Thesis for the degree
Doctor of Philosophy

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
דוקטור לפילוסופיה

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Shachar Lovett

מאת
שחר לובט

פולינומים מעל שדות סופיים ושימושים במדעי המחשב

Polynomials over finite fields with applications to Computer Science

Advisor:
Prof. Omer Reingold
Prof. Ran Raz

מנחה:
פרופ' עומר ריינגולד
פרופ' רן רז

July 2010

תמוז התש"ע

Contents

Abstract

This work studies multivariate polynomials over finite fields and their applications in computer science. The study of polynomials in computer science is not new. Polynomials have found applications in the areas of algorithm design, cryptography, circuit lower bounds, computational learning and coding theory, to name a few key examples. In this work we continue the research of polynomials in computer science, and focus on four main areas: we study polynomials as a model of computation, where we explore several fundamental problems, such as the relation between exact and approximate computation, the resources required to generate hard functions, and the importance of the base field; we study distributions which are pseudorandom for polynomials, and provide as applications lower bounds for bounded depth circuits; we study coding theoretic questions related to codes arising from polynomials, such as Reed-Muller and BCH codes; and we study property testing for polynomials, which is motivated by recent advances in additive combinatorics. In all these areas, we establish new results which emerge from an improved understanding of the fundamental properties of polynomials over finite fields.

תקציר

אנו חוקרים בעבודה זו פולינומים במספר משתנים המוגדרים מעל שדות סופיים, וכן שימושים של פולינומים אלו במדעי המחשב. המחקר של פולינומים במדעי המחשב אינו חדש. פולינומים באו לידי שימוש בתחומים כגון תכנון אלגוריתמים, הצפנה, חסמים תחתונים למעגלים, למידה חישובית ותורת הקודים. בעבודה זו אנו ממשיכים את המחקר של פולינומים במדעי המחשב, כאשר אנו מתמקדים בארבעה תחומים עיקריים: אנו חוקרים פולינומים כמודל חישובי, ובוחנים מספר בעיות יסוד כגון הקשר בין חישוב מדויק ומקורב, המשאבים הדרושים כדי לבנות פונקציות הקשות לחישוב, וכן חשיבות שדה הבסיס; אנו חוקרים התפלגויות פסאודו-אקראיות עבור פולינומים, וכשימוש מוכיחים חסמים תחתונים למעגלים בעומק חסום; אנו חוקרים בעיות בתורת הקודים הקשורות לקודים הנובעים מפולינומים, כגון קודי Reed-Muller או BCH; וכן אנו חוקרים בחינת תכונות עבור פולינומים, בעיה אשר מונעת מהתקדמויות חדשות בקומבינטוריקה אדיטיבית. בכל התחומים הללו, אנו מביאים תוצאות חדשות אשר נובעות מההבנה המשופרת של תכונותיהם הבסיסיות של פולינומים מעל שדות סופיים.

Part I

Introduction

Chapter 1

Overview

This work studies polynomials over finite fields and their applications in computer science. Polynomials are basic mathematical objects, studied in various areas such as algebra, analysis and combinatorics. The study of polynomials in computer science is also not new. Polynomials have found applications in the areas of algorithm design, cryptography, circuit lower bounds, computational learning and coding theory, to name a few key examples. In this work we continue the line of research of the role of polynomials over finite fields in computer science, where we focus our attention on four different research areas.

Our first line of research views polynomials as a computational model and studies several fundamental problems in this model. Our second line of research studies pseudorandomness for polynomials, that is, distributions which cannot be distinguished from the uniform distribution by polynomials. The third line of research studies coding theoretic problems related to codes arising from polynomials, such as Reed-Muller and BCH codes. The fourth line of research studies property testing for polynomials, that is, testing whether a given function is close to a polynomial. This fourth line of research is motivated by several recent developments in additive combinatorics.

It is important to note that all the research presented here focuses on polynomials defined over *constant-sized* finite fields, for example over $\mathbb{F}_2$. This is different from another line of research arising in arithmetic complexity, which studies polynomials defined by arithmetics circuits, which are usually defined over very large (or infinite) fields.

We proceed to review some of the mathematical background required for this work and the previous works in complexity theory related to polynomials. We then provide a high-level description of the results presented in this work. A detailed description of these results is provided in the next chapter.

Finite fields. A finite field is a finite set $\mathbb{F}$ endowed with two operations: addition and multiplication, which are commutative, transitive and distributive. The simplest example for a finite field is $\mathbb{F}_2 = \{0, 1\}$, the field of 2 elements, where addition corresponds to *exclusive-or* and multiplication to the *and* operation. More generally, for any prime p there exists a (unique) finite field with p elements $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$, where addition and multiplication corresponds to these operations *modulo* p and for every prime power $q = p^t$ there exists a (unique) finite field $\mathbb{F}_q$ with q elements. In this work we focus on polynomials defined over

finite fields (i.e. no infinite fields).

Polynomials. Let $\mathbb{F}$ be a finite field. A multi-variate polynomial over $\mathbb{F}$ in n variables $x_1, \ldots, x_n$ is a linear combination of monomials, where a monomial is a product of several variables. The total degree of a polynomial is the maximal number of variables participating in a monomial, counting multiplicities. For example,

$$f(x_1, x_2, x_3, x_4) = x_1 + 2x_2x_3^2 + x_1x_3 + x_2x_4$$

is a polynomial in 4 variables of total degree 3. In general, we can write polynomials as

$$f(x_1, \ldots, x_n) = \sum_{e_1, \ldots, e_n \geq 0} \alpha_{e_1, \ldots, e_n} x_1^{e_1} \ldots x_n^{e_n}$$

where $\alpha_{e_1,\ldots,e_n} \in \mathbb{F}$ and the total degree of f is the maximal $e_1 + \ldots + e_n$ for which $\alpha_{e_1,\ldots,e_n} \neq 0$. Unless stated otherwise, whenever we speak of the degree of f we mean the total degree of f, and we denote it by $\deg(f)$.

Polynomial representation of functions Working over finite fields, it is often convenient to identify functions with polynomials. Fix a finite field $\mathbb{F}_q$. We can think about polynomials $f(x_1, \ldots, x_n)$ over $\mathbb{F}_q$ as computing a function $f : \mathbb{F}_q^n \to \mathbb{F}_q$ and identify polynomials that compute the same function. To this end, we use the following identity: for any element $a \in \mathbb{F}_q$ we have that $a^q = a$. When q is prime this is known as Fermat's little theorem. Thus, we can reduce any individual degree of a variable in a monomial modulo q and not effect the function computed by the polynomial. This gives a canonical representation for functions $f : \mathbb{F}_q^n \to \mathbb{F}_q$ as polynomials. That is, for any function $f : \mathbb{F}_q^n \to \mathbb{F}_q$ there exists a unique set of coefficients $\{\alpha_{e_1,\ldots,e_n} \in \mathbb{F}_q : 0 \leq e_1, \ldots, e_n \leq q-1\}$ such that

$$f(x_1, \ldots, x_n) = \sum_{0 \leq e_1, \ldots, e_n \leq q-1} \alpha_{e_1, \ldots, e_n} x_1^{e_1} \ldots x_n^{e_n}.$$

We refer to this as the *polynomial representation* of the function f.

Representations of functions as polynomials have been studied intensively in computer science [NS92, Pat92, Bei93, BBR94]. This algebraic view of functions has found applications in diverse areas including circuit lower bounds [Raz87, Smo87, BRS, ABFR94], computational learning [KM93, LMN93, KS01, MOS03] and explicit combinatorial constructions [Gro00, Gro02, Gop06b, Efr09]. As a purely algebraic model of computation, polynomial representations lead to some natural complexity measures such as exact degree, approximation degree and sparsity needed to represent a function.

Present work In this work, we focus on four main areas, where polynomials over finite fields seem to be the most prevalent.

1. **Polynomials as a computational model:** polynomials form a very natural model of computation in the area of algebraic algorithms, such as matrix multiplication, integer

and polynomial factorization, discrete Fourier transforms and more. Computational models for polynomials have been intensively studied in the area of arithmetic circuit complexity, where the focus is commonly on high degree polynomials over large (or infinite) fields. In this work we focus on polynomials over small fields and study various questions relating to such polynomials when viewed as a computational model. In particular, we study the relation between exact computation and approximate computation as well as functions which are hard for this model of computation. Polynomials can be used to simulate many computational models, either exactly or with some small error. We study the computational power of polynomials in these settings and in particular the importance of the base field over which the polynomials are defined.

2. **Pseudorandom generators for polynomials**: We study a specific problem relating to the computational power of polynomials: can one construct small efficient distributions which can "fool" polynomials, in the sense that low-degree polynomials cannot distinguish such distributions from genuine uniform distributions? Such distributions can then by used to "fool" other computational models which can be simulated by polynomials (for example small depth circuits). In this work we study this problem in two contexts: when the polynomials are defined over a finite field (which is the more natural algebraic setting) and when the polynomials are defined only over binary inputs (which is the setting required for proving results for circuit lower bounds).

3. **Polynomials in coding theory**: Polynomials form the basis for very natural codes, such as Reed-Solomon, Reed-Muller and BCH codes. Despite being well-studied codes, many fundamental problems in coding theory which relate to these codes remain unsolved. In this work we study the relation between properties of polynomials and the properties of these codes, and manage to achieve new results in coding theory based on the improved understanding of properties of polynomials. These coding theoretic results include: an asymptotic estimate on the weight distribution and list-decoding size of Reed-Muller codes; a surprising sparsity result on the weights of low-degree Reed-Muller codes; and a new local-decoding results for affine invariant codes, which are an extension of dual-BCH codes.

4. **Property testing for polynomials**: Property testing is an area which studies whether one can infer global properties of objects just by looking at small fragments of these objects. Polynomials fit this framework when one tries to estimate whether a specific given function is close to a low-degree polynomial and wishes to do so by only examining a small number of evaluations of this function. Such problems arise naturally in the area of PCPs (Probabilistically Checkable Proofs) and also in the field of additive combinatorics, where it is related to several recent advances and is one of the key problems in this area. In this work we prove hardness results on the problem of estimating whether a function is close to a low-degree polynomial.

In the following chapter we provide a detailed summary of the results conveyed in this work. For each research area, we survey its history, its larger framework and related results obtained in this work. The full details regarding each specific work are presented in the

subsequent chapters and are grouped according to the herein outlined four main areas. We also provide a list of open problems for further research.

Chapter 2

Summary of the results

2.1 Polynomials as a computational model

Polynomials form a very natural computational model, when one studies algebraic problems, either with regards to specific algorithms, or for proving lower bounds for various computational tasks. Consider an $n \times n$ matrix. Its determinant is a polynomial in the n^2 elements of the matrix of degree n. It can be efficiently computed in polynomial time using Gaussian elimination. On the other hand, the permanent of the matrix is also a polynomial of degree n in n^2 variables, however no known efficient algorithm is known for computing it. In fact, the best algorithm requires time exponential in n. Thus, computing the determinant is "easy", while computing the permanent is believed to be "hard".

In order to define formally the computational resources required to compute a polynomial, it is common to consider a circuit computing the polynomial. The inputs to the circuit are the variables of the polynomial (and possibly constants in the field), and each gate may compute either the sum or the product of its inputs. The complexity of the circuit is measured by its size (the number of gates it contain) and its depth (the maximal length between an input and the output of the circuit). This model of computation has been studied extensively in the field of *arithmetic computational complexity*. The common framework in this field is to assume that the base field is very large, and to explore the structure of circuits computing various polynomials.

The main focus of this work is on polynomials over small finite fields, for example over $\mathbb{F}_2$, the field of two elements. This is a completely different framework, and in fact, many results which are proved in arithmetic computational complexity are false for small finite fields, and are true only if the base finite field is large enough. The most basic complexity measure for polynomials (in particular over small finite fields) is their degree. One may consider other, more delicate, complexity measures, such as the sparsity of the polynomial (the number of monomials with non-zero coefficient) or the minimal size of a circuit computing it. Still, many basic problems relating to the degree of a polynomial as a computational complexity measure are still far from being understood. Thus, we feel safe to limit our discussion to the degree of a polynomial as a measure for its complexity.

We study three questions in regards to the computational power of polynomials. The first one relates to the difference between exact computation and approximation, which we explore

in Subsection 2.1.1. The second relates to the problem of finding functions which cannot be approximated by low-degree polynomials, which we discuss in Subsection 2.1.2. The third relates to understanding the importance of the base field in polynomial representation of functions. We study this problem in Subsection 2.1.3.

2.1.1 Exact computation vs. approximation

In general, the problem of exactly computing a function is harder than just approximating the functions. It is often the case that the computational resources required to approximate a function are far below those required to exactly compute it. In a joint work with Tali Kaufman [KL08] we show that when one studies this distinction in the computational model of polynomials, then these two problems of exact computation and approximation are in fact equivalent, at least from a qualitative point of view.

We start by formally defining the notion of approximation. In order to keep this discussion as simple as possible, let us focus on functions over $\mathbb{F}_2$, i.e. functions $f : \mathbb{F}_2^n \to \mathbb{F}_2$. We say a function $g : \mathbb{F}_2^n \to \mathbb{F}_2$ can approximate f with error ϵ if

$$\Pr_{x_1,\ldots,x_n}[g(x_1,\ldots,x_n) = f(x_1,\ldots,x_n)] \geq \frac{1+\epsilon}{2}.$$

Otherwise put, the function g has a *bias* of ϵ over a random guess for the output of f.

Consider now the problem of approximation for polynomials by lower-degree polynomials (that is, approximation by "weaker" functions in our model of computation). Consider for example the polynomial

$$f(x_1,\ldots,x_n) = x_1 + x_2(x_3 + \ldots + x_n).$$

This is a quadratic polynomial. However, it can be approximated by the linear polynomial x_1, since

$$\Pr_{x_1,\ldots,x_n}[f(x_1,\ldots,x_n) = x_1] = 3/4.$$

Hence, the polynomial f can be approximated by a polynomial of lower degree. In fact, the polynomial f can be computed exactly by a function of 3 linear polynomial: x_1, x_2 and $x_3 + \ldots + x_n$. So, in this specific example, f was approximated by a single linear polynomial, but in fact was also computed exactly by a function of a small number of linear polynomials.

This main problem we study is whether this is always the case. We prove that indeed it is. If f can be approximated with error ϵ by a lower degree polynomial g, then in fact one may find a constant number k of lower degree polynomials (where the constant k depends only on the degree and the error of approximation, but crucially not on the number of variables), such that f can be computed by these polynomials.

Theorem (Approximation implies exact computation by several polynomials)**.** *Let $f(x_1,\ldots,x_n)$ be a degree-d polynomial over $\mathbb{F}_2$. Assume there exists a polynomial $g(x_1,\ldots,x_n)$ over $\mathbb{F}_2$ of degree at most $d-1$, such that*

$$\Pr_{x_1,\ldots,x_n}[f(x_1,\ldots,x_n) = g(x_1,\ldots,x_n)] \geq \frac{1+\epsilon}{2}.$$

Then there exist $k = k(d, \epsilon)$ polynomials $g_1, \ldots, g_k$ of degree at most $d-1$, and a composing function $G : \mathbb{F}_2^k \to \mathbb{F}_2$, such that

$$f(x_1, \ldots, x_n) = G(g_1(x_1, \ldots, x_n), \ldots, g_k(x_1, \ldots, x_n)).$$

The theorem can also be extended over arbitrary prime finite fields. It generalizes a previous result of Green and Tao [GT07] which held only for large fields, i.e. when $|\mathbb{F}| > d$. Our result, on the other hand, holds for all constant finite fields and all constant degrees, and in particular over $\mathbb{F}_2$ which is of significance in computer science.

The theorem can be used to prove the following corollary. Consider the following computational model: functions which can be computed by a small number of low-degree polynomials. Say a function has degree (d, k) if it can be computed by some function on k polynomials of degree d. That is, $f : \mathbb{F}_2^n \to \mathbb{F}_2$ has degree (d, k), if there exist k polynomials $g_1, \ldots, g_k$ of degree at most d, and some composing function $G : \mathbb{F}_2^k \to \mathbb{F}_2$ such that

$$f(x_1, \ldots, x_n) = G(g_1(x_1, \ldots, x_n), \ldots, g_k(x_1, \ldots, x_n)).$$

We show that if a polynomial of degree d can be approximated by a function of degree $(d-1, k)$, then it can be computed exactly by a function of degree $(d-1, k')$, where k' depend only on k, d, ϵ. Crucially, it does not depend on the number of variables n. That is, if we fix d and think of k, ϵ as arbitrary constants, then exact computation and approximation are qualitatively equivalent.

Corollary (Equivalence of approximation and computation by several polynomials)**.** *Let $f(x_1, \ldots, x_n)$ be a degree-d polynomial over $\mathbb{F}_2$. Assume there exists a function $g : \mathbb{F}_2^n \to \mathbb{F}_2$ of degree $(d-1, k)$, such that*

$$\Pr_{x_1, \ldots, x_n}[f(x_1, \ldots, x_n) = g(x_1, \ldots, x_n)] \geq \frac{1+\epsilon}{2}.$$

Then there exists $k' = k'(k, d, \epsilon)$ polynomials $g_1, \ldots, g_{k'}$ of degree at most $d-1$, and a composing function $G : \mathbb{F}_2^{k'} \to \mathbb{F}_2$, such that

$$f(x_1, \ldots, x_n) = G(g_1(x_1, \ldots, x_n), \ldots, g_{k'}(x_1, \ldots, x_n)).$$

As before, the corollary can also be extended over arbitrary prime finite fields. The proofs of the theorem and corollary are given in the full paper, which can be found in Chapter 4.

2.1.2 Hard functions for polynomials

Two boolean functions $f, g : \{0,1\}^n \to \{0,1\}$ are said to be *ϵ-correlated* if

$$\Pr[f(x) = g(x)] \geq \frac{1+\epsilon}{2}.$$

A function $f : \mathbb{F}^n \to \mathbb{F}$ is said to be *ϵ-correlated* with a set of functions $\mathcal{F} \subseteq \mathbb{F}^n \to \mathbb{F}$ if it is ϵ-correlated with at least one function $g \in \mathcal{F}$. We are interested in functions that have a low correlation with the set of degree d polynomials; namely, functions that cannot

be approximated by any polynomial of total degree at most d. How *complex* must such a function be? We use the most natural measure for complexity in these settings, which is the degree of the function when considered as a polynomial.

