into 5 blocks of 3, 6 blocks of 2, and 8 blocks of 1. For $k = 4$, we can use $n = 455$ for the same result.

The determinant estimation is tight within a single block $A_i$ when $A_iA_i^\top = J_{m_i} + (k-1)I_{m_i}$. Because $A_iA_i^\top$ is a Gram matrix, and there is a column of $A_i$ that is all 1, we simply need to place the other $k-1$ ones in each row such that no two are in the same column.

In the case where Equation 2.5 is tight that is described above, we have that $\sum\limits_{j=i}^{n}(kj - i+1)x_j = (k-i+1)n$ for each $1 \leq i \leq k$. If we consider all the rows in all the $x_i$ blocks of size $i$, there are $ix_i$ rows where we need to place $k-1$ ones in each. Since none of the chosen rows can have any more ones in them there are $\sum_{j=i}^{n} x_j$ rows we can't use. As in the proof of Lemma 2.10, we know that $a_i = \frac{1-\sum_{j=i+1}^{k} a_j}{(k-1)i+1}$. Rescaling by $n$ gives that $n - \sum_{j=i}^{k} x_j = ix_i(k-1)$, so we only need to put a single one in each column for all $x_i$ blocks!

This means that our determinant estimation can be tight inside every block. In addition, this means that if we have two blocks of size $i$, and take a vector from each block, they will always be perpendicular. This means that $B_1B_2 = 0$ in Fischer's inequality for these blocks as well, so our bound is tight there.

Since each set of blocks puts a one in every row that wasn't chosen, clearly this part of Fischer's inequality isn't tight when the blocks are different sizes. Looking closer shows that each column also has $k$ non-zero entries in this example.

This suggests that the first type of error is where any improvement could be made to this bound, as opposed to the other two. However, it is likely a more nuanced approach will need to be used to balance these different errors.

To see this, we randomly generated many examples of this form for $k = 3$, and found that the maximum of their determinants never exceeded the asymptotic lower bound given by the Fano plane (which is $24^{\frac{n}{7}}$).

This suggests that the projective plane (or a block matrix of them) has the (asymptotically) largest determinant of any matrix in $T(n, k)$. When examining the execution of the greedy algorithm on a projective plane, the decomposition is 1 block of size $k$ and $k-1$ blocks of size $k-1$. Only the $i = k$ inequality is tight here, and so the projective plane is nowhere close the LP bound.

This block decomposition still has that the determinant estimation is tight within each block, but there still is significant error coming from the volume submultiplicativity. Even if we set $x_j$ to the number of blocks from aboce, the volume bound we get for the projective plane of with $k$ ones in each row (order $k-1$) is $(k-1)^{k^2-k+1}\frac{2k-1}{k-1}\left(\frac{2k-2}{k-1}\right)^{k-1} = (2k-1)2^{k-1}(k-1)^{k^2-k-1}$. This is much larger than the actual volume which is $k^2(k-1)^{k^2-k}$, and the loss is approximately a factor of $\frac{2^k}{k^2}$.

## 2.6 Related Classes of Matrices

Ryser [21], Bruhn and Rautenbach [5], and Scheinerman [24] also defined two other related classes of matrices, which they entitled $S(n,k)$ and $T(n,k)$ as follows:

**Definition 2.13.** $S(n,k) = \{A \in R(n,k) : A^\top \in R(n,k)\}$.

**Definition 2.14.** $T(n,k) = \left\{A \in M_{n\times n}(\{0,1\}) : \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} A_{ij} = kn\right\}$.

Since $S(n,k)$ is the set of matrices with $k$ ones in each row and $k$ ones in each column, and $T(n,k)$ is the set of matrices with a total of $kn$ ones, clearly $S(n,k) \subset R(n,k) \subset T(n,k)$.

When asking for the maximum determinants of these other classes of matrices we define the following notation:

**Definition 2.15.** $M_S(n,k) = \max_{A\in S(n,k)} |\det(A)|$

**Definition 2.16.** $M_T(n,k) = \max_{A\in T(n,k)} |\det(A)|$

By inclusion we have that $M_S(n,k) \leq M(n,k) \leq M_T(n,k)$. Using concavity and Hadamard's bound we can see that $M_T(n,k) \leq \left(\sqrt{k}\right)^n$ as well. Ryser [21], Bruhn and Rautenbach [5], and Scheinerman [24] each showed slightly more.

The same theorem by Ryser [21] used above applies to any matrix in $T(n,k)$, giving that $M_T(n,k) \leq k(k-\lambda)^{\frac{n-1}{2}}$ for $\lambda = \frac{k(k-1)}{n-1}$, slightly improving over Hadamard. In addition, they showed the tight examples are exactly the combinatorial designs with those parameters, indirectly proving that $M_S(n,k) = M(n,k) = M_T(n,k)$ when such a combinatorial design exists.

While this is an improvement on Hadamard, in the fixed $k$ regime the question of equality remained open until Bruhn and Rautenbach [5] showed that $M_T(n,2) \leq (\sqrt[6]{6})^n$, giving an exponential improvement on Hadamard's bound. Scheinerman [24] was able to generalize this to $M_T(n,k) \leq c_k^n$ for $c_k = \sqrt{k^2-1}^{\frac{k-1}{2k}} k^{\frac{1}{2k}}$.

A linear programming approach similar to Theorem 2.4 can be used here giving the following result:

**Theorem 2.17.** *Let $\kappa$ be the optimal objective value for the following linear program:*

$$\begin{array}{lll} \textit{maximize} & \kappa = \sum\limits_{l=2}^{n} p_l \ln(l-1) + \sum\limits_{j=1}^{n} t_{jl} \ln\left(1 + \frac{j}{l-1}\right) & \\ \textit{subject to} & \sum\limits_{j=i}^{n} l j t_{jl} \geq l p_l - i + 1, & l = 1, ..., n, i = 2, ..., n \\ & \sum\limits_{j=1}^{n} j t_{jl} = p_l & l = 1, ..., n \\ & \sum\limits_{l=1}^{n} l p_l = k & \\ & \sum\limits_{l=1}^{n} p_l = 1 & \\ & p_l, x_{jl} \geq 0, & j = 1, ..., n, l = 1, ..., n \end{array} \tag{2.10}$$

*Then $M_T(n,k)^2 \leq e^{\kappa n}$*

*Proof.* The methodology used here is similar to Theorem 2.6, except that now we can have anywhere from 1 to $n$ ones per row. To that end, we break up the matrix into $n$ sets of rows, where each set contains all the rows with a certain number of ones. We denote the proportion of rows with $l$ ones by $p_l$, giving the last two constraints.

Then we let $t_{jl}$ be the number of blocks of size $j$ in the $l$th set, and derive the remaining constraints in the same way as in the proof of Lemma 2.6. Since we can bound the determinant by the product of the volumes of each set of rows for each $l$, we simply take the logarithm to get our objective function. □

Scheinerman's bound [24] can be rewritten as a dual feasible solution in a similar way to Theorem 2.8, meaning that Theorem 2.10 is an improvement, although further analysis is needed to quantify the difference.

It is possible to use the improvements used for Theorem 2.11 to improve the bounds within each set of rows, although there is still error from using Fisher's inequality to combine the bounds from each set.

Scheinerman [24] also conjectured that $\lim_{n\to\infty} M_S(n,k)^{\frac{1}{n}} = \rho_k = \lim_{n\to\infty} M_T(n,k)^{\frac{1}{n}}$ for all $k \geq 2$, citing computational evidence that the projective plane was determinant maximizing in all three subsets of matrices.

In the previous section, the construction of a matrix in $R(n,k)$ for which all estimations are tight save that from Fisher's inequality is also contained in $S(n,k)$, further suggesting that $S(n,k)$ likely contains the maximizing examples and thus $M_S(n,k) = M(n,k)$. Even though these matrices minimize some of the errors in the algorithm, experimentally their determinants are smaller than that of a block matrix of projective planes, implying that those are the determinant maximizing matrices, and suggesting Scheinerman's conjecture is true.

# Chapter 3

# Supports of Binary Polynomials

## 3.1 Polynomials on $\mathbb{F}_2$

Finding and counting the zeros of polynomials over various structures has been used throughout many areas of mathematics, ranging from algebra to complex analysis. Computer science also leans heavily on the mathematics of polynomials, with applications to circuits or codes.

When discussing circuit complexity, we often represent the input wires to a circuit as variables that take on values in $\mathbb{F}_2$ to simulate boolean logic. The gates of the circuit can now be represented as arithmetic operations in $\mathbb{F}_2$, and thus the entire circuit can be written as a polynomial. This essential idea has been used for proving many results in complexity theory, including some in PCPs and other probabilistic complexity classes. Model examples are Razborov-Smolensky [20],[27] which proves lower bounds on $AC^0$ circuit complexity.

In addition, the area of error correcting codes often relies on polynomials to construct good rate codes with large distance. Reed-Muller, Reed-Solomon, and BCH codes all utilize the properties of polynomials in finite fields to get their properties. This paper builds on the methods in [11] and [12], which are also focused on Reed-Muller codes.