A simple probabilistic argument shows that for any constant $\delta < 1$ and for $d < \delta n$, a random function has an exponentially small correlation with degree $d-1$ polynomials. However, a random function is complex since, with high probability, its degree is at least $n-2$. In a joint work with Ido Ben-Eliezer and Rani Hod [BEHL09], we study how well a random degree d polynomial can be approximated by any lower degree polynomial, and show that with very high probability a random polynomial of degree d cannot be approximated by polynomials of lower degree in a strong sense. Thus, if we want to find functions that are uncorrelated with degree $d-1$ polynomials, considering degree d polynomials is enough.

The correlation of a typical degree d polynomial with the set of lower degree polynomials is a natural question in arithmetic complexity. More generally, the study of the correlation of functions with the set of low degree polynomials is interesting from both coding theory and complexity theory points of view.

Complexity Theory. Approximation of functions by low degree polynomials is one of the main tools used in proving lower bounds for constant depth circuits. For example, [Raz87] and [Smo87] provided an explicit function MOD_3 that cannot be computed by a constant depth circuit with a subexponential number of AND, OR and XOR gates. The proof combines two arguments:

1. Any constant depth circuit of subexponential size has a very high correlation (that is, $1-o(1)$) with some polynomial of degree n^{ϵ};

2. Such a low degree polynomial has a correlation of at most $2/3$ with MOD_3. (In fact, this is true for any polynomial of degree at most $\epsilon\sqrt{n}$ for some constant ϵ.)

The best known constructions of explicit functions that cannot be approximated by low degree polynomials (see, e.g., [BSK08, BNS, Raz87, Smo87, VW08] fall into two categories:

- For large degree bounds (i.e., bounds as $d < n^{\Omega(1)}$), there exists a symmetric function with a correlation of at most $O(1/\sqrt{n})$ with degree $O(\sqrt{n})$ polynomials;

- For small degree bounds (i.e., bounds as $d < \log n$) there are explicit functions having a correlation of at most $\exp(-n/c^d)$ with degree d polynomials for some constants c (best known is $c = 2$).

Certain applications, for instance, pseudorandom generator constructions via the Nisan–Wigderson construction [NW94], require a function having an exponentially small correlation with low degree polynomials. This is only known for degrees up to $\log n$, while for larger degrees the best known correlation bound is only polynomial in $1/n$. Finding explicit functions with a better correlation is an ongoing quest with limited success. For more details, see a survey by [Vio09].

Coding Theory. The Reed–Muller code $\mathrm{RM}(n,d)$ is a linear code in which codewords correspond to polynomials (over $\mathbb{F}$) in n variables of total degree at most d. This family of codes is one of the most studied objects in coding theory (see, e.g., [MS83]). Nevertheless, determining the weight distribution of these codes (for $d \geq 3$) is a long standing open problem. Interpreted in this language, our result gives a new tail estimate on the weight distribution of Reed–Muller codes.

Our results We show that, with very high probability, a random degree d polynomial has an exponentially small correlation with polynomials of lower degree, that is, of degree at most $d-1$. We prove this for degrees ranging from a constant up to $\delta_{\max} n$, where $0 < \delta_{\max} < 1$ is an absolute constant. All results hold for large enough n.

Theorem (Random degree d polynomial cannot be approximated by lower degree polynomials)**.** *There exist constants $0 < \delta_{\max} < 1$ and $c, c' > 0$ such that the following holds. For every $d \leq \delta_{\max} n$ let f be a random n-variate polynomial of degree d. Then the probability that f has a correlation $2^{-cn/d}$ with polynomials of degree at most $d-1$ is at most $2^{-c'\binom{n}{\leq d}}$, where $\binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i}$.*

The proof of the theorem is given in the full paper, which can be found in Chapter 5.

2.1.3 Representation of boolean functions as polynomials in different characteristics

The study of polynomial representations of boolean functions dates at least as far back as the 1960's, when they arose in various contexts including switching theory [Mur71], voting theory [Cho61] and machine learning [MP68]. Representations of boolean functions over finite fields, especially over $\mathbb{F}_2$ were studied by coding theorists in the context of Reed-Muller codes, see [MS83, Chapters 13-14] and the references therein. The codewords of the code $\mathrm{RM}_2(d,n)$ are all boolean functions $f : \{0,1\}^n \to \{0,1\}$ where $\deg_2(f) \leq d$, while received words are arbitrary functions f.

Polynomial representations have proved especially useful in circuit complexity [Bei93] where a natural lower bound technique is to relate concrete complexity measures (such as circuit-size) which we wish to bound, to purely algebraic complexity measures. Examples of this paradigm include the Razborov-Smolensky lower bounds for $\mathrm{AC}_0[p]$ [Raz87, Smo87], which relates the circuit size to the polynomial degree needed to approximate f over $\mathbb{F}_p$, and the work of Beigel et al. [BRS] and Aspnes et al. [ABFR94] which relate AC_0 circuit size with approximations by real polynomials.

Polynomial representations are among the most powerful tools in computational learning. The best learning algorithms for many basic concept classes, including but not limited to decision trees [KM93], DNF formulae [KS01], AC_0 circuits [LMN93, JCJS02], juntas [MOS03] and halfspaces [ARKS, KKMS05] all proceed by showing that the concept class to be learned has some *nice* polynomial representation. In particular, the algorithm for learning juntas of [MOS03] exploits a connection between $\deg_2(f)$ and the sparsity of its Fourier polynomial.

Polynomial representations of boolean functions have also found applications to constructing combinatorial objects such as set systems [Gro00, Gro02], Ramsey graphs [Gro00, Gop06b] and locally decodable codes [Efr09].

As a purely algebraic model of computation, polynomial representations lead to some natural complexity measures such as exact degree, approximation degree and sparsity needed to represent a function. In this chapter, we are primarily concerned with the polynomial degree of a function, when studied over different characteristics. Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function. The degree of f in characteristic k, denoted $\deg_k(f)$, is the minimal degree of a polynomial over $\mathbb{Z}_k$ computing f; that is, it is the degree of the unique multilinear polynomial $P(X_1, \ldots, X_n) \in \mathbb{Z}_k[X_1, \ldots, X_n]$ such that $P(x) = f(x)$ for every $x \in \{0,1\}^n$. We denote by $\deg(f)$ the degree of f over $\mathbb{R}$. We note that if $k'|k$ then $\deg_{k'}(f) \leq \deg_k(f)$ (since the polynomial over $\mathbb{Z}_k$ can be reduced modulo k') and that for the same reason, $\deg_k(f) \leq \deg(f)$ for all k.

The degree of a boolean function may be very dependent on the characteristic. Consider for example the parity function,

$$\text{Parity}(x_1, \ldots, x_n) = x_1 \oplus x_2 \oplus \ldots \oplus x_n.$$

The parity function can be expressed by a linear function over $\mathbb{F}_2$; it is simply the sum of the bits modulo 2. However, when considered modulo any odd k, the polynomial degree of parity is n, since the (unique) multilinear polynomial computing it is given by

$$\text{Parity}(x_1, \ldots, x_n) = \frac{1}{2}\Big(1 - (1-2x_1)(1-2x_2)\ldots(1-2x_n)\Big).$$

Thus, we have $\deg_2(\text{Parity}) = 1$ and $\deg_k(\text{Parity}) = n$ for all odd k. The main question studied is whether this is the general case. In a joint work with Parikshit Gopalan and Amir Shpilka [GLS09] we show that this is an instance of a more general principle:

A function on all n variables which has low degree in characteristic p is bound to have high degree in every other prime characteristic $q \neq p$.

Theorem (A boolean function with low degree modulo p must have high degree modulo all other q)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which depends on all n variables. Let $p \neq q$ be distinct primes. Then*

$$\deg_q(f) \geq \frac{n}{\lceil \log_2 p \rceil \deg_p(f) p^{2\deg_p(f)}}.$$

This gives a lower bound of $\Omega(n^{1-o(1)})$ on $\deg_q(f)$ as long as $\deg_p(f) = o(\log n)$. This bound is close to the best possible, as there exist functions on all n variables (such as the addressing function [NS92]) where $\deg(f) \leq \log n$ and hence $\deg_p(f) \leq \log n$ for all characteristics p. Thus, one cannot get nontrivial lower bounds on $\deg_q(f)$ once $\deg_p(f)$ exceeds $\log n$.

Nisan and Szegedy showed that any function on n variables must have degree at least $\deg(f) \geq \log n - O(\log \log n)$ [NS92]. An interesting consequence of our main theorem is the following analog of the Nisan-Szegedy bound for non-prime power moduli.

Corollary (Boolean functions must have high degree modulo composites)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which depends on all n variables. Suppose m is not a prime power, and p is its smallest prime divisor. We have*

$$\deg_m(f) \geq \frac{1}{2}\log_p n - \log_p \log_p n - \frac{1}{2}\log_p \lceil \log_2 p \rceil \ .$$

This corollary is interesting as it illuminates a sharp difference between degrees over composite numbers and over primes. A simple way to construct boolean functions of degree $O(1)$ over $\mathbb{F}_p$ is to take any constant degree polynomial $P(x_1, \ldots, x_n) \in \mathbb{F}_p[x_1, \ldots, x_n]$ and raise it to the power $p-1$. This construction fails for composite m since there is no analog of Fermat's little theorem. The corollary shows that indeed any polynomial modulo m computing a boolean function requires degree $\Omega(\log n)$, as it does over the reals.

We can also obtain the following stronger bound for the real degree of f (i.e. $\deg(f)$) in the case that f has low degree modulo some prime p. The proof of the following theorem follows an adaptation of the original proof of Nisan and Szegedy.

Lemma (Boolean functions with low degree modulo p must have high real degree)**.** *Let p be a prime and $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which depends on all n variables. Then*

$$\deg(f) \geq \frac{n}{2^{\deg_p(f)}} \ .$$

The results above show a very basic relation between the degrees of boolean functions over different characteristics. A natural question to ask is what happens if we relax the requirement and only consider polynomials over $\mathbb{F}_q$ that approximate a low degree polynomial over $\mathbb{F}_p$. Similarly to the case of degree 1 polynomials that was studied in [Smo87], we prove that low degree polynomials modulo p are hard to even approximate by polynomials in other characteristics.

Theorem (Boolean functions with low degree modulo p cannot be well-approximated by low-degree polynomials modulo q)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a function depending on all n variables with $\deg_p(f) = d$. Then, for any $q \neq p$ and any $\mathbb{F}_q$ polynomial $Q(x_1, \ldots, x_n) : \mathbb{F}_q^n \to \{0,1\}$, satisfying $\deg_q(Q) = o\left(\sqrt{\frac{n}{dp^{3d}}}\right)$, it holds that*

$$\Pr_{x \in \{0,1\}^n}[f(x) = Q(x)] \leq 1 - \epsilon p^{-d} \ ,$$

where ϵ depends only on p, q.

We note that both the error bound of $1 - p^{-O(d)}$ and the degree bound of $o(\sqrt{n})$ are close to optimal; there are polynomials of degree d over $\mathbb{F}_p$ that are 0 on the boolean hypercube with probability $1 - 2^{-d}$, hence they have trivial approximations over $\mathbb{F}_q$. Secondly, the Mod_p function (and indeed every symmetric function) can be $1-\epsilon$ approximated by polynomials of degree $c(\epsilon)\sqrt{n}$ over $\mathbb{F}_q$ [BGL06], despite being hard to approximate for polynomials of lower degree.

As a corollary of the theorem for inapproximability of low degree polynomials over $\mathbb{F}_p$ by polynomials over $\mathbb{F}_q$, we get that if a boolean function has low degree modulo p, then

the function requires large $\mathrm{AC}_0[q]$ circuits for any prime $q \neq p$. Several of the known lower bounds for $\mathrm{AC}_0[q]$ are for functions like Par and the Mod_{p^k} function where $p \neq q$ that are easily seen to be low-degree polynomials in some characteristic. Our result generalizes this to give a very general class of hard functions for $\mathrm{AC}_0[q]$, namely all functions that have degree $o(\log n)$ modulo $p \neq q$.

Theorem (Boolean functions with low degree modulo p require large $\mathrm{AC}_0[q]$ circuits)**.** *Let p, q be distinct primes. Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which depends on all n variables with $\deg_p(f) = o(\log_p n)$. Then any $\mathrm{AC}_0[q]$ circuit of depth t computing f requires size at least $\exp(n^{(1-o(1))/2t})$.*

It is not hard to see that most known lower bounds for $AC_0[q]$ circuits follow from the theorem above. For example, the lower bound for Mod_{p^k} of [Smo87] follows from the observation that $\deg_p(\mathsf{Mod}_{p^k}) \leq p^k$ (see e.g. [BGL06]). Additionally, it gives several new lower bounds, for instance it shows that every quadratic form on n variables over $\mathbb{F}_2$ requires large $\mathrm{AC}_0[q]$ circuits, for $q \neq 2$. Though we note that this theorem does not imply Razborov's lower bound for $\mathsf{Majority}$.

Summarizing, our results show that for a boolean function, having low degree mod p, or even being close to a low degree polynomial mod p, is a "singular" event, in the sense it can only occur for at most one characteristic p. The proofs of the theorems above are given in the full paper, which can be found in Chapter 6.

2.2 Pseudorandom generators for low-degree polynomials

Pseudorandomness is the theory of generating objects that "look random" despite being constructed using little or no randomness. A primary application of pseudorandomness is to address the question: *Are randomized algorithms more powerful than deterministic ones?* That is, how does randomization trade off with other computational resources? Can every randomized algorithm be converted into a deterministic one with only a polynomial slowdown (i.e. does BPP = P) or with only a constant-factor increase in space (i.e. does RL = L)? The study of both these questions has relied on pseudorandom generators that fool algorithms of limited computational powers.

A *pseudorandom distribution* is a distribution over a domain, which cannot be distinguished from the uniform distribution over this domain by "weak" alogrithms. A *pseudorandom generator* (PRG for short) is an explicit function whose output distribution is a pseudorandom distribution. Formally, a PRG over a domain D (one should think of D as either the boolean cube $\{0,1\}^n$ or as $\mathbb{F}_p^n$) for a family of *tests* $\mathcal{T}$ is an explicit function $G : D^r \to D^n$ such that no test $T \in \mathcal{T}$ can distinguish a random output of G from truly uniform input elements in D^n. Namely,

$$\max_{T \in \mathcal{T}} \left| \Pr_{x \in D^r}[T(G(x)) = 0] - \Pr_{x \in D^n}[T(x) = 0] \right| \leq \epsilon \, .$$

Ideally, one would like to have the *seed* r as short as possible and the error ϵ to be as small as possible. A pseudorandom generator is considered efficient if the seed length is $O(\log n)$

(as in this case, for some applications, one can enumerate over all seeds to find a 'good' one). Pseudorandom generators have been a major object of study in theoretical computer science for several decades, and have found applications in the area of computational complexity, cryptography, algorithms design and more (see [Gol08, AB09]).

A family of tests that was widely considered in the literature is low degree polynomials over finite fields. Before stating the formal definition of a PRG for low degree polynomials we fix some notation: let f be a function, and $\mathcal{D}$ a distribution over the inputs of f. We denote by $f(\mathcal{D})$ the output distribution of f given inputs sampled according to $\mathcal{D}$. For a set S we denote by $f(S)$ the output distribution given that the inputs are uniformly sampled in S (for example, $f(\{0,1\}^n)$ is the distribution of f over uniform input bits). The statistical distance between two distributions $\mathcal{D}', \mathcal{D}''$ is given by $\mathrm{sd}(\mathcal{D}', \mathcal{D}'') = \frac{1}{2}\sum_e |\Pr_{\mathcal{D}'}[e] - \Pr_{\mathcal{D}''}[e]|$.

Definition (Pseudorandom generators for degree d polynomials)**.** *A distribution $\mathcal{D}$ taking values in $\mathbb{F}_p^n$ is* pseudorandom for degree d polynomials over $\mathbb{F}_p$ *with error ϵ if, for any degree d polynomial $f(x_1, \ldots, x_n)$ over $\mathbb{F}_p$, the distributions $f(\mathcal{D})$ and $f(\mathbb{F}_p^n)$ are ϵ-close in statistical distance. A function $G : \{0,1\}^r \to \mathbb{F}_p^n$ is a* pseudorandom generator for degree d polynomials over $\mathbb{F}_p$*, if the output distribution of G, given uniformly sampled seeds, is a pseudorandom distribution for degree d polynomials.*

We discuss pseudorandom generators in Subsection 2.2.1, and prove related lower bounds for such generators in Subsection 2.2.2. We also study a variant of this model, where we consider polynomials evaluated only on the boolean cube. In this case, a pseudorandom distribution is a distribution over $\{0,1\}^n$ (instead of $\mathbb{F}_p^n$) which cannot be distinguished from the uniform distribution over $\{0,1\}^n$ by low-degree polynomials. The motivation for this model is that it has tight relations to circuit lower bounds.

Definition (Pseudorandom bit-generators for degree d polynomials)**.** *A distribution $\mathcal{D}$ taking values in $\{0,1\}^n$ is* pseudorandom for degree d polynomials over $\mathbb{F}_p$ *with error ϵ if, for any degree d polynomial $f(x_1, \ldots, x_n)$ over $\mathbb{F}_p$, the distributions $f(\mathcal{D})$ and $f(\{0,1\}^n)$ are ϵ-close in statistical distance. A function $G : \{0,1\}^r \to \{0,1\}^n$ is a* pseudorandom bit-generator for degree d polynomials over $\mathbb{F}_p$*, if the output distribution of G, given uniformly sampled seeds, is a pseudorandom distribution for degree d polynomials.*

We give a construction of a pseudorandom bit-generator for linear polynomials in Subsection 2.2.3, and a construction for arbitrary constant degrees in Subsection 2.2.4, where we also use it to construction an efficient and explicit pseudorandom generators for the circuit class $\mathrm{CC}_0[p]$, the class of constant depth circuits composed from Mod_p gates.

2.2.1 Pseudorandom generators for low-degree polynomials

Fix a finite field $\mathbb{F}_p$. We study the problem of constructing explicit and efficient pseudorandom generators against low-degree multivariate polynomials over $\mathbb{F}_p$.

The case of pseudorandom generators against linear polynomials, usually called *small-bias generators* or *epsilon-biased generators*, was first studied (over $\mathbb{F}_2$) by Naor and Naor [NN93], who gave PRGs with $O(\log n)$ seed length. This was later generalized by Alon, Goldreich, Håstad and Peralta [AGHP90] to arbitrary fields. These constructions have a seed length

which is optimal up to a constant multiplicative factor. The construction of small-bias generators is a major tool in derandomization, PCPs and lower bounds (see [BSSVW03] and the references within for details regarding small-bias generators).