In all these applications, an essential result that is often used for polynomials over finite fields is the Schwartz-Zippel Lemma [30],[25],[7]:

**Theorem 3.1.** *If $f \in \mathbb{F}_q[X_1, \ldots, X_n]$ is a nonzero polynomial of total degree at most $d$, then*

$$\underset{\vec{x} \in \mathbb{F}_q^n}{P}(f(\vec{x}) = 0) \leq \frac{d}{q}$$

However, when $q = 2$, this bound only is meaningful for linear polynomials, so the following generalization can be used:

**Theorem 3.2.** *If $f \in \mathbb{F}_2[X_1, \ldots, X_n]$ is a nonzero polynomial of total degree at most $d$, then*

$$\underset{\vec{x} \in \mathbb{F}_2^n}{P}(f(\vec{x}) = 0) \leq 1 - 2^{-d}$$

Both of these bounds are tight, exhibited by the polynomial $f(\vec{X}) = \prod_{i=1}^{d} X_i$. Kasami and Tokura [11] focused on the $q = 2$ case and fully characterized all tight examples for Theorem 3.2:

**Theorem 3.3.** *If $f \in \mathbb{F}_2[X_1, \ldots, X_n]$ is a nonzero polynomial of total degree at most $d$ where $P_{\vec{x} \in \mathbb{F}_2^n}(f(\vec{x}) = 0) = 1 - 2^{-d}$, then $f$ is the product of $d$ linearly independent linear functions.*

In this chapter we develop similar results for when $\vec{x}$ is constrained to within a certain Hamming weight, both giving a bound and characterizing the tight examples.

## 3.2 Preliminaries

Given a polynomial $f \in \mathbb{F}_2[X_1, \ldots, X_n]$ and a non-negative integer $r$ no larger than $n$, we define $|f|_r = |\{\vec{x} \in \mathbb{F}_2 : \|\vec{x}\|_H \leq r, f(\vec{x}) = 1\}|$. This is similar to the notation of Kasami and Tokura [11] , although the subscript has a different meaning.

Since we only are considering the evaluations of polynomials in $\mathbb{F}_2[X_1, \ldots, X_n]$ on $\mathbb{F}_2^n$, we can utilize that $x_i^2 = x_i$ for every $i$ and $\vec{x} \in \mathbb{F}_2^n$. We can multilinearize polynomials by reducing them modulo $X_i^2 + X_i$ and the support is unchanged. To this end, we treat all polynomials as elements of the polynomial ring $M_n$, as defined below:

**Definition 3.4.** $M_n := \mathbb{F}_2[X_1, \ldots, X_n]/\langle X_i^2 + X_i : i \in [n]\rangle$

Even though the following theorems are written for polynomials in $M_n$, since multilinearization can only reduce the total degree of a polynomial, the same results also apply to polynomials in $\mathbb{F}_2[X_1, \ldots, X_n]$. This allows us to rewrite Theorem 3.3 as follows:

**Theorem 3.5.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_n = 2^{n-d}$, then $f$ is the product of $d$ independent linear functions.*

To prove this, we prove a division lemma that uses the following notion of divisibility:

**Definition 3.6.** *A polynomial $g \in M_n$ is ml-divisible by a polynomial $f \in M_n$ if and only if there is a polynomial $h \in M_n$ with $\deg(h) \leq \deg(g) - \deg(f)$ such that $g = f \cdot h$ in $M_n$.*

**Lemma 3.7.** *If $f, l \in M_n$ are polynomials such that $\deg(l) = 1, \deg(f) \leq n$, and $f(\vec{x}) = 0 \, \forall \vec{x} \in \mathbb{F}_2^n : l(\vec{x}) = 0$, then $l$ ml-divides $f$.*

*Proof.* Since $l$ depends on at least one variable, there is a change of variables such that $l(\vec{X}) = X_n$. The linearity of $l$ means that the degree of $f$ is preserved in the new variables, and thus all the assumptions still hold.

This means that $f$ is zero when $X_n = 0$, and (potentially) non-zero only on the subspace $X_n = 1$. There is some $h \in M_{n-1}$ such that $h(y_1, \dots, y_{n-1}) = f(y_1, \dots, y_{n-1}, 1)$ for all $\vec{y} \in \vec{F}_2^{n-1}$. Trivially this means that $f(\vec{x}) = l(\vec{x})h(\vec{x})$, and so since both polynomials are of degree less than or equal to $n$, they are equal in $M_n$. This means that $\deg(h) = \deg(f) - 1$, and this is preserved when changing back to the original variables. □

We also utilise the following simple lemma:

**Lemma 3.8.** *If $f \in M_n$ is ml-divisible by $l_1, l_2, \dots, l_t \in M_n$ where $\deg(l_i) = 1 \, \forall \, i \in [t]$ and all $l_i$ are linearly independent, then $f$ is ml-divisible by $\prod_{i=1}^{t} l_i$.*

*Proof.* Since the $l_i$ are linearly independent, there is an affine change of coordinates $\Phi : \mathbb{F}_2^n \to \mathbb{F}_2^n$ that maps $l_i \circ \Phi = X_i$ for every $i \in [t]$.

Now $g = f \circ \Phi$ must be ml-divisible by all of the $l_i \circ \Phi$, and so we can let $h \in M_{n-t}$ be the unique function such that $h(y_1, \dots, y_{n-t}) = f(1, \dots, 1, y_1, \dots, y_{n-t})$.

Similar to the proof of Lemma 3.7, we can see that $g = h \cdot \prod_{i=1}^{t} X_i$, and so reversing the change of variables gives that $\prod_{i=1}^{t} l_i$ ml-divides $f$. □

The proof of Theorem 3.5 is fairly simple using Lemmas 3.7 and 3.8:

*Proof.* This proof goes via induction on $n$. The base case is $n = 0$, where the only nonzero function is identically 1. Trivially this satisfies all the conditions and conclusions of the theorem.

Otherwise we can relabel the variables such that $f$ depends on the variable $X_n$. Now define two functions $f_0, f_1 \in M_{n-1}$ such that $f(\vec{X}) = f_0(\vec{X}) + X_n f_1(\vec{X})$. Note that $f_0$ is of degree at most $d$, while $f_1$ is of degree at most $d-1$ and is nonzero.

If $x_n$ is zero, then $f(\vec{x}) = f_0(\vec{x})$. Otherwise $f(\vec{x}) = f_0(\vec{x}) + f_1(\vec{x})$, so by separating the cube into these two halves we see that:

$$|f|_n = |f_0|_{n-1} + |f_0 + f_1|_{n-1} = 2|f_0|_{n-1} + |f_1|_{n-1} - 2|f_0 f_1|_{n-1}$$

Clearly $|f_0|_{n-1} \geq |f_0 f_1|_{n-1}$, so we can see that $|f_1|_{n-1} \leq 2^{n-d}$. However, since $f_1$ is nonzero of degree $d-1$, the Lemma 3.2 gives that $|f_1|_{n-1} \geq 2^{(n-1)-(d-1)}$. By our inductive assumption, this means that $f_1$ is the product of $d-1$ independent linear functions $L_1, \ldots, L_{n-1}$.

Using the equation above, we can now see that $|f_0|_{n-1} = |f_0 f_1|_{n-1}$, meaning $f_0(\vec{x}) = 0$ when $f_1(\vec{x}) = 0$. Since $f_1$ is the product of $L_1, \ldots, L_{n-1}$, each of these linear functions have the same property. By Lemma 3.7, each of the $L_i$ ml-divide $f_0$, so we can use Lemma 3.8 to see that $f_0 = h \cdot \prod_{i=1}^{d-1} L_i$ for some $\deg(h) \leq 1$.

This means that $f = (h + X_n) \cdot \prod_{i=1}^{d-1} L_i$. Since all the $L_i$ do not depend on $X_n$, it is clear that $h + X_n$ is a linear function independent of all the $L_i$, regardless of $h$. □

The main theorems in this paper follow a similar structure to attain more general results. We will use the following notation to denote the numbers used in the following theorems: Let $\binom{a}{\leq b} = \sum_{i=0}^{b} \binom{a}{i}$ for any integral $a$ and $b$. Now we prove the following analog of the Schwartz-Zippel lemma:

**Theorem 3.9.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$, then $|f|_r \geq \binom{n-d}{\leq r-d}$ for any $r$.*

*Proof.* This is can proven via induction on $d$, with the base case being $d = 0$. The only nonzero polynomial of degree $d$ is identically 1, which trivially satisfies all the constraints.

Otherwise we assume $f$ depends on some variable, which we may assume is $X_n$. Then define $f_0$ and $f_1$ as above, only now we note that $|f|_r = |f_0|_r + |f_0 + f_1|_{r-1} = |f_0|_r + |f_0|_{r-1} + |f_1|_{r-1} - 2|f_0 f_1|_{r-1}$ since we need to account for the Hamming weight when $x_n = 1$.

We note that trivially $|f_0|_r \geq |f_0|_{r-1} \geq |f_0 f_1|_{r-1}$, so we know that $|f|_r \geq |f_1|_{r-1}$. Since $f_1$ is a non-zero polynomial of degree $d-1$ on $n-1$ variables, by induction we know that $|f_1|_{r-1} \geq \binom{(n-1)-(d-1)}{\leq (r-1)-(d-1)} = \binom{n-d}{\leq r-d}$. □

When $r = n$, we recover Theorem 3.2. This also provides a more exact proof of a lemma used in [14], showing that any nonzero polynomial of degree $d$ cannot vanish on any ball of radius $\geq d$.