The generalization of the problem to constant-degree polynomials was first studied by Luby, Velickovic, and Wigderson [LVW93]. Their results apply, in fact, to the more general model of constant depth circuits. In the context of constant degree polynomials, they give an explicit construction of a PRG requiring $\exp(O(\sqrt{\log n/\epsilon}))$ random bits.

Bogdanov [Bog05] studied the problem of constructing PRG for polynomials over large fields. He gave a construction of a PRG in large fields, where the minimum field size required for his construction is polynomial in the degree, the required error and the log of the number of variables. In these settings, his construction is optimal up to polynomial factors. The proof of his result uses techniques and results from algebraic geometry and computational algebra.

A breakthrough result for small finite fields was obtained by Bogdanov and Viola [BV07]. They presented a novel approach for constructing a PRG for low-degree polynomials over small fields. Their construction is the sum of d independent small-bias generators. That is, if $G_1 : \{0,1\}^r \to \{0,1\}^n$ is a small-bias generator, then their generator $G_d : \{0,1\}^{rd} \to \{0,1\}^n$ is given by

$$G_d(s_1, \ldots, s_d) = G_1(s_1) \oplus G_1(s_2) \oplus \ldots \oplus G_1(s_d),$$

where $s_1, \ldots, s_d \in \{0,1\}^r$ and the exclusive-or is performed coordinate-wise. Bogdanov and Viola showed that, if a conjecture in additive combinatorics known as the *inverse conjecture for the Gowers norm* holds, then their construction is a PRG for degree-d polynomials. At the time, the inverse conjecture for the Gowers norm was known to hold only for degrees 2 and 3, and was conjectured to hold for all constant degrees. Thus, their construction was known to be correct only for quadratic and cubic polynomials.

Our work [Lov08] was inspired by the work of Bogdanov and Viola, with the goal of making their construction unconditional, i.e., not relying on any unproven conjectures. We prove that the sum of 2^d independent small-bias generators is pseudorandom against degree-d polynomials, without relying on any unproven conjectures.

Theorem (Pseudorandom generator for degree d polynomials)**.** *There exists a universal constant $c > 0$ such that the following holds. Let G_1 be a small-bias generator with error $\epsilon^{2^{cd}}$. Then the sum of 2^d independent copies of G_1 is pseudorandom against degree-d polynomials with error ϵ. In particular, this gives a pseudorandom generator for degree-d polynomials with error ϵ using $2^{cd} \log(|\mathbb{F}|n/\epsilon)$ random bits for the seed.*

The proof of the theorem is given in the full paper, which can be found in Chapter 7. Subsequent to this work, there were advances on two fronts.

First, the inverse conjecture for the Gowers norm was shown to be false for degrees $d \geq 4$ by Green and Tao [GT07] and independently by Lovett, Meshulam, and Samorodnitsky [LMS08]. A more refined inverse conjecture for the Gowers norm was suggested, and it was proved by Bergelson, Tao and Ziegler [BTZ09, TZ09].

Additionally, Viola [Vio08] proved the correctness of the original construction of [BV07] without using the inverse conjecture for the Gowers norm, or any other unproven conjectures, thus making the original construction of [BV07] unconditionally correct. The result presented

here can thus be seen as an intermediate step in a sequence of works. The proof of Viola uses some of the techniques developed in this work, in addition to some of the original techniques introduced in [BV07] and some clever new ideas.

2.2.2 Explicit lower bound for fooling polynomials by the sum of small-bias generators

Fix a finite field $\mathbb{F}_p$. Small-bias distributions are distributions over $\mathbb{F}_p^n$ which are pseudorandom against all linear functions; that is, any non-zero linear functional in the field elements is distributed almost uniformly. Viola [Vio08], following works of Bogdanov and Viola [BV07] and Lovett [Lov08] showed that the sum of d copies of small-bias generators is a pseudorandom generator against degree d polynomials.

Bogdanov and Viola have shown already in their original work [BV07] that taking d copies is essentially tight with respect to the number of copies needed; using a counting argument, they show that for a fixed error, any generator with output length n that fools all degree d polynomials must have seed length at least $d \cdot \log n - O(1)$. Thus, for every generator with shorter seed, there exists a polynomial of degree at most d that distinguishes a random output of the generator from truly random field elements. The seed length of a small-bias generator with bias ϵ is $O(\log(n/\epsilon))$. Thus, this shows that the sum of $d-1$ small-bias generators (with sufficiently large bias) cannot fool degree d polynomials.

This argument does not rule out the possibility that if one chooses small-bias generators with very small bias (say, even exponentially small in n), then one may get a generator for degree d polynomials by summing less than d copies of the small-bias generators. In a joint work with Yoav Tzur [LT09], we show that even when the bias of the small-bias generators can be exponentially small, still d independent copies are required in order to fool polynomials of degree d.

We give an explicit construction of a small-bias generator, and an explicit polynomial of degree $d+1$, such that this polynomial always evaluates to zero on inputs which are sums of d copies of the small-bias generator, and is almost uniform when evaluated over uniform inputs. Furthermore, our small-bias generator construction allows for exponentially small bias, whereas the proof of [BV07] allows only polynomially small bias.

Theorem (Sum of d small-bias generators cannot fool polynomials of degree $d+1$)**.** *For every $n, d \in \mathbb{N}$ and $\ell \geq 2d+1$, there exists an explicit small-bias generator $G_1 : \mathbb{F}_p^{2n} \to \mathbb{F}_p^{\ell n}$ with bias $\epsilon \leq \ell/p^n$ and with the following property. Let $G_d : (\mathbb{F}_p^{2n})^d \to \mathbb{F}_p^{\ell n}$ be the sum of d independent samples of G_1. Then there exists an explicit polynomial $f(x_1, \ldots, x_{\ell n})$ of degree $d+1$ over $\mathbb{F}_p$ such that*

- *The polynomial f evaluates to zero on any output of G_d.*
- *The distribution of f, when applied to uniform inputs, is $\frac{d}{p^n}$-close to the uniform distribution over $\mathbb{F}_p$.*

The proof the theorem is given in the full paper, which can be found in Chapter 8. We note that Alon et al. [ABEK08] showed almost tight lower bounds for the size of the sample space required to fool all degree d polynomials with given error ϵ. Their bounds relate to the

size of a general sample space, and not to the specific construction we are interested in, the sum of d independent samples of a small-bias distribution. Thus, their result does not imply the bounds of the form we are interested in - that we need to sum d copies of small-bias generator (and not less) in order to fool degree d polynomials.

2.2.3 Pseudorandom bit-generators for modular sums

The motivation for studying pseudorandom bit-generators for modular sums (i.e. linear functions over finite fields evaluated over bits) is motivated by the goal of constructing efficient pseudorandom generators against small-space computations, an instance of which is modular sums.

Small-space algorithms are algorithms which use little memory in order to perform calculations. An important open problem is whether randomization aids in computational tasks. While this problem is still wide open when considering polynomial-time computations, the analog of this problem for small-space computations has seen quite a few advances, and seems to be at reach. In particular, pseudorandom generators that fool space-bounded algorithms [AKS, BNS, Nis92, INW94] were highly instrumental in the study of the RL vs. L problem (e.g. used in the best known derandomization of RL [SZ99]).

While the currently available space-bounded generators are extremely powerful tools, their seed length is still suboptimal. For example, if we want to fool a $\log n$-space algorithm then known generators require $\log^2 n$ truly random bits (the seed) in order to generate up to polynomially many pseudorandom bits. On the other hand, for several interesting special cases we do know generators with almost optimal seed length. The special case which serves as a motivation for our work is that of small-biased generators [NN93]. These generators produce n bits $X_1, X_2, \ldots, X_n$ that fool all linear tests modulo 2. In other words, for each subset T of the bits, the sum $\Sigma_{i \in T} X_i \mod 2$ is uniformly distributed up to bias ϵ. Explicit constructions of ϵ-biased generators are known with seed-length $O(\log(n/\epsilon))$, which is optimal up to the hidden constant [NN93]. Even though linear tests may seem very limited, ϵ-biased generators have turned out to be very versatile and useful derandomization tools [NN93, MNN94, HPS93, Nao92, AM95, AR94, BSSVW03, BV07, Lov08, Vio08].

Given the several applications of distributions that fool linear tests modulo 2, it is natural to consider the question of fooling modular sums for larger moduli. It turns out that the notion of small-biased generators can be generalized to larger fields. Such generators produce a sequence $X_1, X_2, \ldots, X_n$ of elements in a field $\mathbb{F}$ that fool every linear test over $\mathbb{F}$ [Kat89, AIK$^+$90, RSW93, EGL$^+$98, AM95].

In a joint work with Omer Reingold, Luca Trevisan and Salil Vadhan [LRTV09], instead, we consider a different generalization of ϵ-biased generators where we insist on *bit*-generators. Namely we would like to generate a sequence $X_1, X_2, \ldots, X_n$ of bits that fool every linear test modulo a given number M. For every sequence $a_1, a_2, \ldots, a_n$ of integers in $\mathbb{Z}_M = \{0, 1, \ldots, M-1\}$ we want the sum $\Sigma_i a_i \cdot X_i \mod M$ to have almost the same distribution (up to statistical distance at most ϵ) as in the case where the X_i's are uniform and independent random bits. (Note that this distribution may be far from the uniform distribution over $\mathbb{Z}_M$, particularly when only a few a_i's are nonzero.) It turns out that even for $M = 3$ and even if we limit all the a_i's to be either ones or zeros, the best generators that were known prior to

this work are generators that fool general space-bounded computations [Nis92, INW94], and required a seed of length $O(\log^2 n)$. Therefore, obtaining better pseudorandom bit generators that fool modular sums may be considered a necessary step towards improved space-bounded generators. In addition, we consider this notion to be a natural generalization of small-bias generators, which is a central derandomization tool.

We give two constructions of pseudorandom bit generators that fool modular sums. Similarly to [MST06], each construction is actually comprised of two generators: one that fools summations $\Sigma_i a_i \cdot X_i$ in which only relatively few coefficients a_i are not zero (the "low-weight" case) and one that fools summations $\Sigma_i a_i \cdot X_i$ such that many coefficients a_i are not zero (the "high weight" case). The motivation is that fooling low-weight sums and fooling high-weight sums are tasks of a different nature. In the high-weight case, if R_i are truly random bits, then $\Sigma_i a_i \cdot R_i \mod M$ is almost uniformly distributed in $\mathbb{Z}_M$. Thus, in analyzing our generator, we just need to argue that $\Sigma_i a_i \cdot X_i \mod M$ is close to uniform, where $X_1, \ldots, X_n$ is the output of the generator.

On the other hand, in the low-weight case the distribution may be far from uniform and therefore we may need to imitate the behavior of a random sequence of bits more closely.

In each construction, we shall present two generators: one that is pseudorandom against low-weight sums, and one that is pseudorandom against high-weight sums. We shall then combine them by evaluating them on independently chosen seeds and XORing the two resulting sequences. We sketch below the two constructions. For full details, we refer to the full paper which can be found at Chapter 9.

Construction Based on Pseudorandom Generators for Polynomials In our first construction, we handle the case of $M = 3$ and any other fixed prime modulus M (in fact, our construction works also for any fixed prime power). For these cases, our seed length is $O(\log(n/\epsilon))$ as in the case of ϵ-biased generators (but the hidden constant depends exponentially on M).

As mentioned above, for every fixed finite field $\mathbb{F}$, there are nearly-optimal known generators that construct a small-bias distribution $X_1, \ldots, X_n$ of *field elements*, while our goal is to generate *bits*. A natural approach to construct a bit generator would be to sample a sequence of field elements $X_1, \ldots, X_n$ from a small bias distribution, to pick a function $g : \mathbb{F} \to \{0, 1\}$ appropriately, and to output the bits sequence $g(X_1), \ldots, g(X_n)$. Unfortunately the small bias property for $g(X_1), \ldots, g(X_n)$ does not seem to follow from the small bias property of $X_1, \ldots, X_n$.

If, however, we start from a sequence of field elements $X_1, \ldots, X_n$ that fools *polynomials* over $\mathbb{F}$, then we can make such an approach work, because g can be chosen to be itself a polynomial (of degree $\Theta(|\mathbb{F}|)$). However, note that when $|\mathbb{F}|$ is odd, g cannot be balanced, and thus $g(X_1), \ldots, g(X_n)$ are only indistinguishable from independent *biased* coins. Thus, this approach only works when the sum has sufficiently high weight so that both biased and unbiased random bits will yield a sum that is almost uniformly distributed over $|\mathbb{F}|$; specifically we need at least k non-zero coefficients a_i, where $k = O(M^2 \log 1/\epsilon)$. For fixed M, there are known constructions [BV07, Lov08, Vio08] of pseudorandom generators that fool polynomials of degree d over $\mathbb{F} = \mathbb{Z}_M$, M prime, and which only require seed length $O_{M,d}(\log n/\epsilon)$.

In order to fool low-weight sums, we observe that a bit generator $X_1, \ldots, X_n$ which is ϵ-almost k-wise independent fools, by definition, every sum $\sum_i a_i X_i \bmod M$ of weight at most k, and that such generators are known which require only seed length $O(\log n + k + \log 1/\epsilon)$.

A similar construction was independently discovered by Meka and Zuckerman [MZ09]

Construction Based on the INW Generator In our second construction, we give a pseudorandom bit generator that fools sums modulo *any* given M (not necessarily prime) with seed length $O(\log n + \log(M/\epsilon)\log(M\log(1/\epsilon)))$. In both the low-weight and high-weight cases, this generator relies on versions of the Impagliazzo–Nisan–Wigderson [INW94] pseudorandom generator for space-bounded computation. Of course, modular sums are a special case of space-bounded computations, and thus we could directly apply the INW generator. But this would require seed length larger than $\log^2 n$. We obtain better bounds by more indirect use of the INW generator inside our construction.

The most interesting technical contribution underlying this construction is a new analysis of the derandomized graph squaring operation of Rozenman and Vadhan [RV05], which captures the effect of using the INW generator to derandomize random walks on graphs. Here we study the analogue of derandomized squaring for taking products of two distinct Cayley graphs over an abelian group (namely $\mathbb{Z}_M$). The advantage of the new analysis is that it handles graphs that have distinct bounds on their expansion, and works for bounding each eigenvalue separately. This is then used to produce pseudorandom-walks where each step is taken on a different abelian Cayley graph (rather than pseudorandom walks on a single graph as in [RTV06, RV05]).

2.2.4 Pseudorandom bit-generators for low-degree polynomials

We consider the problem of constructing pseudorandom bit-generators against degree d polynomials, generalizing the previous construction from linear polynomials to arbitrary constant degree polynomials. The motivation for this is that such generators will fool in particular the circuit class $\mathrm{CC}_0[p]$, which is the class of constant depth circuits consisting of only Mod_p gates. We start by surveying the current state of knowledge with regards to pseudorandom generators against constant depth circuits. We then proceed to show the relating between pseudorandom bit-generators and the class $\mathrm{CC}_0[p]$, and then sketch in high level the construction of pseudorandom bit-generators against low-degree polynomials.

The family of constant depth circuits which probably received the most attention in computational complexity is the class AC_0. This is the class of constant-depth circuits with unbounded fan-in AND, OR and NOT gates. Håstad [Hås86] showed that the PARITY function cannot be approximated by any polynomial size AC_0 circuit. I.e., that no polynomial size AC_0 circuit agrees with parity on more than $\frac{1}{2} + \exp(-n)$ fraction of inputs. In other words, the *correlation* of PARITY with AC_0 is exponentially small. This result was later used by Nisan [Nis91] for constructing efficient pseudorandom generators for AC_0 (these pseudorandom generators use $r = \mathrm{polylog(n)}$ bits). Recently, following a breakthrough by Bazzi [Baz07], Braverman [Bra09] showed that any polylog-wise independent distribution is pseudorandom for AC_0 circuits, thus settling a conjecture of Linial and Nisan [LN90]. $\mathrm{AC}_0[p]$ is another well studied class of circuits, consisting of all constant-depth circuits with

unbounded fan-in AND, OR, NOT and MOD_p gates (a MOD_p gate outputs 1 if the sum of its inputs is divisible by p, and 0 otherwise). In contrast to the impressive success in constructing pseudorandom generators for AC_0, no PRGs are known for $\mathrm{AC}_0[p]$. One reason is that no strong correlation lower bounds are known for this class. Razborov and Smolensky [Raz87, Smo87] proved exponential lower bounds for $\mathrm{AC}_0[p]$ circuits and their results also imply correlation lower bounds, albeit those are much weaker than the ones known for AC_0. Namely, [Raz87, Smo87] showed that the MOD_q function has polynomially small correlation with $\mathrm{AC}_0[p]$ when p and q are co-prime. The class of $\mathrm{AC}_0[m]$ where m is not a prime power is only very weakly understood; in particular, currently we cannot separate it from NP!

Motivated by the problem of constructing pseudorandom generators for $\mathrm{AC}_0[p]$, in a joint work with Partha Mukhopadhyay and Amir Shpilka [LMS10] we study a natural subclass - $\mathrm{CC}_0[p]$ circuits. The class $\mathrm{CC}_0[p]$ is the class of constant depth circuits using only MOD_p gates. While exponential lower bounds for this class follow from the work of Smolensky [Smo87], no pseudorandom generator better than the one constructed in [LVW93] (whose seed length is $r = \exp(\sqrt{\log n})$) is known for it. Our main result is an explicit pseudorandom generator fooling any $\mathrm{CC}_0[p]$ circuit while using only $r = O(\log n)$ random bits, for any fixed error $\epsilon > 0$. Actually, our construction gives pseudorandom bit-generators for low-degree polynomials over finite fields, from which the result for $\mathrm{CC}_0[p]$ follows: Let $\mathbb{F}_p$ be a prime finite field. The MOD_p function can be computed by a degree $p-1$ polynomial over $\mathbb{F}_p$

$$\mathrm{MOD}_p(x_1, \ldots, x_n) = (x_1 + \ldots + x_n)^{p-1} \pmod{p} .$$

Hence, any depth k circuit in $\mathrm{CC}_0[p]$ can be computed by a polynomial over $\mathbb{F}_p$ of degree $d = (p-1)k$. Thus, in order to fool $\mathrm{CC}_0[p]$ we have to fool the distribution induced by low degree polynomials over $\mathbb{F}_p$, when evaluated on inputs from the boolean cube. In other words, we have to generalize the aforementioned results of [LRTV09, MZ09] from linear polynomials to any constant degree polynomials. This motivates the definition of bit-pseudorandom generators for polynomials. We recall the definition for the convenience of the reader.