In addition, we can see that these bounds are tight for every $r, d \leq n$: For a given $d$, the function $f(\vec{X}) = \prod_{i=1}^{d} X_i$ is tight for every $0 \leq r \leq n$. However, this is not the only minimizing polynomial, as described in the following theorem:

**Theorem 3.10.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_r = \binom{n-d}{\leq r-d}$, then $f$ is of one of the following forms depending on how large $r$ is relative to $n$ and $d$:*

1. *If $r = n$ then $f(\vec{X}) = \prod_{i=1}^{d} L_i(\vec{X})$ for $L_i(\vec{X})$ linear functions that are mutually independent.*

2. *If $r = n-1$ then $f(\vec{X}) = \prod_{i=1}^{d} L_i(\vec{X})$ for $L_i(\vec{X})$ linear functions that are mutually independent and $L_i(\vec{1}) = 1$ for all $i$.*

3. *If $d < r < n-1$, then either $f(\vec{X}) = \prod_{i \in S} X_i$ for some $S \subset [n]$ with $|S| = d$ or $f(\vec{X}) = (r + \sum_{i=1}^{n} X_i) \prod_{i \in S} X_i$ for some $S \subset [n]$ and $|S| = d-1$.*

4. *If $r = d$, then $f(\vec{X}) = \sum_{T \supseteq S : |T| \leq d} \prod_{i \in T} X_i$ for some $S \subset [n]$ with $|S| \leq d$.*

5. *If $r < d$, then $f(\vec{x})$ is a nonzero linear combination of the monomials $\prod_{i \in S} X_i$ for $r < |S| \leq d$.*

The theorem is proved in full in the following section through various lemmas.

## 3.3 Structure of Nonzero Polynomials with Minimal Support on Vectors of Hamming Weight Less Than $r$

To prove the previous structure theorem, we address each case independently. The first case was already proven by Kasami and Tokura [11], and a proof is shown in Theorem 3.5. The second case is proven similarly:

**Lemma 3.11.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_{n-1} = \binom{n-d}{\leq n-d-1} = 2^{n-d} - 1$, then $f(\vec{X}) = \prod_{i=1}^{d} L_i(\vec{X})$ for $L_i(\vec{X})$ linear functions that are mutually independent and $L_i(\vec{1}) = 1$ for all $i$.*

*Proof.* Since $f$ is a nonzero polynomial, Theorem 3.2 gives that $|f|_n \geq 2^{n-d}$. But since the only point not counted in $|f|_{n-1}$ is $\vec{1}$, we have that $|f|_n \leq |f|_{n-1} + 1$. This gives that $|f|_n = 2^{n-d}$ as well as $f(\vec{1}) = 1$.

By Theorem 3.5, the first equality shows $f(\vec{X}) = \prod_{i=1}^{d} L_i(\vec{X})$ for $L_i(\vec{X})$ linear functions that are mutually independent. Since $f(\vec{1}) = \prod_{i=1}^{d} L_i(\vec{1}) = 1, L_i(\vec{1}) = 1$ for every $i$. □

When $r < n - 1$, we have much fewer minimizing polynomials, and they come in exactly two forms. While it is easy to see that the product of $d$ variables has exactly the minimum support size, we have another specific linear factor that is allowed: $(r + \sum_{i=1}^{n} X_i)$. This factor makes $f$ vanish on points whose Hamming weight is the same parity as $r$, which means that the support (when $f$ is degree $d$) is $\sum_{i=1}^{\lfloor \frac{r-d}{2} \rfloor} \binom{n-d+1}{r-d-2i}$. Using the binomial identity $\binom{a}{b} = \binom{a-1}{b} + \binom{a-1}{b-1}$, we can verify that this is equal to $\binom{n-d}{\leq r-d}$.

To prove these are the only minimizing polynomials, we cannot use the same methodology as Lemma 3.11. Instead we use a method similar to Theorem 3.9.

**Lemma 3.12.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_r = \binom{n-d}{\leq r-d}$ for $d < r < n-1$, then either $f(\vec{X}) = \prod_{i \in S} X_i$ for some $S \subset [n]$ with $|S| = d$ or $f(\vec{X}) = (r + \sum_{i=1}^n X_i) \prod_{i \in S} X_i$ for some $S \subset [n]$ and $|S| = d-1$.*

In order to prove this, we also need a division lemma similar to Lemma 3.7:

**Lemma 3.13.** *Let $f \in M_n$ with $\deg(f) \leq t$ and $l \in \{X_i : 1 \leq i \leq n\}$ or $l(\vec{X}) = t + \sum_{i=1}^n X_i$. If $f(\vec{x}) = 0 \,\forall\, \|\vec{x}\|_H \leq t$ with $l(\vec{x}) = 0$, then $l$ ml-divides $f$.*

*Proof.* We begin with the case where $l = X_i$ for some $i$. Assume for the sake of contradiction that there is a monomial in $f$ of the form $\prod_{i \in T} X_i$ for some $T \subset [n] \setminus \{i\}$. Since $\deg(f) \leq t$, we have that $|T| \leq t$.

In this case, there is a minimal monomial of this type, so let $T$ be minimal. This means that $f(\vec{1}_T) = 1$, since no other monomial evaluates to 1. However, $l(\vec{1}_T) = 0$ and $\|\vec{1}_T\| = |T| \leq t$, so by our assumptions $f(\vec{1}_T) = 0$.

This contradiction gives that there are no monomials in $f$ that do not contain $X_i$, so by factoring out that variable from each monomial we get that $f(\vec{x}) = X_i h(\vec{X})$ with $\deg(h) \leq \deg(f) - 1$.

If $l = t + \sum_{i=1}^n X_i$, we let $\Psi$ be a change of basis given by the involution:

$$\Psi(X_1, \ldots, X_{n-1}, X_n) = (X_1, \ldots, X_{n-1}, t + \sum_{i=1}^n X_i)$$

This change of basis preserves degree, but $\Psi(\vec{x})$ need not have the same Hamming weight as $\vec{x}$. However, as long as $\|\vec{x}\|_H \leq t$, we have that $\|\Psi(\vec{x})\|_H \leq t$, as $\Psi(\vec{x})_n = 0$ if $\|\vec{x}\|_H = t$.

This means that $l(\Psi(\vec{X}))$ and $f(\Psi(\vec{X}))$ satisfy the same assumptions of the lemma, only now we also have that $l(\Psi(\vec{X})) = X_n$. By the proof above, $f(\Psi(\vec{X})) = l(\Psi(\vec{X}))h(\Psi(\vec{X}))$, so $f(\vec{X}) = l(\vec{X})h(\vec{X})$. Since the change of variables preserves degree, we still have $\deg(h) \leq \deg(f) - 1$. □

To prove Lemma 3.12, Lemma 3.13 is used in a similar way as Lemma 3.7 is in Theorem 3.5, although more computation is required to perform the inductive step:

*Proof of Lemma 3.12.* The proof is by induction on $d$, where the base case is $d = 0$. The only non-zero polynomial of degree 0 is identically 1, and this trivially is in the first form allowed.

Similar to Theorem 3.5, assume that $f(\vec{X})$ depends on $X_n$, and define the polynomials $f_0$ and $f_1$ as before. As before we know that $|f_0|_r \geq |f_0|_{r-1} \geq |f_0 f_1|_{r-1}$, so $|f|_r \geq |f_1|_{r-1}$. However, the result of that theorem says $|f_1| \geq \binom{n-d}{\leq r-d}$, meaning it is in fact equality. Since $f_1$ is of lower degree, we can use our theorem by induction. Because $d-1 < r-1 < n-2$, we have 2 cases:

1. $f_1(\vec{X}) = \prod_{i \in S'} X_i$ for some $S' \subset [n-1]$ with $|S'| = d-1$
2. $f_1(\vec{X}) = \left(r - 1 + \sum_{i=1}^{n-1} X_i\right) \prod_{i \in S'} X_i$ for some $S' \subset [n-1]$ and $|S'| = d-2$

Since $|f_1|_{r-1} = |f|_r$, we know that $|f_0|_r = |f_0|_{r-1} = |f_0 f_1|_{r-1}$. The first equality means that $f_0(\vec{x}) = 0$ when $\|\vec{x}\|_H = r$, while the second says $f_0(\vec{x}) = 0$ when $\|\vec{x}\|_H \leq r-1$ and $f_1(\vec{x}) = 0$.

Since $\deg(f_0) = d \leq r-1$ and $f_1$ is the product of $d-1$ linear factors of the forms used in Lemma 3.13, we can apply the lemma to each factor with $t = r-1$. This means that each factor ml-divides $f_0$, so by Lemma 3.8, $f_1$ also ml-divides $f_0$. Since $f_1$ is degree $d-1$, this means that $f_0 = h \cdot f_1$ where $\deg(h) \leq 1$.