Definition 2.1 (Bit-pseudorandom distributions for degree d polynomials)**.** *A distribution $\mathcal{D}$ taking values in $\{0,1\}^n$ is* bit-pseudorandom for degree d polynomials over $\mathbb{F}_p$ *with error ϵ if, for any degree d polynomial $f(x_1, \ldots, x_n)$ over $\mathbb{F}_p$, the distributions $f(\mathcal{D})$ and $f(\{0,1\}^n)$ are ϵ-close in statistical distance. A function $G : \{0,1\}^r \to \{0,1\}^n$ is a* bit-pseudorandom generator for degree d polynomials over $\mathbb{F}_p$ *if the output distribution of G over a uniform seed is a bit-pseudorandom distribution for degree d polynomials.*

Notice the difference between this definition and the definition of pseudorandom distributions is that in the latter case, one has to fool the distribution of the polynomial when evaluated over the entire space and not just over the boolean cube.

As mentioned above, PRGs for polynomials over small finite fields were studied in several works [LVW93, BV07, Lov08, Vio08]. The best result to date is by Viola.

Theorem (Theorem 1 in [Vio08])**.** *There exists an explicit and efficient function $G : \{0,1\}^r \to \mathbb{F}_p^n$ for $r = O(d \cdot \log(pn) + 2^d \cdot \log(1/\epsilon))$ such that $G(\{0,1\}^r)$ is pseudorandom for degree d polynomials over $\mathbb{F}_p$ with error ϵ.*

The problem of construction bit-pseudorandom generators for linear polynomials (i.e. the case of $d = 1$) was first studied by [LRTV09, MZ09] in the context of small-space computations. Before describing their generator we need a few notations. For $a = (a_1, \ldots, a_n) \in \mathbb{F}_p^n$ define $a^{p-1} = (a_1^{p-1}, \ldots, a_n^{p-1}) \in \{0,1\}^n$ to be the $p-1$ power of a. Similarly for a distribution $\mathcal{D} \subset \mathbb{F}_p^n$, define $\mathcal{D}^{p-1} \subset \{0,1\}^n$ by raising each element of $\mathcal{D}$ to the $p-1$ power. [LRTV09, MZ09] discovered the following construction for a bit-pseudorandom generator for linear polynomials over $\mathbb{F}_p$: the bitwise-XOR of the $p-1$ power of a pseudorandom distribution for degree $(p-1)$ polynomial over $\mathbb{F}_p$, and a k-wise independent distribution.

Our main result extends their result to any constant degree polynomial. We prove that the following is a bit-pseudorandom distribution for degree d polynomials over $\mathbb{F}_p$: the bitwise-XOR of the $p-1$ power of a pseudorandom distribution for degree $((p-1)d)$ polynomials over $\mathbb{F}_p$, and a k-wise independent distribution.

Theorem (Pseudorandom bit-generators for degree d polynomials)**.** *Let $\mathbb{F}_p$ be an odd prime finite field, $d \geq 1$ an integer and $\epsilon > 0$ an error parameter. Then there exist $\delta = \delta(p, d, \epsilon)$ and $k = k(p, d, \epsilon)$ such that the following holds. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $((p-1)d)$ polynomials with error δ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution. Then, the bitwise-XOR of the two distributions $\mathcal{D}^{p-1} \oplus K$ is a bit-pseudorandom distribution for degree d polynomials over $\mathbb{F}_p$ with error ϵ. The parameters k, δ satisfy*

$$k(p, d, \epsilon), \delta(p, d, \epsilon)^{-1} \leq \exp^{(2d+1)}(\epsilon^{-c_{p,d}})$$

where $\exp^{(t)}$ is the t-times iterated exponential function, and $c_{p,d} > 0$ is some constant which depends on p and d.

An immediate corollary is that there exists an efficient and explicit pseudorandom generator $G : \{0,1\}^r \to \{0,1\}^n$ fooling any depth-k circuit in $\mathrm{CC}_0[p]$ with error ϵ, where $r = c_{p,k,\epsilon} \cdot \log n$.

Corollary (Pseudorandom generators for $\mathrm{CC}_0[p]$)**.** *Let p be an odd prime number and $\epsilon > 0$ an error parameter. For any $k > 0$ there exists an explicit pseudorandom generator $G : \{0,1\}^r \to \{0,1\}^n$, where $r = c_{p,k,\epsilon} \cdot \log n$, such that for any depth k circuit $C \in \mathrm{CC}_0[p]$, the statistical distance between the two distributions $C(\{0,1\}^n)$ and $C(G(\{0,1\}^r))$ is at most ϵ.*

The proof of our main theorem is based on two new structural results for low degree polynomials, over finite fields, which may be of independent interest:

The first result is on the Fourier spectrum of such polynomials. Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a function. The α-Fourier coefficient of f, for $\alpha \in \mathbb{F}_p^n$, is defined as

$$\widehat{f}(\alpha) = \mathbb{E}_{x \in \mathbb{F}_p^n} \left[\omega^{f(x) - \langle x, \alpha \rangle} \right],$$

where $\omega = e^{2\pi i/p}$ is a primitive p-root of unity, and $\langle x, \alpha \rangle = \sum_{i=1}^n x_i \alpha_i$ is the inner product of x and α. The structural result we prove is that the Fourier coefficients of any low-degree polynomial cannot be spread over many disjoint sets. In other words, we show that one can always find a small set $S \subset [n]$ such that almost all Fourier coefficients intersect S (that is,

have some nonzero entry inside S). We note that while our main theorem is interesting only for odd p (as for $p = 2$ it reduces to the case of pseudorandom distributions), this structural result is non-trivial also for polynomials over $\mathbb{F}_2$.

Theorem (The Fourier spectrum of low-degree polynomials over finite fields)**.** *For every prime finite field $\mathbb{F}_p$, degree $d \geq 1$ and error $\epsilon > 0$ there exists a constant $C(d, \epsilon) \leq (1/\epsilon)^{O(4^d)}$ such that the following holds. Let $f(x_1, \ldots, x_n)$ be a degree d polynomial over $\mathbb{F}_p$. Then there exists a subset $S \subset [n]$ of size at most $|S| \leq C(d, \epsilon)$ such that*

$$\sum_{\alpha \in \mathbb{F}_p^n : \alpha \neq 0, \alpha_S = 0} |\hat{f}(\alpha)|^2 \leq \epsilon \, ,$$

where α_S is the restriction of α to coordinates in S. In words, almost all nonzero Fourier coefficients of f intersect S.

Our second structural result concerns the structure of polynomials with the following property. Denote with $\mathcal{U}_p$ the distribution over $\{0,1\}^n$ where each bit is chosen independently to be 0 with probability $1/p$ and 1 with probability $1-1/p$. We call $\mathcal{U}_p$ the *p-biased distribution*. We show that if the distributions $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ are ϵ-far, then f can be approximated, over $\{0,1\}^n$, by a function of a small number of lower degree polynomials. To formally state our theorem we need some definitions.

Definition (Bit-Rank)**.** *Let $g : \{0,1\}^n \to \mathbb{F}_p$ be a function. The d-bit-rank of g, denoted* $\mathrm{bitrank}_d(g)$*, is the minimal number of degree d polynomials over $\mathbb{F}_p$ required to compute g over $\{0,1\}^n$. That is,* $\mathrm{bitrank}_d(g) = k$ *where k is the minimal number such that there exist k degree d polynomials $f_1, \ldots, f_k : \mathbb{F}_p^n \to \mathbb{F}_p$ and a function $\Gamma : \mathbb{F}_p^k \to \mathbb{F}_p$ such that for all $x \in \{0,1\}^n$*

$$g(x) = \Gamma(f_1(x), \ldots, f_k(x)).$$

The definition of bit-rank differs from the more common definition of rank, which is the minimal number of degree d polynomials required to compute a function $g : \mathbb{F}_p^n \to \mathbb{F}_p$ over the entire input domain $\mathbb{F}_p^n$. For example, consider the function $g(x) = \sum_{i \neq j} x_i x_j$ over $\mathbb{F}_p$ for $p > 2$. We have that the 1-bit-rank of g is 1, as for all $x \in \{0,1\}^n$

$$g(x) = (x_1 + \ldots + x_n)^2 - (x_1^2 + \ldots + x_n^2) = (x_1 + \ldots + x_n)^2 - (x_1 + \ldots + x_n) \, .$$

Thus, for $x \in \{0,1\}^n$, $g(x)$ is determined by the linear function $\ell(x) = x_1 + \ldots + x_n$. Notice that as a quadratic polynomial over $\mathbb{F}_p$, the rank of g (i.e. the minimal number of linear functions required to compute g on inputs from $\mathbb{F}_p^n$) is either $n-1$ or n, depending on p.

Our second structural result is the following.

Theorem (Structure of bit-biased polynomials)**.** *Let $f(x_1, \ldots, x_n)$ be a degree d polynomial over $\mathbb{F}_p$ such that the statistical distance between the distributions $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ is at least ϵ. Then, for every $\delta > 0$, there exists a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that* $\Pr_{x \in \{0,1\}^n}[g(x) \neq f(x)] \leq \delta$ *and* $\mathrm{bitrank}_d(g) \leq p^{O(c)}$ *where $c = (p/\delta\epsilon)^{O(4^{(p-1)(d+1)})}$.*

The proofs of all theorems are given in the full paper, which can be found in Chapter 10.

2.3 Polynomials in coding theory

A code over a finite field $\mathbb{F}_p$ is a large subset $\mathcal{C} \subset \mathbb{F}_p^n$, with the guarantee that any two distinct codewords in $\mathcal{C}$ are "far apart"; that is, they have only a small number of coordinates where they agree. Formally, $\mathcal{C}$ is said to be an (n, k, d) code if $\mathcal{C} \subset \mathbb{F}_p^n$, $|\mathcal{C}| = p^k$ and for every two distinct codewords $c', c'' \in \mathcal{C}$ we have that the distance between them is at least d, where the distance between c' and c'' is defined as

$$\mathrm{dist}(c', c'') = |\{1 \leq i \leq n : c'_i \neq c''_i\}|.$$

The parameter n is called the *length* of the code; k is the *rate* of the code; and d is the *minimal distance* of the code. A code is said to be *linear* if the set $\mathcal{C}$ forms a linear space. Almost all codes studies are linear, and we will restrict our attention from now on only to linear codes. We note that the minimal distance in linear codes is equivalent to the *minimal weight* of a nonzero codeword, where the weight of a codeword $c \in \mathcal{C}$ is defined as

$$\mathrm{wt}(c) = |\{1 \leq i \leq n : c_i \neq 0\}|.$$

The main interest in coding theory is to find good codes, which simultaneously have good rate $k = \Omega(n)$ and good distance $d = \Omega(n)$ (when the base field $\mathbb{F}_p$ is large compared to n, even better distance can be obtained), and to find efficient decoding procedures for correcting words close to codewords to the correct codewords.

Polynomials form the basis for some of the more basic and intensively studied families of codes, namely Reed-Solomon, Reed-Muller and BCH codes. Reed-Solomon codes correspond to evaluations of univariate polynomials $f : \mathbb{F}_p \to \mathbb{F}_p$ of bounded degree; Reed-Muller codes correspond to evaluations of multivariate polynomials $f : \mathbb{F}_{2^n} \to \mathbb{F}_2$ of bounded total degree; Generalized Reed-Muller codes are extension of this to $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$; and BCH codes correspond to evaluations of univariate polynomials $f : \mathbb{F}_{p^n} \to \mathbb{F}_{p^n}$ of bounded degree.

The minimal distance of all the aforementioned codes is relatively well understood, but slightly more complicated properties are yet to be fully determined, despite the simplicity of the codes definitions. One of these properties is the *weight distribution* of the codes: this is the number of codewords of every prescribed weight. In the case of Reed-Muller codes this is equivalent to the following question: how many polynomials $f : \mathbb{F}_p^n \to \mathbb{F}_p$ of total degree at most d are there with a prescribed number of zeros. The weight distribution of Reed-Muller codes is fully understood for polynomials of degree 1 and 2, but is unsettled already for cubic polynomials. In a joint work with Tali Kaufman and Ely Porat [KLP10], we provide the first tight asymptotic estimation for this number when studying polynomials over $\mathbb{F}_2$. Prior to our work, this problem was settled only for small distances (up to 2.5 times the minimal distance of the code, by work of Azumi et al. [AKT76]). Furthermore, we extend this to obtain tight bounds also on the list-decoding size of Reed-Muller codes: this is the maximal number of codewords within a prescribed distance from some element $g \in \mathbb{F}_2^n$ (which is not necessarily a codeword). Combining this with previous results of Gopalan et al. [GKZ08] we get a list-decoding algorithm. We sketch this result in more details in Subsection 2.3.1.

A second work [Lov09] studies the weight distribution of Reed-Muller codes of constant degree but of unbounded number of variables. That is, we consider the set $A_p(d) = \{\Pr_{x \in \mathbb{F}_p^n}[f(x) = 0]\}$ where $f : \mathbb{F}_p^n \to \mathbb{F}_p$ ranges over all polynomials over $\mathbb{F}_p$ in n variables of

total degree at most d. We prove a somewhat surprising property about the zero probabilities of bounded degree polynomials: the set $A_p(d)$ is not dense; in fact all its limit points are of the form $\frac{\ell}{p^k}$ where $\ell, k \in \mathbb{N}$. We discuss this result in Subsection 2.3.2.

A third work with Tali Kaufman [KL19] studies affine invariant codes. These are codes $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ which are invariant under affine transformations. That is, if $f(x) \in \mathcal{C}$ then also $f(ax+b) \in \mathcal{C}$ for any $a \in \mathbb{F}_{p^n} \setminus \{0\}, b \in \mathbb{F}_{p^n}$. Understanding these families of codes turns out to be tightly related to understanding BCH codes and their duals. Previous work of Kaufman and Sudan [KS08] showed that any such code which is sparse (that is, where $\mathcal{C} = \text{poly}(n)$) is locally testable. We extend this result to exponentially large codes, that is, codes of size $|\mathcal{C}| \leq p^{p^{O(n)}}$. We sketch this in Subsection 2.3.3.

2.3.1 Weight Distribution and List-Decoding Size of Reed-Muller Codes

The weight distribution of an error correcting code counts, for every given weight parameter, the number of codewords with weight bounded by the given parameter. The weight distribution of a code is the main characteristic of the code, and governs the behavior of the code, from both theoretical and practical aspects.

Understanding the weight distribution of Reed-Muller codes is a 30-year-old standing open question in coding theory. The last progress on this question was made by Azumi, Kasami and Tokura [AKT76] that characterized the codewords of Reed-Muller codes of weight up to twice the minimal distance of the code, and hence obtained bounds for the weight distribution that apply till twice the minimal distance of the code. In this work we study the weight distribution of Reed Muller codes and provide asymptotically tight bounds that apply to all distances.

The problem of list-decoding an error correcting code is the following: given a received word and a distance parameter find all codewords of the code that are within the given distance from the received word. List-decoding is a generalization of the more common notion of unique decoding in which the given distance parameter ensures that there can be at most one codeword of the code that is within the given distance from the received word. The notion of list-decoding has numerous practical and theoretical implications. The breakthrough results in this field are due to Goldreich and Levin [GL89] and Sudan [Sud97] who gave efficient list decoding algorithms for the Hadamard code and the Reed-Solomon code. See surveys by Guruswami [Gur04] and Sudan [Sud00] for further details. In complexity, list-decodable codes are used to perform hardness amplification of functions [STV99]. In cryptography, list-decodable codes are used to construct hard-core predicates from one way functions [GL89]. In learning theory, list decoding of Hadamard codes implies learning parities with noise [KM93].

In this work we study the question of list-decoding Reed-Muller codes. Specifically, we are interested in bounding the list sizes obtained for different distance parameters for the list-decoding problem. Our work provides asymptotically tight bounds that apply to all distances. The improved bounds, imply improved algorithms for list-decoding Reed-Muller codes.

Our results are obtained by making a new connection between computer science tech-

niques used for studying low-degree polynomials and the discussed coding theory questions. Using this connection we manage to progress significantly towards resolving these two important open problems.

Our proofs are technically relatively simple. We view this as evidence to the importance of this new connection, since these were considered as open problems, resistent to the more common coding theory tools. We view this as the main innovation of our work.

Reed–Muller codes Reed-Muller codes are a fundamental and well studied family of codes. $\mathrm{RM}(n,d)$ is a linear code, whose codewords $f \in \mathrm{RM}(n,d) : \mathbb{F}_2^n \to \mathbb{F}_2$ are evaluations of polynomials in n variables of total degree at most d over $\mathbb{F}_2$. In this work we study the code $\mathrm{RM}(n,d)$ when $d \ll n$, and are interested in particular in the case of constant d.

The following facts regarding $\mathrm{RM}(n,d)$ are straight-forward: It has block length of 2^n, dimension $\sum_{i \le d} \binom{n}{i}$ and minimum relative distance $\frac{2^{n-d}}{2^n} = 2^{-d}$.

Weight distribution of Reed-Muller codes We now formally define the weight distribution of a code, and discuss previous known bounds for the weight distribution of Reed-Muller codes.

Definition (Relative weight)**.** *The relative weight of a function/codeword $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is the fraction of non-zero elements,*

$$\mathrm{wt}(f) = \frac{1}{2^n} |\{x \in \mathbb{F}_2^n : f(x) = 1\}|$$

Definition (Accumulative weight distribution)**.** *The accumulative weight distribution of $\mathrm{RM}(n,d)$ at a relative weight α is the number of codewords up to this weight, i.e.*

$$A(\alpha) = |\{p \in \mathrm{RM}(n,d) : \mathrm{wt}(p) \le \alpha\}|$$

where $0 \le \alpha \le 1$.

It is well-known that for any $p \in \mathrm{RM}(n,d)$ which is not identically zero, $\mathrm{wt}(p) \ge 2^{-d}$. Thus, $A(2^{-d} - \epsilon) = 1$ for any $\epsilon > 0$. Kasami and Tokura [KT70] characterized the codewords in $\mathrm{RM}(n,d)$ of weight up to twice the minimal distance of the code (i.e up to distance 2^{1-d}). Based on their characterization one could conclude the following.