Let $h(\vec{X}) = a + \sum_{i=1}^{n-1} a_i X_i$ for constants $a$ and $a_i$. Now the first equality from above says that if $R \subset [n-1]$ is a subset of size $r$, we have that:

$$f_0(\vec{1}_R) = f_1(\vec{1}_R) h(\vec{1}_R) = f_1(\vec{1}_R)\left(a + \sum_{i \in R} a_i\right) = 0$$

Regardless of which case we are in, if $S' \subset R$, then $f_1(\vec{1}_R) = 1$. This means that $\left(a + \sum_{i \in R} a_i\right) = 0$ as long as $S' \subset R$.

Now choose any $i \neq j \in [n-1] \setminus S'$. Because $r < n-1$, we can pick an $R$ with $|R| = r$ such that $i \in R$ and $j \notin R$. Then we can set $R' = R \cup \{j\} \setminus \{i\}$. Now we have that:

$$\left(a + \sum_{i \in R} a_i\right) + \left(a + \sum_{i \in R'} a_i\right) = a_i + a_j = 0$$

Since $i$ and $j$ were chosen arbitrarily this means that for some constant $c, a_i = c$ for every $i \in [n-1] \setminus S'$. This means that:

$$h(\vec{X}) = a + \sum_{i \in S'} a_i X_i + c \sum_{i \in [n-1] \setminus S'} X_i$$

Since $h$ is being multiplied by $f_1$, $h$ is multiplied by $X_i$ for each $i \in S'$, and thus every choice of the $a_i$ induces the same function up when multilinearized modulo $X_i^2 - X_i$. We choose $a_i = c$ for each $i \in S'$, meaning that $h(\vec{1}_R) = a + cr = 0$ for $S' \subset R$. This means $a = cr$ and thus:

$$h(\vec{X}) = c \left( r + \sum_{i=1}^{n-1} X_i \right)$$

If $c = 0$, then $f_0 = 0$, and thus $f = X_n f_1$, which is in one of the two acceptable forms, regardless of which form $f_1$ takes. However, if $c = 1$, then $f_0(\vec{X}) = \left(r + \sum_{i=1}^{n-1} X_i\right) f_1(\vec{X})$. If $\left(r - 1 + \sum_{i=1}^{n-1} X_i\right)$ divides $f_1$, then both $\left(r + \sum_{i=1}^{n-1} X_i\right)$ and its conjugate divide $f_0$, and thus $f_0$ is identically 0 on $\mathbb{F}_2^n$ which reduces to the prior case. Otherwise, $f_1(\vec{X}) = \prod_{i \in S'} X_i$ for some $S' \subset [n-1]$ with $|S'| = d - 1$, so $f(\vec{X}) = \prod_{i \in S} X_i \left(r + \sum_{i=1}^{n} X_i\right)$ for $S = S'$. □

At the threshold $r \leq d$, the paradigm changes again and the minimizing functions are easier to determine:

**Lemma 3.14.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_d = \binom{n-d}{\leq d-d} = 1$, then $f(\vec{X}) = \sum_{T \supseteq S : |T| \leq d} \prod_{i \in T} X_i$ for some $S \subset [n]$ with $|S| \leq d$.*

When $r = d$, there is a unique point with weight less than $d$ that entirely determines the structure of the polynomial. The proof follows from that observation:

*Proof.* Since $|f|_d = 1$, there is a unique $\vec{x}$ with $\|\vec{x}\|_H \leq d$ for which $f(\vec{x}) = 1$. Let $S \subseteq [n]$ be the set for which $\vec{x} = \vec{1}_S$.

Assume that $f$ has the monomial $\prod_{i \in T} x_i$ for some $T \not\supseteq S$. As in the proof of 3.7, take a minimal $T$, and so $f(\vec{1}_T) = 1$. Since $f$ is of maximum degree $d, \|\vec{1}_T\|_H \leq d$, which is a contradiction.

Now we can show that $\prod_{i \in T} X_i$ is part of $f$ for every $T \supseteq S$ by induction on the size of $T$ up to $|T| = d$. The base case is $T = S$. Since no monomials of with fewer than $|S|$ factors can be part of $f$, the only reason $f(\vec{1}_S)$ can be 1 is that $\prod_{i \in S} x_i$ is in $f$.

Now we assume that $\prod_{i \in T'} x_i$ is part of $f$ for every $S \subseteq T' \subsetneq T$. If $\prod_{i \in T} x_i$ is not part of $f$, then $f(\vec{1}_T) = \sum_{S \subseteq T' \subsetneq T} 1 = 2^{|T|-|S|} - 1 = 1$. Since $|T| \leq d$, this is a contradiction. □

It is possible for a polynomial of degree $d$ to vanish on all points with Hamming weight less than $d$, so when $r < d$, saying $|f|_r = 0$ is just imposing a set of linear constraints on $f$. The resulting minimizing polynomials thus form a subspace which we can easily characterize:

**Lemma 3.15.** *If $f \in M_n$ is a nonzero polynomial of total degree at most $d$ where $|f|_r = \binom{n-d}{\leq r-d} = 0$, then $f(\vec{X})$ is a nonzero linear combination of the monomials $\prod_{i \in S} X_i$ for $r < |S| \leq d$.*

*Proof.* If $r < |S| \leq d$, note that $|\prod_{i \in S} X_i|_r = 0$. This means that if $f$ is in the subspace generated by these monomials, then $|f|_r = 0$.

Now assume that $f$ contains a monomial $\prod_{i \in T} X_i$ for some $|T| < r$. We may again take $T$ to be minimal, and then $f(\vec{1}_T) = 1$, which is a contradiction. This means that the subspace is the set of polynomials of degree at most $d$ that vanish on points of hamming weight at most $r$. □

## 3.4 Division Lemmas

As in [9], factoring polynomials is itself an important problem with applications outside its usage in this paper. We used Lemma 3.7 and Lemma 3.13 to find the linear factors needed for Theorem 3.5 and Theorem 3.10 respectively.

While we assume less about the polynomial to be factored using Lemma 3.7, the lemma only works for specific linear functions: The linear functions with minimal support on vectors of Hamming weight less than $t$. Any other linear function will not induce an automorphism when doing a change of coordinates as in the proof of Lemma 3.13. In fact, we can actually produce a counterexample for each of these other linear functions:

**Lemma 3.16.** *If $l(\vec{X})$ is a linear function that is not $t+\sum_{i=1}^{n} X_i$ or $X_i$ for any $1 \leq i \leq n$, then for any $t < n-1$ there is a function $f$ with $\deg f = t$ and $f(\vec{x}) = 0 \,\forall\, \|\vec{x}\|_H \leq t$ with $l(\vec{x}) = 0$, but $l$ does not ml-divide $f$.*

*Proof.* If $l$ satisfies the conditions above, it is not of the structures described in Lemma 3.12, so it must be that $|l|_t > \binom{n-1}{\leq d-1}$.

Consider all functions $f \in M_n$ such that $\deg(f) \leq t$ and $f(\vec{x}) = 0 \,\forall\, \|\vec{x}\|_H \leq t$ with $l(\vec{x}) = 0$. All these constraints are linear, meaning that this forms a vector space $V \subset M_n$. The dimension of $V$ is then easily computed to be $|l|_t$, as we can think of it as a subspace of the space of polynomials of degree at most $t$, with codimension equal to the number of times $l(\vec{x}) = 0$ in the Hamming ball of radius $t$.

Now consider the vector space $W$ of polynomials $f \in M_n$ such that $l$ ml-divides $f$. Trivially $W \subset V$, but since $f = l \cdot g$ for some $g \in M_{n-1}$ with degree $d-1$, the dimension of $W$ is $\binom{n-1}{\leq d-1} < \dim(V)$, and thus $W \subsetneq V$. Therefore $V \setminus W \neq \emptyset$ and thus there is a polynomial $f \in V \setminus W$ that satisfies all the constraints of the theorem. □

We also provide the following construction to generate an explicit $f$ for every $l$ that satisfies the conditions:

*Constructive Proof.* Since $l$ is a non-trivial linear function, we can write $l(\vec{X}) = a + \sum_{i \in S} X_i$ for some $\emptyset \neq S \subseteq [n]$. Fix some $j \in S$. Now choose a maximal set $T_1 \subseteq S \setminus \{j\}$ such that $|S| \cong a+1 \mod 2$ and $|T_1| \leq t$. This is always possible unless $|S \setminus \{j\}| = 0$ and $a = 0$, which only occurs when $l(\vec{X}) = X_i$ for some $i \in [n]$.

Now Let $T_2$ be any set of $t - |S'|$ elements chosen from $[n] \setminus S$. Since $t < n-1$, the only way this is impossible is if $S = [n]$ and $t \cong a \mod 2$ which is when $l(\vec{X}) = t + \sum_{i=1}^{n} X_i$. Otherwise, let $T = T_1 \cup T_2$, so $|T| = t$.

Now let $f(\vec{X}) = \prod_{i \in T} X_i$. Note that $f$ is zero on all $|\vec{x}|_H \leq t$ except $\vec{1}_T$, so since $l(\vec{1}_T) = a + |S'| = 1$, all that remains is to show that $l$ does not divide $f$.

Because $l$ depends on $X_j$, $l(\vec{1}_{T \cup \{j\}}) = 0$. However $f(\vec{1}_{T \cup \{j\}}) = 1$, meaning that $f \neq l \cdot g$ for any $g \in M_n$. □

The same techniques can be used to prove analogs of Lemma 3.13 and Lemma 3.16

for the case where $t = n-1$, except the linear functions that can be used are those $l$ for which $l(\vec{1}) = 1$.

In Lemma 3.16, we note that the technique used can only construct counterexamples of degree exactly $t$. This is no coincidence, as the following extension of Lemma 3.13 shows:

**Lemma 3.17.** *If* $f, l \in M_n$ *are polynomials such that* $\deg(l) = 1, \deg(f) \leq t-1$*, and* $f(\vec{x}) = 0 \, \forall \, \|\vec{x}\|_H \leq t$ *with* $l(\vec{x}) = 0$*, then* $l$ *ml-divides* $f$*.*

We can prove this in two similar ways, using either of our previous division lemmas. The first proof of this mirrors that of Lemma 3.13, using a similar change of coordinates to prove divisibility:

*Proof using Lemma 3.13.* Without loss of generality, we can assume that $l$ depends on $X_n$, and thus let $\Psi$ be a change of basis given by the involution:

$$\Psi(X_1, \ldots, X_{n-1}, X_n) = (X_1, \ldots, X_{n-1}, l(\vec{X}))$$

While this doesn't induce an automorphism in the same way as Lemma 3.13, we note that only one coordinate is changing, and thus $\|\Psi(\vec{x})\|_H - \|\vec{x}\|_H \leq 1$. This means that $\{\vec{x} : \|\Psi(\vec{x})\|_H \leq t-1\} \subset \{\vec{x} : \|\vec{x}\|_H \leq t\}$, so since $f(\vec{x}) = 0 \, \forall \, \|\vec{x}\|_H \leq t$ with $l(\vec{x}) = 0$, we have that $f(\Psi(\vec{x})) = 0 \, \forall \, \|\Psi(\vec{x})\|_H \leq t-1$ with $l(\Psi(\vec{x})) = 0$.

Now since $\deg(f) \leq t-1$ and $l(\Psi(\vec{X})) = X_n$, by Lemma 3.13, $l$ ml-divides $f$. □

This can also be proved through a clever use of Theorem 3.9:

*Proof using Lemma 3.9.* Consider the function $g = f(1-l)$. Since $f(\vec{x}) = 0$ when $l(\vec{x}) = 0$ and $\|\vec{x}\|_H \leq t$, we can see that $g$ vanishes on every $\vec{x}$ with $\|\vec{x}\|_H \leq t$, and thus $|g|_t = 0$

Since $\deg g \leq \deg f + 1 \leq t$, by Theorem 3.9 we know that $g = 0$ (after multilinearization). This means that $f(\vec{x}) = 0$ when $l(\vec{x}) = 0$ for every $\vec{x} \in \mathbb{F}_2^n$, so by Lemma 3.7 $l$ ml-divides $f$. □

This means that the only possible counterexamples to Lemma 3.13 are of degree exactly $t$. While Lemma 3.17 is not needed in the proof of Theorem 3.12, it may have other applications in future work.

## 3.5 Further Directions

There are a few avenues to explore to extend these results. Kasami and Tokura [11], [12] analyze the structure of polynomials with nearly minimal support. When only considering the support on vectors with weights less than $r$, similar results may be possible.

When $r \leq d$, by taking a proper summation of the functions described in Lemma 3.14 (and Lemma 3.15 if $r < d$), we can create a function with an arbitrary support on $\{\vec{x} : \|\vec{x}\| \leq r\}$, trivially answering the above questions in this regime.

When $r > d$, the question of identifying and categorizing nearly minimal supports is far less trivial. The first challenge with obtaining a structural result is determining which range of $|f|_r$ such a theorem can address, since it is unclear what even the second smallest support is. By considering $f = (X_1 + X_2)X_3 \ldots X_{d+1}$, we can see that $|f| = 2\binom{n-d-1}{r-d-1}$ is possible, though experimentally this is not always the second smallest support size.

The methods used in Lemma 3.12 unfortunately don't generalize well in this regime, since now it is possible that $|f_0|_r + |f_0|_{r-1} - 2|f_0 f_1|_{r-1} > 0$. Although we can rewrite this number as $|\{\vec{x} : \|\vec{x}\|_H = r, f_0(\vec{x}) = 1\}| + 2|f_0(1 - f_1)|_{r-1}$, this still means that:

$$|f|_r - \binom{n-d}{r-d} = |\{\vec{x} : \|\vec{x}\|_H = r, f_0(\vec{x}) = 1\}| + 2|f_0(1 - f_1)|_{r-1} + \left(|f_1|_{r-1} - \binom{n-d}{r-d}\right)$$

This three term summation means that there are potentially many ways to distribute the excess support size, raising the complexity of any categorization proof substantially.

To avoid this problem in the $r = n$ case, Kasami and Tokura [11] decompose $f$ along two variables, writing $f = f_0 + X_{n-1}f_1 + X_n f_2 + X_{n-1}X_n f_3$ for $f_0, f_1, f_2, f_3 \in \mathbb{F}_2[X_1, \ldots, X_{n-2}]$. In the $r < d$ case, we now must consider the changes in $r$ that come from writing $|f|_r = |f_0|_r + |f_0 + f_1|_{r-1} + |f_0 + f_2|_{r-1} + |f_0 + f_1 + f_2 + f_3|_{r-2}$.

Kasami and Tokura [11],[12] also use the fact that any affine change of coordinates in $\mathbb{F}_2^n$ is an automorphism of the cube to find linear factors of $f$, but when $r < n$ this is not possible unless the affine change of coordinates is one of the special ones outlined in Lemma 3.13. However, it may be possible to use Lemma 3.17 to find linear divisors when the degree is low enough, implying that the regime to examine is $r > d+1$ instead of $r > d$.

So far all these questions have been focused on determining the structure and support of a polynomial $f \in \mathbb{F}_2[X_1, \ldots, X_n]$. If the polynomial is over $\mathbb{F}_q$ instead of $\mathbb{F}_2$, we can ask similar questions. The same kind of improvements in Theorem 3.2 can be applied to $\mathbb{F}_q$ to get a similar Schwartz-Zippel type result:

**Theorem 3.18.** *If $f \in \mathbb{F}_q[X_1, \ldots, X_n]$ is a nonzero polynomial of total degree at most $d$, then by writing $d = a(q-1) + b$ for $0 \leq b < q-1$,*

$$P_{\vec{x} \in \mathbb{F}_q^n}(f(\vec{x}) \neq 0) \geq \left(1 - \frac{b}{q}\right) q^{-a}$$

This generalization can be seen to be tight in a similar way to Theorem 3.2: Let $f(\vec{X}) = \prod_{i=1}^{a}(X_i^{q-1} - 1) \prod_{j=1}^{b}(X_b - j)$. This polynomial and all affine transformations of it are tight examples for Theorem 3.18, but it is unknown whether these are the only polynomials that do so.

In projective space, this is known ([28],[26],[15],[6]), but it is not clear how to adapt those algebraic methods to address the affine case. The methods used here don't generalize well into $\mathbb{F}_q$, as even if we write $f = \sum_{i=0}^{q-1} X_n^i f_i$, we can no longer use inclusion-exclusion to analyse $|f_0 + f_1|_r$ in terms of $|f_0|_r$ and $|f_1|_r$. Further algebraic tools are needed to continue this line of inquiry.

# Chapter 4

# Weight Bounds in Dual Reed-Muller Codes

## 4.1 Background and Definitions

In coding theory, the most basic uses of polynomials are the Reed-Muller codes. For a given degree $d$, there are $\binom{n}{\leq d}$ unique monomials on $n$ variables of degree no more than $d$, and thus we can represent $\binom{n}{\leq d}$ bits as a polynomial $f \in \mathbb{F}_2[X_1, \ldots, X_n]$ where $\deg f \leq d$.

The codeword corresponding to a given $f$ is the tuple $(f(\vec{x}) : \vec{x} \in \mathbb{F}_2^n)$. This is a string of length $2^n$, thus encoding at a rate of $\frac{\binom{n}{\leq d}}{2^n}$.

The minimum distance between two codewords $f$ and $g$ is exactly the number of nonzeros in the codeword of $f + g$. Unless $f = g$, this is a non-zero polynomial of degree less than $d$ and so by Theorem 3.2, $f + g$ has at least $2^{n-d}$ non-zeros. This means the minimum distance of the code is $2^{n-d}$.

This gives us a simply described family of linear codes for which we can tailor their efficiency, allowing for many applications, such as in [22],[23],[1],[2],[3], and [13]. Although presented as purely polynomial results, the papers from which the previous chapter drew inspiration [11],[12] stated their results in the context of Reed-Muller codes. Many results on polynomials can be used to make statements about Reed-Muller codes, making them very useful codes. Results on polynomials over other fields find uses in Reed-Solomon or BCH codes.