Corollary (Corollary 10 in [GKZ08])**.**

$$A(2^{1-d} - \epsilon) \le (1/\epsilon)^{2(n+1)}$$

Corollary 11.1 and simple lower bounds (which we show later, see Lemma 11.5) show that $A(\alpha) = 2^{\Theta(n)}$ for $\alpha \in [2^{-d}, 2^{1-d} - \epsilon]$ for any $\epsilon > 0$ (and constant d).

List-decoding size of Reed-Muller codes We now formally define the list-decoding size of a code, and discuss previous known bounds for the list-decoding size of Reed-Muller codes. Moreover we discuss known list-decoding algorithms for Reed-Muller codes. We start with the following definition.

Definition (Relative distance between two functions)**.** *The relative distance between two functions* $f, g : \mathbb{F}_2^n \to \mathbb{F}_2$ *is defined as*

$$\mathrm{dist}(f, g) = \mathbb{P}_{x \in \mathbb{F}_2^n}[f(x) \neq g(x)]$$

Our work focuses on understanding the asymptotic growth of the list size in list-decoding of Reed-Muller codes, as a function of the distance parameter. Specifically we are interested in obtaining bounds on the following.

Definition (List-decoding size)**.** *For a function* $f : \mathbb{F}_2^n \to \mathbb{F}_2$ *let the ball at relative distance* α *around* f *be*

$$B(f, \alpha) = \{p \in \mathrm{RM}(n, d) : \mathrm{dist}(p, f) \leq \alpha\}$$

The list-decoding size of $\mathrm{RM}(n, d)$ *at distance* α*, denoted by* $L(\alpha)$*, is the maximal size of* $B(f, \alpha)$ *over all possible functions* f*, i.e.*

$$L(\alpha) = \max_{f : \mathbb{F}_2^n \to \mathbb{F}_2} |B(f, \alpha)|$$

In a recent work, Gopalan, Klivans and Zuckerman [GKZ08] proved that for distances up to the minimal distance of the code, the list-decoding size of Reed-Muller codes remains constant.

Theorem (Theorem 11 in [GKZ08])**.**

$$L(2^{-d} - \epsilon) \leq O\left((1/\epsilon)^{8d}\right)$$

Their result of bounding the list-decoding size of Reed-Muller codes is inherently limited to work up to the minimum distance of the code, since it uses the structural theorem of Kasami and Takura on Reed-Muller codes [KT70], which implies a bound on the weight distribution of Reed-Muller codes that works up to twice the minimum distance of the code.

Additionally, the work of [GKZ08] has developed a list-decoding algorithm for $\mathrm{RM}(n, d)$ whose running time is polynomial in the worst list-decoding size and in the block length of the code.

Theorem (Theorem 4 in [GKZ08])**.** *Given a distance parameter* α *and a received word* $R : \mathbb{F}_2^n \to \mathbb{F}_2$*, there is an algorithm that runs in time* $poly(2^n, L(\alpha))$ *and produces a list of all* $p \in \mathrm{RM}(n, d)$ *such that* $\mathrm{dist}(p, R) \leq \alpha$*.*

Since Gopalan et al. could obtain non-trivial bounds on the list-decoding size for distance parameter α that is bounded by the minimum distance of the Reed-Muller code, their algorithm running time could be analyzed only for α that is less than the minimum distance of the code. This supports our earlier statement, that the crux of the analysis of list-decoding algorithms is in bounding the list-decoding size.

Our Results The weight distribution of $\mathrm{RM}(n,d)$ codes beyond twice the minimum distance was widely open prior to our work. See e.g. Research Problem (15.1) in [MS83] and the related discussion in that chapter. In this work we provide asymptotic bounds for the weight distribution of $\mathrm{RM}(n,d)$ that applied for all weights $2^{-d} \leq \alpha \leq 1/2$. We state now our results for constant d, where the notation $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ hides constants depending only on d. Our first main result gives exact boundaries on the range of α for which $A(\alpha) = 2^{\Theta(n^\ell)}$, for any $\ell = 1, 2, ..., d$, showing there are "cut-off distances", at which the accumulative weight distribution jumps from $2^{\Theta(n^\ell)}$ to $2^{\Theta(n^{\ell+1})}$.

Theorem (First main theorem - accumulative weight distribution)**.** *Let $1 \leq \ell \leq d-1$ be an integer, and let $\epsilon > 0$. For any $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$*

$$2^{\Omega(n^\ell)} \leq A(\alpha) \leq (1/\epsilon)^{O(n^\ell)}$$

and $A(\alpha) = 2^{\Theta(n^d)}$ for any $\alpha \geq 1/2$.

We also address the more general problem of bounding the list-decoding size. Gopalan et al. [GKZ08] left as an open problem the question of bounding the list-decoding size of Reed-Muller codes beyond the minimal distance. We give tight bounds on the list-decoding size of Reed–Muller codes that apply to all distances. In fact, we show that the behavior of the list-decoding size is asymptotically identical to that of the accumulative weight distribution.

Theorem (Second main theorem - list-decoding size)**.** *Let $1 \leq \ell \leq d-1$ be an integer, and let $\epsilon > 0$. For any $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$*

$$2^{\Omega(n^\ell)} \leq L(\alpha) \leq (1/\epsilon)^{O(n^\ell)}$$

and $L(\alpha) = 2^{\Theta(n^d)}$ for any $\alpha \geq 1/2$.

Using our results combined with the previous results of Gopalan et al. [GKZ08], we obtain the following algorithmic result for list-decoding Reed-Muller codes.

Theorem (List-decoding algorithm)**.** *Let $R : \mathbb{F}_2^n \to \mathbb{F}_2$ be a received word. Let $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$ be a required distance parameter, where $1 \leq \ell \leq d-1$ is integer and $\epsilon > 0$. There exists an algorithm that runs in time $(1/\epsilon)^{O(n^\ell)}$ and produces a list of all $p \in \mathrm{RM}(n,d)$ such that $\mathrm{dist}(p,R) \leq \alpha$.*

The proofs of all the above theorems can be found in the full paper, which is provided in Chapter 11.

2.3.2 Holes in Reed-Muller codes

In this work we study the possible weights of codewords of Generalized Reed–Muller codes. For a prime power q, let $\mathbb{F}_q$ denote the field of q elements. The r^{th}-order Generalized Reed–Muller code over $\mathbb{F}_q$, denoted by $\mathrm{RM}_q(r,m)$, is a linear code over $\mathbb{F}_q$, whose codewords $f \in \mathrm{RM}_q(r,m) : \mathbb{F}_q^m \to \mathbb{F}_q$ are evaluations of polynomials over $\mathbb{F}_q$ in m variables of total

degree at most r. Reed–Muller codes correspond to the special case of $q = 2$. Both Reed–Muller codes and the more general family of Generalized Reed–Muller codes have attracted research for many years; to quote [MS83],

Reed–Muller (or RM) codes are one of the oldest and best understood families of codes.

In this work we study Generalized Reed–Muller codes $\mathrm{RM}_q(r, m)$, when the field $\mathbb{F}_q$ and order r are fixed, and the number of variables m tends to infinity. The basic property that we study is the relative weights of codewords $f \in \mathrm{RM}_q(r, m)$.

Definition (Relative weight)**.** *The* relative weight *of a codeword $f \in \mathrm{RM}_q(r, m)$ is the fraction of non-zero elements,*

$$\mathrm{wt}(f) = \frac{1}{q^m}|\{\mathbf{x} \in \mathbb{F}_q^m : f(\mathbf{x}) \neq 0\}|.$$

We denote by $A_q(r, m)$ the set of all weights of codewords $f \in \mathrm{RM}_q(r, m)$,

$$A_q(r, m) = \{\mathrm{wt}(f) : f \in \mathrm{RM}_q(r, m)\}.$$

There are two simple constraints on the values in $A_q(r, m)$. The first constraint relates to the the fact that the code is finite - since a relative weight is the fraction of inputs $\mathbf{x}$ for which $f(\mathbf{x}) \neq 0$, all values in $A_q(r, m)$ are rational of the form $\frac{\ell}{q^m}$. The second one relates to the minimal distance of the code.

Definition (Minimal distance)**.** *The* minimal relative distance *of a code $\mathcal{C}$ is the minimal weight of a non-zero codeword $f \in \mathcal{C}$.*

The minimal distance of Generalized Reed–Muller codes is well-known (see for example [MS83]).

Fact. *Let $r = (q-1)a + b$ where $0 \leq b \leq q-1$. The minimal relative distance of* $\mathrm{RM}_q(r, m)$ *is*

$$\delta_q(r) = \frac{1}{q^a}\left(1 - \frac{b}{q}\right).$$

We are interested in the set of possible weights of $A_q(r, m)$ for fixed q and r when $m \to \infty$. Clearly $A_q(r, m) \subset A_q(r, m')$ when $m < m'$. Thus it makes sense to look at the limit

$$A_q(r) = \bigcup_{m=1}^{\infty} A_q(r, m).$$

Our main object of study is the set $A_q(r)$. A priory, one would think that the set $A_q(r)$ is dense inside the permissible range, given by the minimal distance of the code. However, our main result shows that the truth is quite far from this. First we define q-rational numbers.

Definition (q-rational numbers)**.** *A rational number $\alpha \in [0, 1]$ is* q-rational *if it is of the form $\alpha = \frac{\ell}{q^k}$ for some integers ℓ, k.*

Note that if $q = p^t$ for a prime p, then q-rational numbers and p-rational numbers define the same set.

Theorem (Main theorem)**.** *Let $\alpha \in [0,1]$ be a number which is not q-rational. Then there exists some $\epsilon > 0$ such that $A_q(r)$ contains no value in the range $(\alpha-\epsilon, \alpha+\epsilon)$. Equivalently, there is no sequence of polynomials $f_1, f_2, \ldots$ over $\mathbb{F}_q$ of degree at most r, each possibly on a different number of variables, such that $\lim_{k\to\infty} \mathrm{wt}(f_k) = \alpha$.*

For example, there is no sequence of polynomials $f_1, f_2, \ldots$ over $\mathbb{F}_3$ of total degree at most 17, such that $\lim_{k\to\infty} \mathrm{wt}(f_k) = \frac{1}{2}$. The following is an immediate corollary of our main theorem.

Corollary. *Let $\overline{C}$ denote the closure of $C \subset [0,1]$. We have*

$$\bigcup_{r=1}^{\infty} \overline{A_q(r)} = \{\textit{the set of } q\textit{-rationals}\}$$

The proofs of all the above results can be found in the full paper, which is given in Chapter 12.

2.3.3 Testing of exponentially large codes, by a new extension to Weil bound for character sums

We study in this work families of locally testable codes. Let $\mathbb{F}_N = \mathbb{F}_{p^n}$ be a finite field, where we think of p as either constant or small. A code is a family of functions $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$. All codes we consider in this work are linear. The dimension of a code is $\dim(\mathcal{C}) = \log_p(|\mathcal{C}|)$.

A code is *locally testable* if there is a randomized algorithm, which when given as input a function $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$, probes f in a small number of locations and determines (with high probability) whether $f \in \mathcal{C}$ or f is far[1] from all codewords of $\mathcal{C}$. A code is q-locally testable if the number of probes is at most q, where q is sublinear in the code length, i.e. $q = o(N)$.

Most of the study of locally testable codes has been focused on codes testable with constant query complexity (i.e. $q = O(1)$) or with poly-logarithmic query complexity (i.e. $q = (\log N)^{O(1)}$). They appear as low-degree tests in the $IP = PSPACE$, $MIP = NEXP$ and $PCP = NP$ theorems, and indeed the work of [GS06] (which was later partly derandomized by [BSSVW03]) elucidates their role as the "combinatorial heart" of PCPs.

In general, there is a tradeoff between the rate of the code $\dim(\mathcal{C})/N$ and the query complexity of testing this code. A major open problem in this field is whether one can enjoy the best of both worlds: a code of constant rate which is locally testable with a constant query complexity.

One line of research focuses on constructing explicit codes which try to approach this optimal tradeoff. The best results to date are by Ben-Sasson and Sudan [BSS05] and Dinur [Din07] (see also Meir [Mei08]) which achieve an explicit binary code of rate $\frac{1}{(\log N)^{O(1)}}$ which is testable using a constant number of probes.

A second line of research focuses on characterization of general families of codes that are locally testable [BLR93, RS93, NAR03, JPRZ04, KR04, KS08, KS07, KL05, GKS09, KS10]. Many results in this field apply only to *sparse* codes over binary fields $\mathbb{F}_{2^n}$, which are codes

[1] If f has distance ϵ from $\mathcal{C}$, i.e. if $\min_{g\in\mathcal{C}} \Pr_{x\in\mathbb{F}_{p^n}}[f(x) \neq g(x)] = \epsilon$, we require the local test to reject f with probability at least $\Omega(\epsilon)$.

of dimension $O(\log N)$ [KL05, KS07, GKS09, KS10]. Another example is *Generalized Reed-Muller codes* which are the family of polynomials $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$ of total degree at most d. These codes are testable using $p^{\frac{d}{p-1}} = exp(d)$ queries, while having dimension $O(n^d)$ [NAR03, JPRZ04, KR04]. Such codes can be locally testable with sublinear number of queries for $d \leq O(\log n)$, which gives codes of quasi-logarithmic dimension $\dim(\mathcal{C}) \leq (\log N)^{\log\log N}$.

Our work falls into the latter line of research. We exhibit a general family of codes of almost optimal dimension $\dim(\mathcal{C}) = N^{\Omega(1)}$ which are locally testable with sublinear query complexity. We achieve this by studying *affine invariant codes*. A code $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ is affine invariant if it is invariant under affine transformation of the coordinates of input space. That is, if $f(x) \in \mathcal{C}$ then also $g(x) = f(ax+b) \in \mathcal{C}$ for any $a, b \in \mathbb{F}_{p^n}, a \neq 0$. Previous results [GKS09] showed that sparse affine invariant codes (i.e., codes of size $p^{O(n)}$) are locally testable. We significantly extend this to codes of up to exponential size, i.e. of size at most $p^{p^{\Omega(n)}}$.

Theorem (Main result)**.** *Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be a linear code which is affine invariant of dimension $\dim(\mathcal{C}) \leq p^{\alpha n}$, where $\alpha > 0$ is an absolute constant. Then $\mathcal{C}$ is locally testable with query complexity $q = poly(\dim(\mathcal{C})/n) = o(p^n)$. In particular, any sparse affine invariant code (i.e. with $\dim(\mathcal{C}) = O(n)$) is locally testable with constant query complexity $q = O(1)$. The parameter α can be chosen to be any $\alpha < 1/32$ for large enough n.*

The proof of the theorem can be found in the full paper, which is given in Chapter 13.

This generalizes previous works in several aspects: our result applies to codes of exponential size $exp(N^\alpha)$, while previous results apply only to codes of polynomial size $N^{O(1)}$ or quasi-polynomial size $exp(\log N^{\log\log N})$. Previous results on sparse codes applied only to binary fields $\mathbb{F}_{2^n}$, while our result applies to any field of small characteristic. Note that a recent result of Ben-Sasson and Sudan [BSS09, Sud10] shows that affine invariant codes that are testable with constant number of queries can not have exponential rate. Thus, our testing result of exponentially large codes can not be improved to testing with constant locality.

The main new ingredients in our work is a Fourier-analytic approach for estimating the weight distribution of affine invariant codes, and a new extension of the Weil bound for character sums of low-degree polynomials. We start by describing our new result for character sums for polynomials, and then discuss its relation to proving local testability of affine invariant codes. The proof of our new extension for the Weil bound relies on techniques borrowed from additive combinatorics. This demonstrates yet another connection between additive combinatorics and theoretical computer science. Such connections were used before to establish results regarding pseudorandom generators [BV07, Lov08, Vio08] and list-decoding of codes [KLP10].

2.4 Property testing for polynomials

Property testing is a field which studies when can properties of "large" objects can be revealed just by looking on small fractions of these objects. Natural objects studied in this framework are functions, graphs and hypergraphs. In general, a property is locally testable

is one can distinguish between objects having this property and objects which are "far" from having this property. In our setting, we consider the problem of property of locally testing for polynomials. That is, given a function $f : \mathbb{F}_p^n \to \mathbb{F}_p$, we wish to decide (while querying f only on a small number of evaluations) whether f is a low-degree polynomial, or is far from being a low-degree polynomials (in the sense that many evaluations of f must be changed in order to make f a low-degree polynomial). That is, we wish to query f only in a few locations, and still be able to distinguish degree-d polynomials from functions which are far from degree-d polynomials. Such questions, beside being natural in the setting of property testing, also arise in the area of PCP (Probabilistically Checkable Proofs) constructions, where they are known as "low degree tests", and they form the heart of many PCP constructions.

Being a low-degree polynomial, although being a global property of a function, can be verified locally. A function is a degree-d polynomial iff taking any $d+1$ derivatives of it makes the function vanish. This is true in the analytical setting, and also for polynomials over finite fields, where we use discrete directional derivatives.

The derivative of $f(x)$ in direction $y \in \mathbb{F}_p^n$ is defined as $f_y(x) = f(x+y) - f(x)$. It is easy to see that if f is a degree-d polynomial, then for any y, $f_y(x)$ is a polynomial in x of degree at most $d-1$. Define $f_{y_1,\ldots,y_k}$ to be the iterative derivative of f in directions $y_1, \ldots, y_k$. Then it is clear by the above argument, that if f is a degree-d polynomial, then $f_{y_1,\ldots,y_{d+1}} \equiv 0$ for any set of directions $y_1, \ldots, y_{d+1}$. In fact, the reverse implication is also true: if for any $y_1, \ldots, y_{d+1}$ we have that $f_{y_1,\ldots,y_{d+1}} \equiv 0$, then f must be a degree-d polynomial.

This raises a very natural test. To test if a function is close to degree-d polynomials, compute a random $d+1$ iterated derivative of it, and accept the function if the derivative is zero. We refer to this test as the *derivatives test*. Our previous discussion is equivalent to the statement: f is a degree-d polynomial iff the derivatives test always accept f. Alon et al. [AGHP90] showed that a robust version of this statement also holds - if f is accepted by the derivatives test with probability at least $1-\epsilon$ (where ϵ is small enough as a function of p and d), then the function has distance $O(\epsilon)$ to a unique degree-d polynomial (and in fact the function f can be decoded to this polynomial).

2.4.1 The Gowers norm

Very similar ideas, in a different context, were also studied by Gowers [Gow01] in his seminal work on the new proof for Szemerdi's theorem. He defined the (now known as) Gowers norm of a function, which is related to the acceptance probability of the derivatives test. Let $\omega_p = e^{\frac{2\pi i}{p}}$ be a p root of unity. The d-Gowers norm of $f : \mathbb{F}_p^n \to \mathbb{F}_p$ is defined as:

$$\|f\|_{U^d} = \left(\mathbb{E}_{x,y_1,\ldots,y_d \in \mathbb{F}_p^n} [\omega_p^{f_{y_1,\ldots,y_d}(x)}] \right)^{1/2^d}$$

Gowers has proved that $\|\cdot\|_{U^d}$ is indeed a norm. This proves that if a function f is somewhat close to a degree-$(d-1)$ polynomial, then the Gowers norm, or equivalently the derivatives test, accepts f with noticeable probability over a random function. It turns out that instead of studying distance of functions, it it more analytically convenient to study their correlation. The correlation of two functions $f, g : \mathbb{F}_p^n \to \mathbb{F}_p$ is defined as

$$correl(f,g) = \mathbb{E}_{x \in \mathbb{F}_p^n} [\omega_p^{f(x)-g(x)}]$$

Correlation and distance are equivalent up to a constant - two functions have distance $(1 - \frac{1}{p}) - \epsilon$, iff for some $c \in \mathbb{F}_p \setminus \{0\}$, $correl(cf, cg) \geq \Theta(\epsilon)$.

In this language, Gowers showed that if f has correlation ϵ with some degree-$(d-1)$ polynomial, then $\|f\|_{U^d} \geq \epsilon$ (equivalently, the d-derivatives test accepts f with probability at least $\frac{1}{p} + \Theta(\epsilon^{2^d})$). This is in contrast to a random function, which is inherently far from all degree-d polynomials, for which the Gowers norm is $o(1)$, and the acceptance probability of the derivatives test is $\frac{1}{p} + o(1)$.

Thus, a natural question is whether the opposite direction is true as well: If we know that $\|f\|_{U^d} \geq \epsilon$, does this tells us that f has ϵ' correlation with some degree-$(d-1)$ polynomial? This was conjectured to be true, and came to be known as the *Inverse Conjecture for the Gowers norm* (abbreviated ICGN).

Although the ICGN has application in Computer Science, as a possible candidate for testing correlation with low-degree polynomials, its main use is in the area of Additive Combinatorics, where this is the major tool allowing one to study linear progressions (and in general linear structures) in arbitrary sets. For example, it allowed Gowers [Gow01] to prove that sets with positive density in the integers must contain arithmetic progressions of any length.

For correlation with linear polynomials, i.e. $\|\cdot\|_{U^2}$, the ICGN is true, by the linearity testing result of Blum et al. [BLR93] and Bellare et al. [BCH+95]. In this case we get $\epsilon' = poly(\epsilon)$. For correlation with quadratic polynomials, i.e. $\|\cdot\|_{U^3}$, the ICGN is also true. This is by results of Green and Tao [GT08] and Samorodnitsky [Sam07]. The proof for this case is more involved, and the ϵ' obtained is weaker: $\epsilon' = exp(-1/\epsilon)$.

However, it turns out the ICGN is false for $\|\cdot\|_{U^4}$. This was obtained by Green and Tao [GT07] and independently by myself, Meshulam and Samorodnitsky [LMS08]. There is a simple example of a degree-4 polynomial, whose 4-Gowers norm is high (so if the ICGN was true, we would expect it to be close to cubic polynomials), but this polynomial has correlation at most $exp(-n)$ with all cubic polynomials. For more details we refer to the full paper which is given in Chapter 14.

A new version of the conjecture was recently proved by Bergelson et al. [BTZ09]. This new conjecture states that functions with noticeable d-Gowers norm is correlated to *non-classical polynomials* of degree $d-1$. A non-classical polynomial of degree $d-1$ is a function $g : \mathbb{F}_p^n \to \mathbb{Z}_{p^k}$, such that $g_{y_1,\ldots,y_d} \equiv 0$ for all directions (notice that in $g_y(x) = g(x+y) - g(x)$, the sum inside the brackets is taken in $\mathbb{F}_p$, while the difference outside is taken in $\mathbb{Z}_{p^k}$). This generalizes the definition of polynomials (which match the $k = 1$ case), and it turns out that there exist non-classical polynomials which are not polynomials. Bergelson et al. prove that if $\|f\|_{U^d} \geq \epsilon$, then f has correlation ϵ' with some non-classical polynomial g. The definition of correlation is generalized in the natural way to non-classical polynomials: for $f : \mathbb{F}_p^n \to \mathbb{F}_p$ and $g : \mathbb{F}_p^n \to \mathbb{Z}_{p^k}$,

$$correl(f, g) = \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega_p^{f(x)} \cdot \omega_{p^k}^{-g(x)}]$$

The caveat of the Bergelson et al. argument is that it is purely existential - for every constant ϵ there exists some constant ϵ', where they cannot determine an exact relation. This is because their proof is based on Ergodic Theory, which has strong tools to prove qualitative results, such as the one describe above, but is not equipped as such to give quantitative result.

2.4.2 Lower bounds for linearity testing

The simplest version of property testing for polynomials is testing whether a function is close to a linear function or not. This corresponds to the case of degree 1 polynomials. We study in this section general tests for linearity, which extend the basic derivatives test. The main question is what is the relation between the number of queries the test makes and the properties of the test.

We study the relation between the number of queries and soundness of adaptive linearity tests. A linearity test (over the field $\mathbb{F}_2$ for example) is a randomized algorithm which has oracle access to the truth table of a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$, and needs to distinguish between the following two extreme cases:

1. f is linear
2. f is far from linear functions

A function f is called *linear* if it can be written as $f(x_1, ..., x_n) = a_1x_1 + ... + a_nx_n$, with $a_1, ..., a_n \in \mathbb{F}_2$. The correlation of two functions $f, g : \mathbb{F}_2^n \to \mathbb{F}_2$ is defined as $correl(f, g) = |\mathbb{P}_{\mathbf{x}}[f(\mathbf{x}) = g(\mathbf{x})] - \mathbb{P}_{\mathbf{x}}[f(\mathbf{x}) \neq g(\mathbf{x})]|$. f is far from linear functions if it has small correlation with all linear functions.

Linearity tests were first introduced by Blum, Luby and Rubenfeld in [BLR93]. They presented the following test (coined the BLR test), which makes only 3 queries to f:

1. Choose $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$ at random
2. Verify that $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$.

Bellare et al. [BCH+95] gave a tight analysis of the BLR test. It is obvious that the BLR test always accepts a linear function. They have shown that if the test accepts a function f with probability $1/2 + \epsilon$, then f has correlation at least 2ϵ with some linear function.

For a linearity test, we define that it has *completeness* c if it accepts any linear function with probability of at least c. A test has *perfect completeness* if $c = 1$. A linearity test has *soundness* s if it accepts any function f with correlation at most ϵ with all linear functions, with probability of at most $s + \epsilon'$, where $\epsilon' \to 0$ when $\epsilon \to 0$. We define the *query complexity* q of a test as the maximal number of queries it performs. In the case of the BLR test, it has perfect completeness, soundness $s = 1/2$ (with $\epsilon' = 2\epsilon$) and query complexity $q = 3$.

If one repeats a linearity test with query complexity q and soundness s independently t times, the query complexity grows to $q' = qt$ while the soundness reduces to $s' = s^t$. So, it makes sense to define the *amortized query complexity* $\bar{q}$ of a test as $\bar{q} = q/\log_2(1/s)$. Independent repetition of a test doesn't change it's amortized query complexity. Notice that the BLR test has amortized query complexity $\bar{q} = 3$.

Linearity tests are a key ingredient in the PCP theorem, started in the works of Arora and Safra [AS98] and Arora, Lund, Motwani, Sudan and Szegedy [ALM+98]. In order to improve PCP constructions, linearity tests were studied in order to improve their amortized query complexity.

Samorodnitsky and Trevisan [ST00] have generalized the basic BLR linearity test. They introduced the *Complete Graph Test*. The Complete Graph Test (on k vertices) is:

1. Choose $\mathbf{x}_1, ..., \mathbf{x}_k \in \{0,1\}^n$ independently
2. Verify $f(\mathbf{x}_i + \mathbf{x}_j) = f(\mathbf{x}_i) + f(\mathbf{x}_j)$ for all i, j

This test has perfect completeness and query complexity $q = \binom{k}{2} + k$. They show that all the $\binom{k}{2}$ tests that the Complete Graph Test performs are essentially independent, i.e. that the test has soundness $s = 2^{-\binom{k}{2}}$. This makes this test have amortized query complexity $\bar{q} = 1 + \theta(1/\sqrt{q})$. They show that this test is optimal among the family of Hyper-Graph Tests (see [ST00] for definition of this family of linearity tests), and raise the question of whether the Complete Graph Test is optimal among all linearity tests, i.e. does a test with the same query complexity but with better soundness exist?

They partially answer this question in [ST06], where (among many other results) they show that no non-adaptive linearity test can perform better than the Complete Graph Test. A test is called *non-adaptive* if it first chooses q locations in the truth table of f, then queries them, and based on the results accept or rejects f. Otherwise, a test is called *adaptive*. An adaptive test may decide on its query locations based on the values of f in previous queries.

The proof technique of [ST06] uses the algebraic analysis of the Gowers Norm of certain functions. The Gowers Norm is a measure of local closeness of a function to a low degree polynomial. For more details regarding the definition and properties of the Gowers Norm, see [GT08] and [Sam07].

Ben-Sasson, Harsha and Raskhodnikova prove in [BSHR05] that any adaptive linearity test with completeness c, soundness s and query complexity q can be transformed into a non-adaptive linearity test with the same query complexity, perfect completeness and soundness $s' = s + 1 - c$. Combining their result with the result of [ST06] proves the lower bound also for adaptive linearity tests.

We also prove the same optimal lower bound for adaptive linearity test, but our proof technique is arguably simpler and more direct than the one used in [ST06]. We also study, like [ST06], the behavior of linearity tests on quadratic functions. However, instead of employing algebraic analysis of the Gowers Norm of certain functions, we provide a more direct combinatorial proof, studying the behavior of linearity tests on random quadratic functions. This proof technique also lets us prove directly the lower bound also for adaptive linearity tests. The result is given in details in the full paper, which can be found in Chapter 15.

Chapter 3

Open problems

Our work sheds some light on the properties of polynomials in various aspects: considered as a computational model, in coding theory and the structure for functions which are close to polynomials. Still, many important questions remain unsolved, most of them are currently far beyond the power of current techniques. We will explore various open problems which arise in these contexts, some of which may be more approachable than others.

3.1 Structure of biased polynomials

To better understand low-degree polynomials, it is interesting to understand how generic (i.e "random") polynomials behave. In particular, it would be interesting to find some distinction between random and non-random polynomials. We would like to have a separation of the form: either a polynomial "appears random", or it has some specific structure (and thus is non-random).

A natural measure for randomness of a polynomial $f(x) : \mathbb{F}_p^n \to \mathbb{F}_p$, is its distribution, i.e. the number of elements mapped to each image,

$$|\{x : f(x) = a\}|, \quad a \in \mathbb{F}_p$$

It is easy to see that most polynomials give rise to a uniform map: they map around $1/p$ of the elements to each $a \in \mathbb{F}_p$. A polynomial which fails to do so, is inherently not random. We would like to show that such polynomial must be highly structured. If we could show this, then any polynomial is either "random" (in the sense its image is close to uniform), or structured.

Moreover, if one succeeds to establish such a dichotomy, than in fact any polynomial f which can be approximated by a lower degree polynomial g, can in fact be represented as the sum of a lower degree polynomial, and a "structured" polynomial $f - g$ (this holds as the distribution of the difference $f - g$ must be non-uniform if g approximates f).

In the joint work with Tali Kaufman [KL08] we showed that qualitatively this is indeed the case. We show that if for a degree-d polynomial $f(x)$, the distribution $\{f(x) : x \in \mathbb{F}_p^n\}$ has distance at least ϵ from the uniform distribution (where distance is taken for example in statistical distance), then in fact there exist $c = c(\epsilon)$ degree-$(d-1)$ polynomials $g_1, \ldots, g_c$:

$\mathbb{F}_p^n \to \mathbb{F}_p$, and a combiner function $F : \mathbb{F}_p^c \to \mathbb{F}_p$, such that:

$$f(x) = F(g_1(x), \ldots, g_c(x))$$

This generalizes a previous work of Green and Tao [GT07], which was restricted to degrees bounded the field characteristic.

The value of c we (and also Green and Tao) obtain, increases rapidly with d. Its growth resembles that of the Ackerman function, iterated d times. In a subsequent work, Haramaty and Shpilka [HS10] showed this number can be reduced to $\log 1/\epsilon$ for $d = 3$ and $poly(1/\epsilon)$ for $d = 4$.

However, we have no example in which the number of lower degree polynomials required exceeds $O(\log 1/\epsilon)$. This raises the following natural question, which relates to the structure of polynomials which on one hand behave "non-random", but has no simple structural explanation for this.

Problem 3.1. *Fix a field $\mathbb{F}_p$ and a degree d. Is it true there exists a constant $C = C(p, d)$ such that the following holds: Every degree-d polynomial $f : \mathbb{F}_p^n \to \mathbb{F}_p$, whose distribution has distance of at least ϵ from the uniform distribution over $\mathbb{F}_p$, can be decomposed as a function of at most $C \cdot \log 1/\epsilon$ polynomials of degree at most $d - 1$.*

Answering this problem in either direction will improve dramatically our understanding of the relation between approximation and structure of low-degree polynomials.

3.2 Explicit functions which cannot be approximated by polynomials

One of the more intriguing questions is which functions cannot be computed or approximated by low-degree polynomials. Finding functions which cannot be approximated by low-degree polynomials $f : \mathbb{F}_p^n \to \mathbb{F}_p$, even of moderately large degree $d = \log^{O(1)} n$, have far reaching applications in complexity, as they are related to bounding the power of modular counting with regards to other basic computational operations.

Consider for simplicity the case of polynomials and functions over $\mathbb{F}_2$. The correlation of two functions $f, g : \mathbb{F}_2^n \to \mathbb{F}_2$ is defined as

$$correl(f, g) = \Pr[f = g] - \Pr[f \neq g]$$

, and the correlation of f with degree d polynomials is the maximal correlation of f with any degree d polynomial g. Approximation of functions by low degree polynomials is one of the main tools used in proving lower bounds for constant depth circuits. For example, Razborov and Smolensky [Raz87, Smo87] provided an explicit function MOD_3 that cannot be computed by a constant depth circuit with a subexponential number of AND, OR and XOR gates. The proof combines two arguments:

1. Any constant depth circuit of subexponential size has a very high correlation (that is, $1 - o(1)$) with some polynomial of degree n^ϵ;

2. Such a low degree polynomial has a correlation of at most $2/3$ with MOD_3. (In fact, this is true for any polynomial of degree at most $\epsilon\sqrt{n}$ for some constant ϵ.)

The best known constructions of explicit functions that cannot be approximated by low degree polynomials (see, e. g., [BSK08, BNS, Raz87, Smo87, VW08]) fall into two categories:

- For large degree bounds ($d < n^{\Omega(1)}$), there exists a symmetric function with a correlation of at most $O\left(1/\sqrt{n}\right)$ with degree $O\left(\sqrt{n}\right)$ polynomials;
- For small degree bounds ($d < \log n$) there are explicit functions having a correlation of at most $\exp(-n/c^d)$ with degree d polynomials for some constants c (best known is $c = 2$.)

Certain applications, e. g., pseudorandom generator constructions via the Nisan–Wigderson construction [NW94], require a function having an exponentially small correlation with low degree polynomials. This is only known for degrees up to $\log n$, while for larger degrees the best known bound is polynomial in n. Finding explicit functions with a better correlation is an ongoing quest with limited success. The above discussion raises the following natural problem:

Problem 3.2. *Find an explicit function $f : \mathbb{F}_2^n \to \mathbb{F}_2$, whose correlation with polynomials of degree $\log^{O(1)}(n)$ is exponentially small.*

In fact, this problem is unsolved even if one restricts itself to correlation below $n^{-1/2}$. For more details, see a survey by Viola [Vio09]. In the joint work with Ido Ben-Eliezer and Rani Hod [BEHL09] we showed that such a function f can have degree $\log^{O(1)}(n)$. It is not clear though how to find this polynomial explicitly.

Problem 3.3. *Construct using little randomness a degree-$(d+1)$ polynomial, which also has low correlation with all degree-d polynomials.*

3.3 Pseudorandom generators for polynomials

We are interested in explicitly constructing pseudorandom generators (PRG) against low degree polynomials over finite fields. A PRG for degree-d polynomials is a function $G : \{0,1\}^r \to \mathbb{F}_p^n$, such that for every degree-d polynomial $f(x_1, \ldots, x_n)$, the distribution of the outcome of f when applied to a uniform output of G is close to the distribution of the outcome of f when applied to a uniform element in $\mathbb{F}_p^n$. We are interested in PRG that are pseudorandom against all degree-d polynomials, and use as few random bits as possible.

The case of pseudorandom generators against *linear polynomials*, commonly referred to as *small-bias generators*, was first studied (over $\mathbb{F} = \mathbb{F}_2$) by Naor and Naor [**?**] and later by Alon et al. [AGHP90]. Them and others gave explicit constructions, which were later generalized for any finite field. These constructions have seed length which is up to a constant optimal. The construction of small-bias generators was a major tool in derandomization, PCPs constructions and lower bounds. See [BSSVW03] and the references within for more details regarding small-biased generators.

The generalization of the problem for constant degree polynomials was first studied by Luby, Veličković, and Wigderson [LVW93]. Their construction required $exp(O(\sqrt{\log n/\epsilon}))$ random bits for constant degree.

In large fields, a relatively good solution was given by Bogdanov[Bog05]. His PRG fooling degree-d polynomials has seed length polynomially close to the optimal, but works only if the field size is at least polynomial in the degree and log the number of variables. His construction is based on techniques in Algebraic Geometry.