For any linear code $C \subset \mathbb{F}_2^n$, the dual code is $\{\vec{x} \in \mathbb{F}_2^n : \vec{x} \cdot \vec{c} = 0 \,\forall\, c \in C\}$. Equivalently, dual codewords for Reed-Muller codes are sets of points $S \subset \mathbb{F}_2^n$ such that every polynomial of degree (at most) $d$ sums to 0 on $S$.

**Definition 4.1.** *The set $S \subset \mathbb{F}_2^n$ is a dual codeword for polynomials of degree up to $d$*

*if* $\sum_{\vec{x}\in S} f(\vec{x}) = 0 \,\forall\, f \in \mathbb{F}[X_1, \ldots, X_n]$ *with* $\deg f \leq d$.

Each set $S$ corresponds to a $c$ which is a $\mathbb{F}_2$ vector of length $2^n$, where each entry is associated with a $\vec{x} \in \mathbb{F}_2^n$, and is 1 if and only if $\vec{x} \in S$.

## 4.2 Minimum Weights of Dual Codewords

There are a few properties of dual codewords that are easily verified:

1. Trivially $S = \emptyset$ is a dual codeword for any $n$ and $d$.

2. If $d \geq n$, there are no other dual codewords.

3. If $d < n$, then $S = \mathbb{F}_2^n$ is a dual codeword.

To show the second property, choose $\vec{y} \in \mathbb{F}_2^n$ arbitrarily. Note that the polynomial $f(\vec{X}) = \prod_{i=1}^n (X_i + y_i + 1)$ is 0 on every point of $\mathbb{F}_2^n$ except $\vec{y}$. This means that $\sum_{\vec{x}\in S} f(\vec{x}) = 1$ if $\vec{y} \in S$. Since $\vec{y}$ was arbitrary, no nonempty $S$ can be dual codewords.

The last property can be shown by examining each monomial separately and noting that each is nonzero on a subspace of codimension equal to their degree. Since $d < n$, they are each nonzero on a subspace of positive dimension, and thus on an even number of points.

When $d < n - 1$, we can construct more dual codewords by affine transformation: Let $\Psi : \mathbb{F}_2^k \to \mathbb{F}_2^n$ be any affine embedding for $k > d$. If $f$ is any polynomial of degree $d$ on $\mathbb{F}_2^n$, it is also degree $d$ on $\Psi(\mathbb{F}_2^k)$. Since $k > d$, $\mathbb{F}_2^k$ is a dual codeword for any $f \circ \Psi$, and thus $\Psi(\mathbb{F}_2^k)$ is a dual codeword for any $f$.

In fact, if we know a dual codeword $S \subset \mathbb{F}_2^k$, then $\Psi(S) \subset \mathbb{F}_2^n$ is a dual codeword by the same logic. This means that to categorize the structure of dual codewords, we should classify them by their affine dimension.

Now we can pose the question: What is the smallest a dual codeword $S$ for degree $d$ polynomials can be if $S$ has affine dimension $m$? We already constructed a dual codeword with $|S| = 2^m$ (while $m > d$), but, for large enough $m$, there are subexponential size codewords.

In fact, we can construct dual codewords where the size is linear in $m$ as follows: Let $S$ be the union of $k$ affinely independent (affine) subspaces of dimension $d+1$. Each subspace has $d+2$ degrees of freedom, so the $k2^{d+1}$ points are contained in an $m = k(d+2)-1$ dimensional affine subspace. This means that when $k = \frac{m+1}{d+2}$ is integral, $|S| = \frac{2^{d+1}}{d+2}(m+1)$ is possible.

We conjecture that this construction gives an optimal bound:

**Conjecture 4.2.** *If $S$ is a dual codeword for polynomials of degree at most $d \geq 2$ of affine dimension $m$, then $|S| \geq \frac{2^{d+1}}{d+2}(m+1)$.*

We exclude the case when $d = 1$, as it is easy to prove that $|S| \geq m+2$, and that this bound is tight for $m$ even. The lower bound follows from the fact that we need at least $m+1$ points to define an affine subspace of dimension $m$. If $|S| = m+1$, since as any $m$ points define an affine subspace of dimension $m-1$, we can construct a linear function that vanishes on those $m$, but not on the final point of $S$, meaning $S$ is not a dual codeword.

The construction is simple: $S = \{0, \vec{e}_1, \ldots, \vec{e}_m, \vec{e}_1 + \cdots + \vec{e}_m\}$. It is easily verified that any non-trivial linear function sums to 0 on these points, and we require even $m$ to be a dual codeword for the degree 0 polynomial that is identically 1.

In the case that $d = 2$, we can prove that this bound is indeed tight:

**Theorem 4.3.** *If $S$ is a dual codeword for polynomials of degree at most 2 of affine dimension $n$, then $|S| \geq 2(m+1)$.*

To prove this, we use an alternate formulation of the set $S$:

**Definition 4.4.** *The set $S \subset \mathbb{F}_2^n$ is a dual codeword for polynomials of degree up to $d$ if $\sum_{\vec{x} \in S} (\vec{x}, 1)^{\otimes d} = \vec{0}^{\otimes d}$.*

This definition is equivalent to Definition 4.1, as the $i_1, i_2, \ldots, i_d$ entry of $(\vec{x}, 1)^{\otimes d}$ is $\prod_{j=1}^{d} x_{i_j}$ which is the the evaluation of the monomial $\prod_{j=1}^{d} X_{i_j}$ on $\vec{x}$. Since evaluation is over $\mathbb{F}_2$, repeated indices can be removed, allowing us to generate every multilinear monomial of degree up to $d$ (the additional entry allows us to do the same for the constant monomial).

If Definition 4.1 holds, then since each of these monomials is itself a polynomial of degree $d$, we see that every entry of the tensor must be 0 implying Definition 4.4.

On the other hand, if Definition 4.4 holds, then by examining the appropriate entry of the tensor, we can see that the sum of the evaluations on $S$ of every monomial of degree up to $d$ add to 0. Therefore, any polynomial of maximum degree $d$ also has the sum of its evaluations on $S$ add to 0, implying Definition 4.1.

Since 2-tensors of vectors are equivalent to matrices, we can use known matrix properties to prove Theorem 4.3.

*Proof.* Assume for the sake of contradiction that $|S| \leq 2m+1$. If $S$ is contained in an affine subspace of dimension $m$, we can use a set of affine transformations to make $\vec{0} \in S$, as well as $\vec{e}_i \, \forall \, 1 \leq i \leq m$. Now $S \subset \mathbb{F}_2^m$, so we may drop the other coordinates without changing the problem.

Now consider the outer product matrix that corresponds to the 2-tensor $(\vec{e}_i, 1)^{\otimes 2} = (\vec{e}_i, 1)(\vec{e}_i, 1)^\top$. The only nonzero entries are the coordinates in $\{i, m+1\} \times \{i, m+1\}$. Similarly $(\vec{0}, 1)^{\otimes 2}$ is zero except at the $(m+1, m+1)$ position. Adding these $n+1$ matrices together gives a matrix:

$$(\vec{0}, 1)^{\otimes 2} + \sum_{i=1}^{m} (\vec{e}_i, 1)^{\otimes 2} = \left[\begin{array}{c|c} I_m & \vec{1}^\top \\ \hline \vec{1} & m+1 \end{array}\right]$$

By row reduction, we can see that this matrix is full rank. However, since $\sum_{\vec{x} \in S} (\vec{x}, 1)^{\otimes 2} = \vec{0}^{\otimes 2}$, we can see that $\sum_{\vec{x} \in S \setminus \{\vec{0}, \vec{e}_i \, \forall \, i\}} (\vec{x}, 1)^{\otimes 2} = (\vec{0}, 1)^{\otimes 2} + \sum_{i=1}^{m} (\vec{e}_i, 1)^{\otimes 2}$ and thus is the same matrix.

Since $(\vec{x}, 1)^{\otimes 2} = (\vec{x}, 1)(\vec{x}, 1)^\top$ is a rank 1 matrix for any $\vec{x}$, the rank of $\sum_{\vec{x} \in S \setminus \{\vec{0}, \vec{e}_i \, \forall \, i\}} (\vec{x}, 1)^{\otimes 2}$ is no more than $|S| - m - 1 \leq m$. This contradicts that this matrix is full rank. □

Since $(\vec{x}, 1)^{\otimes 2}$ is a subtensor of $(\vec{x}, 1)^{\otimes d}$ for any $d > 2$, the above result proves that $|S|$ must be at least twice $m$ (in fact at least $2m+2$) for $d \geq 3$, but cannot achieve the bounds conjectured. Even using a higher dimensional analog of rank cannot achieve better bounds, as any sum of $m+1$ rank one tensors can be constructed as the sum of those same $m+1$ tensors. Forbidding repeat tensors is difficult with a rank-based approach, and would require new developments.

The following weakening of Conjecture 4.2 would itself be an interesting result:

**Conjecture 4.5.** *If $S$ is a dual codeword for polynomials of degree at most $d \geq 3$ of affine dimension $m$, then $\exists\, \epsilon_d > 0$ such that $|S| \geq (2+\epsilon_d)(m+1)$.*

Consider the function $1_S \in \mathbb{F}_2[X_1, \ldots, X_n]$ such that $1_S(\vec{x}) = 1$ if and only if $\vec{x} \in S$. This means that $\sum_{\vec{x} \in S} f(\vec{x}) = \sum_{\vec{x} \in \mathbb{F}_2^n} f(\vec{x}) 1_S(\vec{x})$, so if $S$ is a dual codeword, we have that $\sum_{\vec{x} \in \mathbb{F}_2^n} f \cdot 1_S(\vec{x}) = 0$ for every $\deg f \leq d$. This means that $\deg(1_S) \leq n-d-1$, so by using Theorem 3.9, we can get a small improvement over Theorem 4.3:

**Theorem 4.6.** *If $S$ is a dual codeword for polynomials of degree at most $d$ of affine dimension $m$, then $|S| \geq 2(m+1) + 2^d - d - 2$.*

*Proof.* If $S$ is of affine dimension $m$, then there is a change of variables such that $S$ is contained in the subspace generated by the first $m$ coordinates. By picking the change of coordinates properly, we may assume without loss of generality $\vec{1} \in S$ and $\{\vec{1} - \vec{e}_i, 1 \leq i \leq n\} \subset S \subseteq \mathbb{F}_2^m$. Again we can drop all but these $m$ coordinates without affecting the result.

If $l \cdot 1_S$ is identically 0 for some (non-zero) linear $l$ on $X_1, \ldots, X_m$, then $S$ is contained in the affine subspace $l(\vec{X}) = 0$. Since $S$ is of affine dimension $m$, $l \cdot 1_S$ is nonzero for every nonzero linear $l$. Since the space of linear functions on $X_1, \ldots, X_m$ is of dimension $m+1$, we can choose any $m$ points in $S$ and find a nonzero linear function $l$ that vanishes on those points.

Choose any $n$ points in $S$ with weight no more than $m-2$, and let $l$ be a nonzero linear function that vanishes on those points. Since $\deg(l \cdot 1_S) \leq m-d$ and not identically zero, by Theorem 3.9, we have that $|l \cdot 1_S|_{m-2} \geq \binom{d}{\leq d-2} = 2^d - d - 1$.

Since $1_S$ must be one whenever $l \cdot 1_S$ is, $|1_S|_{m-2} \geq |l \cdot 1_S|_{m-2}$. In fact, since $l$ was chosen to be 0 on $m$ points in $S$ with weight $d-2$, we actually have $|1_S|_{m-2} \geq m + 2^d - d - 1$.

By our assumptions earlier, all $m+1$ points with weight greater than $m-2$ are in $S$, so therefore:

$$|S| = |1_S|_m = m + 1 + |1_S|_{m-2} \geq 2(m+1) + 2^d - d - 2$$

□

When $d = 2$, Theorem 4.6 matches the bound in Theorem 4.3, as well as the lower bound construction above. For $d \geq 3$, Thoerem 4.6 is only a constant improvement over Theorem 4.3, but not the linear improvement needed for Conjectures 4.2 and 4.5.

It is unlikely this bound is tight, as Theorem 3.10 shows that the polynomials that achieve the bound in Theorem 3.9 have many linear factors, but $1_S$ cannot have any. Another structural result may be needed to make further improvements.

We can use Theorem 4.3 (or Theorem 4.6) to count the number of dual codewords for polynomials of degree at most $d$:

**Theorem 4.7.** *For a given $n$, the number of dual codewords for polynomials of degree $d \geq 2$ of size at most $s$ is $O\left(2^{\frac{s}{2}n}\right)$.*

*Proof.* Consider any dual codeword $S$ of size at most $s$. By Theorem 4.3 $s \geq |S| \geq 2(r+1)$ where $r$ is the affine dimension of $S$. An affine subspace of dimension $r$ can be defined by choosing $r+1$ affinely independent points and there are at most $\binom{2^n}{r+1}$ ways to do this. Once we have chosen an affine subspace, we must choose a dual codeword of size $s$ from the $2^r$ points. Naively, this is bounded by $\binom{2^r}{s}$. Summing these estimations gives the bound:

$$\sum_{r=0}^{\frac{s}{2}-1} \binom{2^n}{r+1}\binom{2^r}{s}$$

For constant $s$, this is $O\left(2^{\frac{s}{2}n}\right)$. □

Not only is this much better than the naive bound of $\binom{2^n}{s} = O\left(2^{sn}\right)$, but it is actually not overestimating too much when Theorem 4.3 is tight at $d = 2$.

However, when $d \geq 3$, this is likely not tight for similar reasons as Theorem 4.6. If proven, Conjecture 4.2 would give a better bound, saying that the number of codewords of size $s$ is at most $O\left(2^{\frac{d+2}{2^{d+1}}sn}\right)$.

Since these dual codewords can themselves be used as Reed-Muller codes (as in the proof of Theorem 4.6), Theorem 4.7 shows that for Reed-Muller codes of where

$d \leq n-3$, we have that the number of codewords of size at most $s$ is $O(N^{\frac{s}{2}})$, where $N = 2^n$ is the length of the code.

## 4.3 Further Directions

There are a few different methods that could be used to improve Theorem 4.6 in the hopes of proving either Conjecture 4.5 or even Conjecture 4.2.

As mentioned before, Theorem 3.10 says that if $l \cdot 1_S$ achieves the bound in Theorem 3.9, it should be the product of linear functions (after multilinearization). We also assumed that $1_S$ was supported on all $\vec{x}$ with Hamming weight $n-1$ and above. It seems impossible to satisfy both of these conditions, suggesting that it may be possible to refine this method.

**Conjecture 4.8.** *Let $f \in \mathbb{F}_2[X_1, \ldots X_n]$ be a nonzero degree $d$ polynomial divisible by at most $t$ linearly independent linear functions. Then $|f|_n \geq \frac{2^{n-d}}{n-d+1}(n-t+1)$.*

Note that when $t = d$, the above conjecture is proved by Theorem 3.2. In addition, when $t = 0$, we are left with Conjecture 4.2. However the two conjectures are actually equivalent, as the $t$ linear factors can be used to create a $\Phi : \mathbb{F}_2^n \to \mathbb{F}_2^{n-t}$ by restricting to the subspace given by setting all the linear factors to 0. Restricting $f$ by $\Phi$ gives that $f$ is a dual codeword of affine dimension $m = n - t$.

While Conjecture 4.8 is equivalent to Conjecture 4.2, the different presentation may facilitate an inductive proof. By Theorem 4.3, we have proven a base case when $d = n-3$, but going further has proven difficult.

Another possible methodology for proving Conjecture 4.2 comes from this alternate view of dual codewords: If $\deg(1_S) \leq n-d-1$, then by examining the monomials we can see that $S$ is the $\mathbb{F}_2$ sum of cubes, where each cube is a point of Hamming weight at least $d+1$ as well as all points below that point in the subset lattice (with $n$ atomic basis elements).

By mapping $\Phi(\vec{x}) = \left(\vec{x}, 1 + \sum_{i=1}^{n}\right) \in \mathbb{F}_2^{n+1}$, we can embed $S$ into the odd Hamming weight subspace of $\mathbb{F}_2^{n+1}$. Since we assumed that $S$ is of affine dimension $m$, we can assume that $\vec{e_i} \in \Phi(S)$ for all $i \in [m+1]$.

The cubes from points of Hamming weight $d+1$ become the odd weight vertices of a cube from a point of Hamming weight $d+2$, and now we can view this as an energy minimization problem. Since $|\{\vec{x} \in \Phi(S) : \|\vec{x}\|_H = 1\}|$ is always less than the affine dimension of $S$, we simply want to say that if $S$ is the $\mathbb{F}_2$ sum of cubes from points of Hamming weight at least $d+2$, then $|\{\vec{x} \in S : \|x\|_H \cong 1 \mod 2\}| \geq \frac{2^{d+1}}{d+2}|\{\vec{x} \in \Phi(S) : \|\vec{x}\|_H = 1\}|$.

This is natural in that if we take cubes that are on pairwise disjoint coordinates, we achieve the tight construction in the previous section. In this regime it is also easy to visualize why there are factors of $d+2$ and $n+1$ in the bound of Conjecture 4.2.

However, proving this for all sums of these cubes is difficult, as computing the intersections of these cubes requires using their polynomial structure. While this lattice helps with visualization of the tight examples, use of the polynomial structure is likely needed to actually prove the bounds.

Another potential avenue for exploration is to more carefully decompose the polynomials in the polynomial approach. We chose exactly $n$ points and an $l$ that vanished on those points, but a randomly chosen linear function should vanish on about half of any set of points.

Alternately, if we allow for nonlinear polynomials, we can choose more than $n$ points. But if $p$ is a polynomial of degree (at least) 2, we have no guarantee that $p \cdot 1_S \neq 0$. In fact, this cannot be true, as then we could show that $|S| = \Omega(n^2)$. Ideally, we would want to choose $p$ from a polynomial subspace of dimension $\left(\frac{2^{d+1}}{d+2} - 1\right)(m+1)$ where $p \cdot 1_S \neq 0$ and $\deg(p \cdot 1_S) \leq m-2$, but no natural constructions seem to appear. Finding a similar subspace of dimension $(1+\epsilon_d)(m+1)$ would be sufficient to prove Conjecture 4.5.