In small fields (which is the important case for circuit complexity, and also the harder case), a sequence of works by Bogdanov and Viola [BV07], myself [Lov08] and Viola [Vio08] proved that the following is a PRG for degree d polynomials with error ϵ: the sum of d independent small-bias generators with error $\epsilon^{\exp(d)}$. This gives a generator which requires $\exp(d)\log n/\epsilon$ random bits in order to fool degree d polynomials. This is no better than trivially picking n elements which $d > \Omega(n)$. On the other hand, by the probabilistic method, there exists a non-explicit PRG for degree-d polynomials required only $O(d \log n/\epsilon)$ bits. This raises the following research question:

Problem 3.4. *Find a PRG for degree-d polynomials over small fields, for $d = \Omega(\log n)$, which has non-trivial seed length (i.e. less randomness than just choosing a random input in $\mathbb{F}_p^n$).*

The model of degree-d polynomials makes sense from the point of view of circuit complexity when $d < O(\log n)$, since such polynomials can be computed with polynomial size circuits (that is, they are the sum of a polynomial number of monomials). However, when the degree exceeds $\Omega(\log n)$ this is no longer true. Thus, it makes sense to study also the family of *sparse polynomials*, which are polynomials with at most a polynomial number of monomials. Note that contrary to the consideration of degree, the notion of sparsity is not invariant under a linear transformation of the input elements, and thus sparsity seems to be more related to the notion of succinctness of the representation of a function. For this reason, establishing results for such polynomials will be a major breakthrough. The following problem captures the essence of the problem.

Problem 3.5. *Find an efficient PRG for polynomials $f : \mathbb{F}_2^n \to \mathbb{F}_2$ of degree $\Omega(\log n)$ but with only a polynomial number of monomials.*

3.4 List decoding size for Generalized Reed-Muller codes

The Generalized Reed-Muller code $RM_p(n, d)$ encodes a degree-d polynomial $f : \mathbb{F}_p^n \to \mathbb{F}_p$ by the list of its evaluations $\left(f(x) : x \in \mathbb{F}_p^n\right)$. When $p = 2$, the codes are simply called Reed-Muller codes.

The most basic property of a codeword is its weight, defined as the number of non-zero elements in it. Understanding the weights of codewords turns to be important in analyzing the decoding properties of linear codes.

There are several properties of the weights of codewords which are commonly studied: the minimal weight of the code (the minimal weight of a non-zero codeword), the weight

distribution of the code (the weight of a uniformly chosen codeword), and divisibility properties (do all weights divide by some factor). Another notion, used mainly in Theoretical Computer Science, is that of list decoding (what is the maximal number of codewords in a ball or prescribed radius).

The properties Generalized Reed-Muller code, although being a relatively simple to define code, are not at all well understood. The minimal distance of Generalized Reed-Muller codes are known. However, the weight distribution is fully characterized only for $d = 1$ and $d = 2$.

For Reed-Muller codes (i.e. for $p = 2$), Azumi et al. [AKT76] classified the codewords of weight at most 2.5 times the minimal distance. The first result on the weight distribution for general distances was obtained in a joint work with Tali Kaufman and Ely Porat [KLP10], where we give nearly tight results for the weight distribution of Reed-Muller codes which apply for all distances.

List-decoding of Reed-Muller codes were studied by Gopalan et al. [GKZ08], who gave bounds on the list-decoding size when the radius is below the minimal distance of the code, and also efficient algorithms to find the codewords in this case. In the case of larger radii, they gave a reduction, showing that if one can give a bound on the list-decoding size, then an efficient algorithm to find the codewords can be obtained. In the joint work with Tali Kaufman and Ely Porat [KLP10] we also give nearly tight bounds for the list-decoding size of Reed-Muller codes, which apply to all radii.

The problem of list-decoding Generalized Reed-Muller codes is still open. In particular, it is conjectured that in balls of radius below the minimal distance of the code, the number of codewords is constant (i.e. independent of the number of variables). This was proved by Gopalan et al. [GKZ08] when $p - 1|d$, and by Gopalan [Gop09] for $d = 2$.

Problem 3.6. *Let ρ be the minimal distance of $RM_p(n, d)$. Show the list-decoding size of $RM_p(n, d)$ in radius $\rho - \epsilon$ is constant (i.e. depends only on p, d, ϵ, and is independent of n). That is, for every function $g : \mathbb{F}_p^n \to \mathbb{F}_p$ the number of degree d polynomials f for which $\Pr[f \neq g] \leq \rho - \epsilon$ is bounded by a constant independent of n.*

3.5 The inverse conjecture for the Gowers norm

Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be some function. We wish to know if f is close to a low-degree polynomial (where the distance between two functions is defined to be the fraction of inputs on which they differ). The goal is to decide this efficiently, without querying all the values of f, and still be able to get the correct answer with good probability. That is, we wish to query f only in a few locations, and still be able to distinguish functions somewhat close to degree-d polynomials, from functions which are not.

Such questions are natural in the area of property testing, where the general problems studied are deciding if an object has some global structure (as being close to a low-degree polynomial) by local views of the object (viewing only a few points of the function in our case). Testing if a function is close to low-degree polynomials is also important in many PCP constructions (although there the field is commonly chosen to be large).

Being a low-degree polynomial, although being a global property, can be verified locally. A function is a degree-d polynomial iff taking any $d + 1$ derivatives of it makes the function

vanish. This is true in the analytical setting, and also for polynomials over finite fields, where we use discrete directional derivatives.

The derivative of $f(x)$ in direction $y \in \mathbb{F}_p^n$ is defined as $f_y(x) = f(x+y) - f(x)$. It is easy to see that if f is a degree-d polynomial, then for any y, f_y is a polynomial of degree at most $d-1$. Define $f_{y_1,\ldots,y_k}$ to be the iterative derivative of f in directions $y_1, \ldots, y_k$. Then it is clear by the above argument, that if f is a degree-d polynomial, then $f_{y_1,\ldots,y_{d+1}} \equiv 0$ for any set of directions $y_1, \ldots, y_{d+1}$. In fact, the reverse implication is also true: if for any $y_1, \ldots, y_{d+1}$ we have that $f_{y_1,\ldots,y_{d+1}} \equiv 0$, then f must be a degree-d polynomial.

This raises a very natural test. To test if a function is close to degree-d polynomials, compute a random $d+1$ iterated derivative of it, and accept the function if the derivative is zero. We refer to this test as the *derivatives test*. Our previous discussion is equivalent to the statement: f is a degree-d polynomial iff the derivatives test always accept f. Alon et. al [AGHP90] showed that a robust version of this statement also holds - if f is accepted by the derivatives test with probability at least $1-\epsilon$ (where $\epsilon < \epsilon_{\max}(\mathbb{F}, d)$), then the function has distance $O(\epsilon)$ to a unique degree-d polynomial (and in fact the function f can be decoded to this polynomial).

Very similar ideas, in a different context, were also studied by Gowers [Gow01] in his seminal work on the new proof for Szemerdi's theorem. He defined the (now known as) Gowers norm of a function, which is related to the acceptance probability of the derivatives test. Let $\omega_p = e^{\frac{2\pi i}{p}}$ be a p root of unity. The d-Gowers norm of $f : \mathbb{F}_p^n \to \mathbb{F}_p$ is defined as:

$$\|f\|_{U^d} = \left(\mathbb{E}_{x,y_1,\ldots,y_d \in \mathbb{F}_p^n}[\omega_p^{f_{y_1,\ldots,y_d}(x)}]\right)^{1/2^d}$$

Gowers has proved that $\|\cdot\|_{U^d}$ is indeed a norm. This proves that if a function f is somewhat close to a degree-$(d-1)$ polynomial , then the Gowers norm, or equivalently the derivatives test, accepts f with noticeable probability over a random function. It turns out that instead of studying distance of functions, it it more analytically convenient to study their correlation. The correlation of two functions $f, g : \mathbb{F}_p^n \to \mathbb{F}_p$ is defined as

$$correl(f,g) = \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega_p^{f(x)-g(x)}]$$

Correlation and distance are equivalent up to a constant - two functions have distance $(1-\frac{1}{p})-\epsilon$, iff for some $c \in \mathbb{F}_p \setminus \{0\}$, $correl(cf, cg) \geq \Theta(\epsilon)$.

In this language, Gowers showed that if f has correlation ϵ with some degree-$(d-1)$ polynomial, then $\|f\|_{U^d} \geq \epsilon$ (equivalently, the d-derivatives test accepts f with probability at least $\frac{1}{p} + \Theta(\epsilon^{2^d})$). This is in contrast to a random function, which is inherently far from all degree-d polynomials, for which the Gowers norm is $o(1)$, and the acceptance probability of the derivatives test is $\frac{1}{p} + o(1)$.

Thus, a natural question is whether the opposite direction is true as well: If we know that $\|f\|_{U^d} \geq \epsilon$, does this tells us that f has ϵ' correlation with some degree-$(d-1)$ polynomial? This was conjectured to be true, and came to be known as the *Inverse Conjecture for the Gowers norm* (abbreviated ICGN).

Although the ICGN has application in Computer Science, as a possible candidate for testing correlation with low-degree polynomials, its main use is in the area of Additive

Combinatorics, where this is the major tool allowing one to study linear progressions (and in general linear structures) in arbitrary sets. For example, it allowed Gowers [Gow01] to prove that sets with positive density in the integers must contain arithmetic progressions of any length.

For correlation with linear polynomials, i.e. $\|\cdot\|_{U^2}$, the ICGN is true, by the linearity testing result of Blum et. al [BLR93] and Bellare et. al [BCH+95]. In this case we get $\epsilon' = poly(\epsilon)$. For correlation with quadratic polynomials, i.e. $\|\cdot\|_{U^3}$, the ICGN is also true. This is by results of Green and Tao [GT08] and Samorodnitsky [Sam07]. The proof for this case is more involved, and the ϵ' obtained is weaker: $\epsilon' = exp(-1/\epsilon)$.

However, it turns out the ICGN is false for $\|\cdot\|_{U^4}$. This was obtained by Green and Tao [GT07] and independently by myself, Meshulam and Samorodnitsky [LMS08]. There is a simple example of a degree-4 polynomial, whose 4-Gowers norm is high (so if the ICGN was true, we would expect it to be close to cubic polynomials), but this polynomial has correlation at most $exp(-n)$ with all cubic polynomials.

A new version of the conjecture was recently proved by Bergelson et. al [BTZ09]. This new conjecture states that functions with noticeable d-Gowers norm is correlated to *non-classical polynomials* of degree $d-1$. A non-classical polynomial of degree $d-1$ is a function $g : \mathbb{F}_p^n \to \mathbb{Z}_{p^k}$, such that $g_{y_1,\ldots,y_d} \equiv 0$ for all directions (notice that in $g_y(x) = g(x+y) - g(x)$, the sum inside the brackets is taken in $\mathbb{F}_p$, while the difference outside is taken in $\mathbb{Z}_{p^k}$). This generalizes the definition of polynomials (which match the $k = 1$ case), and it turns out that there exist non-classical polynomials which are not polynomials. Bergelson et. al prove that if $\|f\|_{U^d} \geq \epsilon$, then f has correlation ϵ' with some non-classical polynomial g. The definition of correlation is generalized in the natural way to non-classical polynomials: for $f : \mathbb{F}_p^n \to \mathbb{F}_p$ and $g : \mathbb{F}_p^n \to \mathbb{Z}_{p^k}$,

$$correl(f, g) = \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega_p^{f(x)} \cdot \omega_{p^k}^{-g(x)}]$$

The caveat of the Bergelson et. al argument is that it is purely existential - for every constant ϵ there exists some constant ϵ', where they cannot determine an exact relation. This is because their proof is based on Ergodic Theory, which has strong tools to prove qualitative results, such as the one describe above, but is not equipped as such to give quantitative result. This raises several research questions.

Problem 3.7. *Find a proof for the revised ICGN, which does not rely on Ergodic Theory. In particular, find some quantitative connection between ϵ' and ϵ.*

A first step in this direction, which may be easier, is

Problem 3.8. *Find a quantitative proof for the revised ICGN, where the function f itself is constrained to be a low-degree polynomial.*

Bergelson et. al result in fact tells us that the natural derivatives test is not a good test for testing if a function is mildly close to low-degree polynomials over finite fields, because it tests instead if a function is close to a low-degree non-classical polynomials. The first case where this difference emerges is when testing if a function over $\mathbb{F}_2$ is mildly close to cubic polynomials. Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function. We wish to test if f is $\frac{1}{2} - \epsilon$ close to cubic polynomials, while querying f in a small number of place.

Problem 3.9. *Let $\epsilon_1 > \epsilon_2 > 0$ be arbitrary. Does there exist a test, querying f at a constant number of indices (which may depend on ϵ_1, ϵ_2), which can differentiate function at least $\frac{1}{2} - \epsilon_1$ close to cubic polynomials, from functions at most $\frac{1}{2} - \epsilon_2$ close to cubic polynomials?*

Note the for linear and quadratic polynomials, the derivatives test uses a constant number of queries, independent of ϵ_1, ϵ_2.

Part II

Polynomials as a computational model

Chapter 4

Worst case to average case reductions for polynomials

A degree-d polynomial p in n variables over a field $\mathbb{F}$ is *equidistributed* if it takes on each of its $|\mathbb{F}|$ values close to equally often, and *biased* otherwise. We say that p has *low rank* if it can be expressed as a function of a small number of lower degree polynomials. Green and Tao [GT07] have shown that over large fields (i.e when $d < |\mathbb{F}|$) a biased polynomial must have low rank. They have also conjectured that bias implies low rank over general fields, but their proof technique fails to show that. In this work we affirmatively answer their conjecture. Using this result we obtain a general worst case to average case reductions for polynomials. That is, we show that a polynomial that can be *approximated* by a few polynomials of bounded degree (i.e. a polynomial with non negligible correlation with a function of few bounded degree polynomials), can be *computed* by a few polynomials of bounded degree. We derive some relations between our results to the construction of pseudorandom generators. Our work provides another evidence to the structure vs. randomness dichotomy.

Joint work with Tali Kaufman.

4.1 Introduction

Let $\mathbb{F}$ be a prime finite field. Let $p : \mathbb{F}^n \to \mathbb{F}$ be a polynomial in n variables over $\mathbb{F}$ of degree at most d. We say that p is *equidistributed* if it takes on each of its $|\mathbb{F}|$ values close to equally often, and *biased* otherwise. We say that p has a *low rank* if it can be expressed as a bounded combination of polynomials of lower degree, and *high rank* otherwise. More formally we consider the following definitions.

Definition 4.1 (bias)**.** *The* bias *of a function $f : \mathbb{F}^n \to \mathbb{F}$ is defined to be*

$$\mathrm{bias}(f) = \mathbb{E}_{X \in \mathbb{F}^n}[\omega^{f(X)}]$$

where ω stands for the $|\mathbb{F}|$ root of unity, i.e. $\omega = e^{\frac{2\pi i}{|\mathbb{F}|}}$.

We use the bias of f as a measure for the distance from uniformity of $f(X) \in \mathbb{F}$ when $X \in \mathbb{F}^n$ is chosen uniformly. The following simple facts explain why we can do so.

Fact 4.1. *Let $X \in \mathbb{F}^n$ be chosen uniformly. Then:*

- *If $f(X) \in \mathbb{F}$ is uniform then* $\text{bias}(f) = 0$
- *If* $\text{bias}(f) \geq \delta > 0$ *then the statistical distance between $f(X)$ and the uniform distribution over $\mathbb{F}$ is at least δ.*
- *If the statistical distance between $f(X)$ and the uniform distribution over $\mathbb{F}$ is δ, then there is some $c \in \mathbb{F}$, $c \neq 0$ s.t.* $\text{bias}(cf) \geq \delta'$ *for* $\delta' = \delta/\sqrt{|\mathbb{F}| - 1}$

Definition 4.2 (rank)**.** *Let $p(X)$ be a degree d polynomial over $\mathbb{F}^n$.* $\text{rank}_{d-1}(p)$ *is the smallest integer k such that there exist degree $d-1$ polynomials $q_1(X), ..., q_k(x)$, and a function $F : \mathbb{F}^k \to \mathbb{F}$, such that $p(X) = F(q_1(X), ..., q_k(X))$.*

Green and Tao [GT07] have shown that over large fields bias implies low rank.

Theorem 4.1 (Theorem 1.7 in [GT07])**.** *Let $p(X)$ be a degree d polynomial over $\mathbb{F}^n$, where $d < |\mathbb{F}|$. If* $\text{bias}(p) \geq \delta > 0$*, then* $\text{rank}_{d-1}(p) \leq c(\mathbb{F}, d, \delta)$.

In their paper, Green and Tao conjecture that the restriction $d < |\mathbb{F}|$ can be removed, but their proof technique breaks down when $d \geq |\mathbb{F}|$. Note that over large fields things might behave differently than over small fields. One important example is the *Inverse Conjecture for the Gowers Norm*. This conjecture roughly says that if the d-derivative of a polynomial is biased then that polynomial has a non-negligible correlation with some polynomial of degree $d-1$. The *Inverse Conjecture for the Gowers Norm* was proven to be true over large fields by [GT07], but was proven to be false over small fields [GT07, LMS08]. One of the main tools used for proving the conjecture over large fields was Theorem 4.1, that was proven over large fields.

One could ask what is the case with the above theorem, whether it remains true over smaller fields or it becomes false there. We show that the [GT07] result is true over general fields. In this respect, as opposed to the *Inverse Conjecture for the Gowers Norm* case,large and small fields behave similarly.

4.1.1 Our Main Results

Our first main theorem is a worst case to average case reduction for polynomials. It says that a polynomial that can be approximated by few polynomials of bounded degree, can be computed by few polynomials of bounded degree. We now move to define this rigorously.

Definition 4.3 (δ-approximation)**.** *We say a function $f : \mathbb{F}^n \to \mathbb{F}$ δ-approximates $p(X)$ if:*

$$|\mathbb{E}_{X \in \mathbb{F}^n}[\omega^{p(X)-f(X)}]| \geq \delta$$

Theorem 4.2 (Worst-case to average case reduction for polynomials of bounded degree)**.** *Let $p(X)$ be a polynomial of degree d, $g_1, ..., g_c$ polynomials of degree k (where d, c, k are constants) and $F : \mathbb{F}^c \to \mathbb{F}$ a function s.t. the composition $G(x) = F(g_1(X), ..., g_c(X))$ δ-approximates p. Then there exist c' polynomials $h_1, ..., h_{c'}$ and a function $F' : \mathbb{F}^{c'} \to \mathbb{F}$ s.t.*

$$F'(h_1(X), ..., h_{c'}(X)) \equiv p(X)$$

Moreover, $c' = c'(\mathbb{F}, d, c, k, \delta)$ (i.e. independent of n) and each h_i is of the form $p(X + a) - p(X)$ or $g_j(X + a)$ for $a \in \mathbb{F}^n$. In particular, if $k \leq d - 1$ then also $\deg(h_i) \leq d - 1$.

Our first main theorem is obtained as a corollary from our second main theorem, Theorem 4.3. This theorem shows that bias implies low rank over *general fields*.