Lastly, note that we can adapt the proof of Theorem 4.6 to prove something slightly different:

**Lemma 4.9.** *If $S$ is a dual codeword for polynomials of degree at most $d$ of affine dimension $n$, then $|1_S|_{n-d} \geq n+1$.*

The proof is similar to that of Theorem 4.6:

*Proof.* As before, we assume without loss of generality $\vec{1} \in S$ and $\{\vec{1} - \vec{e_i}, 1 \leq i \leq n\} \subset S \subseteq \mathcal{F}_2^n$.

Now choose $n$ points in $S$ of Hamming weight at most $n - d$. Let $l$ be a nonzero linear function that vanishes on those points. As before, $l \cdot 1_S$ is a nonzero polynomial of degree at most $n - d$, and thus $|l \cdot 1_S|_{n-d} \geq 1$ by Theorem 3.12. This means that $1_S$ is nonzero on at least one point of Hamming weight at most $n - d$ other than the $n$ chosen in the construction of $l$. □

This lemma says that $|1_S|_{n-d}$ is much larger than the $d + 2$ given by applying Theorem 3.12 to $1_S$. We also know there are $n + 1$ points of Hamming weight $n - 1$ or $n$, proving Theorem 4.3.

There must also be points of Hamming weights between $n-d$ and $n-1$, but Theorem 4.6 only gives a constant number. But knowing that both $|1_S|_{n-d} \geq n + 1$ and that $S$ contains all points with hamming weight at least $n + 1$, it seems impossible that $S$ is that sparse in the gap. Showing that there are $O(n)$ elements of $S$ in this intermediate range would prove Conjecture 4.5, with a sufficiently large constant proving Conjecture 4.2.

# Bibliography

[1] Emmanuel Abbe, Amir Shpilka, and Avi Wigderson. "Reed–Muller codes for random erasures and errors". In: *IEEE Transactions on Information Theory* 61.10 (2015), pp. 5229–5252.

[2] Emmanuel Abbe, Amir Shpilka, and Min Ye. "Reed-Muller Codes: Theory and Algorithms". In: *ArXiv* abs/2002.03317 (2020).

[3] Abhishek Bhowmick and Shachar Lovett. "The List Decoding Radius for Reed Muller codes over Small Fields". In: *IEEE Transactions on Information Theory* (2018).

[4] Joel Brenner and Larry Cummings. "The Hadamard maximum determinant problem". In: *The American Mathematical Monthly* 79.6 (1972), pp. 626–630.

[5] Henning Bruhn and Dieter Rautenbach. "Maximal determinants of combinatorial matrices". In: *Linear Algebra and its Applications* 553 (Sept. 2018), pp. 37–57. ISSN: 0024-3795. DOI: 10.1016/j.laa.2018.04.030. URL: http://dx.doi.org/10.1016/j.laa.2018.04.030.

[6] Alain Couvreur. "An upper bound on the number of rational points of arbitrary projective varieties over finite fields". In: *Proceedings of the American Mathematical Society* 144.9 (Feb. 2016), pp. 3671–3685. ISSN: 1088-6826. DOI: 10.1090/proc/13015. URL: http://dx.doi.org/10.1090/proc/13015.

[7] Richard A DeMillo and Richard J Lipton. *A Probabilistic Remark on Algebraic Program Testing.* Tech. rep. GEORGIA INST OF TECH ATLANTA SCHOOL OF INFORMATION and COMPUTER SCIENCE, 1977.

[8] Shaun Fallat and P Van Den Driessche. “Maximum determinant of (0, 1) matrices with certain constant row and column sums”. In: *Linear and Multilinear Algebra* 42.4 (1997), pp. 303–318.

[9] Michael A Forbes and Amir Shpilka. “Complexity Theory Column 88: Challenges in Polynomial Factorization1”. In: *ACM SIGACT News* 46.4 (2015), pp. 32–49.

[10] J. Hadamard. “Resolution d’une question relative aux determinants”. In: *Bull. Des Sciences Math.* 2 (1893), pp. 240–246.

[11] T. Kasami and N. Tokura. “On the weight structure of Reed-Muller codes”. In: *IEEE Transactions on Information Theory* 16.6 (Nov. 1970), pp. 752–759. ISSN: 1557-9654. DOI: 10.1109/TIT.1970.1054545.

[12] Tadao Kasami, Nobuki Tokura, and Saburo Azumi. “On the Weight Enumeration of Weights Less than 2.5d of Reed-Muller Codes”. In: *Information and Control* 30.4 (1976), pp. 380–395. DOI: 10.1016/S0019-9958(76)90355-7. URL: https://doi.org/10.1016/S0019-9958(76)90355-7.

[13] Tali Kaufman, Shachar Lovett, and Ely Porat. “Weight Distribution and List-Decoding Size of Reed-Muller Codes”. In: *Electronic Colloquium on Computational Complexity* (Sept. 2011).

[14] Swastik Kopparty and Srikanth Srinivasan. “Certifying Polynomials for $\mathrm{AC}^0[\oplus]$ Circuits, with Applications to Lower Bounds and Circuit Compression”. In: *Theory of Computing* 14.12 (2018), pp. 1–24. DOI: 10.4086/toc.2018.v014a012. URL: http://www.theoryofcomputing.org/articles/v014a012.

[15] Gilles Lachaud and Robert Rolland. “On the number of points of algebraic sets over finite fields”. In: *Journal of Pure and Applied Algebra* 219.11 (Nov. 2015), pp. 5117–5136. ISSN: 0022-4049. DOI: 10.1016/j.jpaa.2015.05.008. URL: http://dx.doi.org/10.1016/j.jpaa.2015.05.008.

[16] Chi-Kwong Li, Julia Shih-Jung Lin, and Leiba Rodman. “Determinants of Certain Classes of Zero-One Matrices with Equal Line Sums”. In: *Rocky Mountain Journal of Mathematics* 29.4 (Dec. 1999), pp. 1363–1385. ISSN: 0035-7596. DOI: 10.1216/rmjm/1181070411. URL: http://dx.doi.org/10.1216/rmjm/1181070411.

[17] Ingram Olkin. “A determinantal inequality for correlation matrices”. In: *Statistics & Probability Letters* 88 (May 2014), pp. 88–90. ISSN: 0167-7152. DOI: 10.1016/j.spl.2014.01.012. URL: http://dx.doi.org/10.1016/j.spl.2014.01.012.

[18] William P Orrick. “The maximal {-1, 1}-determinant of order 15”. In: *Metrika* 62.2-3 (2005), pp. 195–219.

[19] William P Orrick and Bruce Solomon. “Large-determinant sign matrices of order 4k+ 1”. In: *arXiv preprint math/0311292* (2003).

[20] Alexander A Razborov. “Lower bounds on the size of bounded depth circuits over a complete basis with logical addition”. In: *Mathematical Notes of the Academy of Sciences of the USSR* 41.4 (1987), pp. 333–338.

[21] H. J. Ryser. “Maximal Determinants In Combinatorial Investigations”. In: *Canadian Journal of Mathematics* 8 (1956), pp. 245–249. ISSN: 1496-4279. DOI: 10.4153/cjm-1956-028-4. URL: http://dx.doi.org/10.4153/cjm-1956-028-4.

[22] Ramprasad Saptharishi, Amir Shpilka, and Ben lee Volk. “Efficiently Decoding Reed-Muller Codes From Random Errors”. In: *IEEE Trans. Information Theory* 63 (2017), pp. 1954–1960.

[23] Ori Sberlo and Amir Shpilka. “On the Performance of Reed-Muller Codes with respect to Random Errors and Erasures”. In: *CoRR* abs/1811.12447 (2018). arXiv: 1811.12447. URL: http://arxiv.org/abs/1811.12447.

[24] Daniel Scheinerman. “Several problems in linear algebraic and additive combinatorics”. PhD thesis. Rutgers University-School of Graduate Studies, 2019.

[25] J. T. Schwartz. “Fast Probabilistic Algorithms for Verification of Polynomial Identities”. In: *J. ACM* 27.4 (Oct. 1980), pp. 701–717. ISSN: 0004-5411. DOI: 10.1145/322217.322225. URL: https://doi.org/10.1145/322217.322225.

[26] Jean-Pierre Serre. “Lettre à M. Tsfasman”. In: 2000.

[27] Roman Smolensky. “Algebraic methods in the theory of lower bounds for Boolean circuit complexity”. In: *Proceedings of the nineteenth annual ACM symposium on Theory of computing.* 1987, pp. 77–82.

[28] Anders Bjært Sørensen. “On the number of rational points on codimension-1 algebraic sets in Pn(Fq)”. In: *Discret. Math.* 135 (1994), pp. 321–334.

[29] James Joseph Sylvester. “LX. Thoughts on inverse orthogonal matrices, simultaneous signsuccessions, and tessellated pavements in two or more colours, with applications to Newton’s rule, ornamental tile-work, and the theory of numbers”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 34.232 (1867), pp. 461–475.

[30] Richard Zippel. “Probabilistic algorithms for sparse polynomials”. In: *International symposium on symbolic and algebraic manipulation.* Springer. 1979, pp. 216–226.