Theorem 4.3 (Bias implies low rank for general fields)**.** *Let $p(X)$ be a degree d polynomial over $\mathbb{F}^n$, s.t. $\mathrm{bias}(p) \geq \delta > 0$. Then $\mathrm{rank}_{d-1}(p) \leq c(\mathbb{F}, d, \delta)$. That is, there exist degree-$(d-1)$ polynomials $q_1(X), ..., q_c(x)$, and a function $F : \mathbb{F}^c \to \mathbb{F}$, s.t. $p(X) = F(q_1(X), ..., q_c(X))$, and $c = c(\mathbb{F}, d, \delta)$. Moreover, $q_1, ..., q_c$ are derivatives of the form $p(X + a) - p(X)$ where $a \in \mathbb{F}^n$.*

Most of the technical part of the paper is dedicated to proving Theorem 4.3. The proof is by induction on the degree d of $p(X)$. Notice that for $d = 1$ it holds trivially. So, we assume Theorem 4.3 to hold for all degrees smaller than d, and prove it for degree d.

4.2 Significance of Results

Worst case to average case reductions for polynomials. Our first main theorem (Theorem 4.2) shows that every polynomial, not necessarily biased, that is approximated by few other bounded degree polynomials, can be computed by few bounded degree polynomials. We view this result as a worst case to average case reduction for polynomials. I.e. in order to show that a polynomial cannot be approximated by few bounded degree polynomials, it would be sufficient to show that the polynomial cannot be computed by few bounded degree polynomials. That later task might be easier. An example when such a scenario is relevant is the following. The papers [GT07, LMS08] that disprove the *Inverse Conjecture for the Gowers Norm* needed to show that the symmetric polynomial S_4 over $\mathbb{F}_2$, i.e. $S_4(x_1, ..., x_n) = \sum_{i<j<k<l} x_i x_j x_k x_l$ cannot be approximated by a degree 3 polynomial. Given the current result it could be sufficient (and maybe easier?) to show that S_4 cannot be computed by a constant number of degree 3 polynomials.

Proof of the Green Tao Conjecture. Our second main theorem (Theorem 4.3) shows that over general fields there is a phenomena that bias implies low rank. Green and Tao [GT07] proved this for large fields. They conjectured it to hold also over small fields. We answer their conjecture affirmatively, by showing that the "bias imply low rank" phenomena is robust and holds for all fields.

On the power of induction and relation to pseudorandom generators. Pseudorandom generator for polynomials of degree-d is an efficient procedure that stretches s field elements into $n \gg s$ field elements that can fool any polynomial of degree d in n variables. Pseudorandom generators are mostly interesting over small fields. One can use our second main theorem to provide an alternative proof to the correctness of the pseudorandom generators of [BV07] that fools degree d polynomials. Specifically, the generator of [BV07] is a XOR of d copies of the generator of Naor and Naor that fools linear functions. The proof of correctness of the [BV07] generator of [Vio08] is by induction. The proof assumes the existence of a pseudorandom generator that fools degree $d-1$ polynomial and constructs from it pseudorandom generator that fools degree d polynomial. The proof of the induction step is based on the following. Either the polynomial is unbiased, and hence the generator could fool it. Alternatively, it is biased, and hence again [Vio08] shows that it can be fooled. By our result here, if the polynomial is biased then it has low rank. One can use the property that a generator that can fool a function in the class can fool any composition of few functions from the class to complete the induction step. This proof method is inspired by the original argument of [BV07] the relied on the *Inverse Conjecture for the Gowers Norm* which turned out to be false. The proof of correctness of Viola [Vio08] is clearly more direct. However, we still feel that the original proof strategy of [BV07] sheds light on the relations between structure and pseudorandomness in the realm of low degree polynomials.

The "bias imply low rank" idea suggests a robust way to construct pseudorandom generators for some complex function classes based on pseudorandom generators for simpler function classes. This would be done in the spirit of the induction above. Either a function is unbiased , in which case it should be easy to claim that it could be fooled based on the induction assumption, or it is a function of few functions of lower complexity. Use now a property that a generator that can fool a function in the class can fool any composition of few functions from the class. Hence, by *induction* we obtain a construction of pseudorandom generator for functions of higher complexity classes (e.g. degree d polynomials) given pseudorandom generators for functions of lower complexity classes (e.g. linear functions).

Diakonikolas et al. [DLM$^+$07] suggest a general methodology to test whether a function on n variables has a concise representation. A tester is a randomized algorithm that should perform *few (even constant many)* queries into a given function such that the following holds. If the function has a concise representation it is accepted, and if it is far from having concise representation it is rejected. The idea of Diakonikolas et al. is to do testing by implicit learning. Their work provides property testers for several concise structures among them are s-sparse polynomials, size-s algebraic circuits and more.

Consider the following definition of concise representation for degree d polynomials. A polynomial of degree d has a concise representation if it is a function of few polynomials of lower degree (i.e. if it has a low rank). We argue that one can use the "bias imply low rank" theorem in order to construct a tester that test for this concise representation. The tester first performs a low degree testing e.g. by [RS93] to test that the given polynomial is of degree at most d, if the degree-tester rejects the tester rejects otherwise, the tester would approximate the bias of the polynomial. If the bias is large then by our first theorem the polynomial has low rank and the tester accepts, otherwise it rejects.

The above tester is robust in the following sense. It suggests a methodology for testing

concise representation of some monotone (with respect to complexity) families (e.g depth d circuits). The concise representation is with respect to the family. I.e. we would like to accept an object if it is in the family, and if it can be represented as a function of few other members of this family, which are of lower complexity. Testing of concise representation for monotone complexity classes could be done given a membership tester for that family, and given that the family obeys the "bias imply low rank" principle. If these two conditions are met, one can construct a tester. The tester first test membership in the family and then estimate the bias. If the bias is high the rank is low and concise representation exists.

Extension to tensors Let $L(x, y)$ be a bilinear form over $\mathbb{F}^n$, i.e. a function of the form

$$L(x, y) = x^t A y$$

where $x, y \in \mathbb{F}^n$ and A is a matrix. There is a close connection between the rank of the matrix and the bias of L. Dixon's Theorem ([MS83]) tells us that the bias of L (and in fact, all non-zero Fourier coefficients of L) has absolute value $c(\mathbb{F})^{-rank(A+A^t)}$. The theory of higher dimensional multilinear forms, i.e. tensors, is much less understood. In particular, there is no single notion of tensor rank. We prove, as a direct corollary of Theorem 4.3, that if we define the rank of a tensor as minimal number of lower degree multilinear forms needed to compute it, then bias imply low rank for tensors.

Theorem 4.4. *Let $L(X_1, ..., X_d)$ be a multilinear form of degree d s.t.* $\text{bias}(L) \geq \delta > 0$. *Then, there exist degree-$(d-1)$ multilinear forms $q_1, ..., q_c$, each operating on $d-1$ variables out of $X_1, ..., X_d$, and a function $F : \mathbb{F}^c \to \mathbb{F}$, s.t.*

$$\begin{aligned} L(X_1, ..., X_d) = F(&q_1(X_1, ..., X_{t_1-1}, X_{t_1+1}, ..., X_d), \\ &..., \\ &q_c(X_1, ..., X_{t_c-1}, X_{t_c+1}, ..., X_d)) \end{aligned}$$

and $c = c(\mathbb{F}, d, \delta)$. Moreover, $q_1, ..., q_c$ are derivatives of L.

Proof. We use Theorem 4.3 on L as a degree d-polynomial, and observe that derivatives of L are sums of d degree-$(d-1)$ multilinear forms in $d-1$ variables of $X_1, ..., X_d$. □

4.2.1 Proof Overview

We will prove that if a degree-d multivariate polynomial over a finite field can be approximated by a function of a constant number of lower degree polynomials, then it in fact be exactly computed by a function of a (larger) constant number of lower degree polynomials. Here and in the paper, constant means independent in the number of variables. In fact, we think of the number of variables as going to infinity, where the rest of the parameters (field size, degree, number of approximating polynomials) as constants. We denote by $p(X)$ a multivariate polynomial, where $X = (x_1, ..., x_n) \in \mathbb{F}^n$.

First we reduce the problem to showing that if a polynomial $p(X)$ is biased, then it can be computed by a function of constant number of lower degree polynomials. The reduction

is straightforward: if $p(X)$ can be approximated by a function $F(g_1(X), ..., g_k(X))$, where $deg(g_i) < deg(p)$ for all i, then there is some linear combination of the g_i's s.t. $p(X) + a_1 g_1(X) + ... + a_k g_k(X)$ is biased, and thus can be computed by a constant number of lower degree polynomials.

We now describe the proof of the main technical part of the paper, that is, if a degree d polynomial $p(X)$ is biased, then it can be calculated by a constant number of degree $d-1$ polynomials (the constant depending only on the field, the degree d, and the bias of p). The proof is by induction on d. We note that the case $d = 1$ is trivial.

Green and Tao prove the same result [GT07], when the degree d is bounded by the field size, $d < |\mathbb{F}|$. The main contribution of this work is extending this proof for all constant degrees. We will follow closely the proof structure of Green and Tao, and we make one significant divergence which allows us to make the result hold for all constant degrees.

The proof starts, as in the case of the work of Green and Tao, with a lemma of Bogdanov and Viola. Bogdanov and Viola [BV07] prove that if a degree-d polynomial $p(X)$ has bias, then it can be well approximated by a constant number of lower degree polynomials. Formally, for every constant $\epsilon > 0$, there is a function F_s and degree $d-1$ polynomials $b_1, ..., b_s$ s.t.

$$\Pr_{X \in \mathbb{F}^N}[p(X) = F_s(b_1(X), ..., b_s(X))] \geq 1 - \epsilon$$

where s depends only on the field $\mathbb{F}$, the degree d and the required approximation error ϵ. Importantly, s doesn't depend on the number of variables. Bogdanov and Viola in fact show an explicit construction of such a function F and polynomials $b_1, ..., b_s$.

The technical heart of this paper, as well as in the work of Green and Tao [GT07], is to show that when the approximation is good enough, it can in fact be made into an exact computation. Note that we can't use the lemma of Bogdanov and Viola directly, since choosing $\epsilon < |\mathbb{F}|^{-N}$ would result in a non-constant s.

Consider the following partition of $\mathbb{F}^n$ given by the joint distribution of the polynomials $(b_1, ..., b_s)$. For every $c = (c_1, ..., c_s) \in \mathbb{F}^s$, define the region

$$R_c = \{x \in \mathbb{F}^n : \forall i \; b_i(x) = c_i\}$$

The function F_s assigns a value to each region. We say that the joint distribution of $(b_1, ..., b_s)$ is *close to uniform*, if all the regions are roughly of the same size. That is, given $\gamma(s) > 0$, for every $c = (c_1, ..., c_s) \in \mathbb{F}^s$,

$$|R_c| = \frac{|\mathbb{F}|^n}{|\mathbb{F}|^s}(1 \pm \gamma(s)).$$

Green and Tao [GT07] show that a set of polynomials $(b_1, ..., b_s)$ that approximates p in the above sense, can be transformed into a larger set of polynomials called a *regular set* $(g_1, ..., g_t)$ that approximates p and such that the joint distribution of $(g_1, ..., g_t)$ is close to uniform, where t depends only on the field $\mathbb{F}$, the degree d and the required approximation error $\gamma(t)$.

Consider now the regions defined by the polynomials $(g_1, ..., g_t)$. Using averaging arguments the polynomial p is almost constant on most regions. We would like to show that in fact p is constant on all regions. We first show that if p is almost constant on a region, it

must be constant on all the region. We then extend this to all regions, assuming p is constant on most regions.

In order to show this, we first recall basic facts regarding derivatives. For a variable $Y \in \mathbb{F}^n$, we define the (discrete) derivative of $p(X)$ in direction Y to be $p_Y(X) = p(X+Y)-p(X)$. It is easy to see that the degree of X strictly reduces when taking derivatives. We define inductively taking multiple derivatives. For $Y_1, ..., Y_{d+1} \in \mathbb{F}^n$, consider the derivative of $p(X)$ in directions $Y_1, ..., Y_{d+1}$:

$$p_{Y_1,...,Y_{d+1}}(X) = \sum_{I \subseteq [d+1]} (-1)^{d+1-|I|} p(X + \sum_{i \in I} Y_i)$$

since p is a degree d polynomial, this derivative is identically zero. This will play an important role in the proof.

Let R_c be some region on which p is almost constant, and fix some $x_0 \in R_c$. Let $F_{|R_c}$ be the value that F assigns to that region. We will show that if $Y_1, ..., Y_{d+1}$ are chosen uniformly and independently, then there is a positive probability that $x_0 + \sum_{i \in I} Y_i \in R_c$ for all $I \subseteq [d+1]$. Moreover, since almost all points in $x' \in R_c$ are "good", i.e. $p(x') = F_{|R_c}$, there is in fact a positive probability that they all fall in the "good" part of R_c, i.e. that $p(x_0 + \sum_{i \in I} Y_i) = F_{|R_c}$ for all $I \neq \phi$. Plugging this into the derivative equation, and using the fact that it is identically zero, will give that also $p(x_0) = F_{|R_c}$. That is, if a region is almost constant, then it must be fully constant.

So, we need to prove that if $Y_1, ..., Y_{d+1}$ are chosen uniformly, there is a positive probability for all $x_0 + \sum_{i \in I} Y_i$ to fall in R_c and in fact to behave like a uniform point in R_c. In order to do so, we need to use the definition of the region R_c.

Consider the joint evaluation of all the polynomials $g_1, ..., g_t$ on all points $(x_0 + \sum_{i \in I} Y_i)$, i.e. the joint distribution in $\mathbb{F}^{(2^{d+1}-1)t}$ of:

$$\left(g_j(x_0 + \sum_{i \in I} Y_i) : j \in [t], I \subseteq [d+1], I \neq \phi \right)$$

where $Y_1, ..., Y_{d+1}$ are uniform and independent in $\mathbb{F}^n$. (Notice we disallow $I = \phi$, because it corresponds to the evaluations $\{g_j(x_0)\}$, which are fixed since they do not depend on any Y_i.)

If this distribution was uniform (over $\mathbb{F}^{(2^{d+1}-1)t}$), or even close enough to uniform, there was a positive probability that for all $j \in [t]$ and $I \subseteq [d+1]$,

$$g_j(x_0 + \sum_{i \in I} Y_i) = g_j(x_0)$$

Hence, all points $x_0 + \sum_{i \in I} Y_i$ would belong to R_c as required.

However, there is no reason why the joint distribution of $\{g_j(x_0 + \sum_{i \in I} Y_i)\}$ should be close to uniform. One obvious reason is that each polynomial g_j is itself a low degree polynomial, of degree at most $d-1$. Thus, for any $K \subseteq [d+1]$ s.t. $|K| > deg(g_j)$, deriving g_j in directions $\{Y_k : k \in K\}$ yields the zero polynomial, and thus we have the following linear relation:

$$\sum_{I \subseteq K} (-1)^{|K|-|I|} g_j(x_0 + \sum_{i \in I} Y_i) \equiv 0$$

Another reason for correlation is that different polynomials among $g_1, ..., g_t$ can be correlative. For example, we could have that $g_5 = g_1 g_2 + g_3 g_4$.

Green and Tao solve this problem by showing that if there are correlations between the polynomials, apart from the aforementioned linear relations, then using interpolation over $\mathbb{F}$ there must exist a linear functional over $a_1 g_1(X) + ... + a_t g_t(X)$ which is biased. This contradicts the fact, achieved in the construction of the g_i's, that the joint distribution of $(g_1(X), ..., g_t(X) : X \in \mathbb{F}^n)$ is extremely close to uniform. They then show that the linear relations can in fact be dealt with. However, their use of interpolation requires that $d < |\mathbb{F}|$.

We solve the problem in a different way, which allows us to make the result hold for all constant degrees. We transform our original set of polynomials $b_1, ..., b_s$ into a *strongly-regular* set of low degree polynomial $h_1, ..., h_t$, in which we can control all the correlations without using interpolation. The basic idea is that every h_j has an effective degree $\Delta(h_j) \leq deg(h_j)$, s.t. in the set

$$\{h_j(X + \sum_{i \in I} Y_i) : \ j \in [t], \ I \subseteq [d+1], \ |I| \leq \Delta(h_j)\}$$

there are no significant correlations, and any $h_k(X + \sum_{i \in K} Y_i)$ for $|K| > \Delta(h_k)$ can be calculated by a function of $\{h_j(X + \sum_{i \in I} Y_i) : \ j \in [t], \ I \subseteq K, \ |I| \leq \Delta(h_j)\}$.

This definition in fact allows us to prove several results showing that certain sets of evaluations are close to uniform, which are required for the proof.

4.2.2 Organization

The rest of the paper is organized as follows. We define required notation in Section 4.3. We define and analyze regularity and strongly regularity of polynomials in Section 4.4. We prove Theorem 4.2 and Theorem 4.3 in Section 4.5.

4.3 Preliminaries

$\mathbb{F}$ if a fixed prime field. We work with constant degree polynomials over $\mathbb{F}^n$. We denote by capital letters $X, Y, ...$ variables in $\mathbb{F}^n$, and by small letters $x, y, a, ...$ values in $\mathbb{F}^n$. We use the notation Pr for probability measure. Degree of a polynomial will always mean total degree. Unless otherwise specified, when we speak of a degree d polynomial, we mean in fact a polynomial of total degree at most d. For a set of variables $Y_1, Y_2, ... \in \mathbb{F}^n$ we denote by $Y_I = \sum_{i \in I} Y_i$, and similarly for a set of values $y_1, y_2, ... \in \mathbb{F}^n$. We write $u = v(1 \pm \epsilon)$ for $u \in [v(1-\epsilon), v(1+\epsilon)]$. When we speak of a *growth function*, we mean any monotone function $\mathcal{F} : \mathbb{N} \to \mathbb{N}$ (for example, $\mathcal{F}(n) = 2^{n^2}$). We shorthand the set $\{1, 2, ..., t\}$ by $[t]$.

Definition 4.4 (close to uniform)**.** *The joint distribution of the polynomials $(g_1, ..., g_s)$ is* γ-close to uniform/almost independent*, for $\gamma = \gamma(s) > 0$, if for every $(c_1, ..., c_s) \in \mathbb{F}^s$,*

$$\Pr_{X \in \mathbb{F}^n}(\forall i \in [c], g_i(X) = c_i) = (1 \pm \gamma(s)) \frac{|\mathbb{F}|^s}{|\mathbb{F}|^n}.$$

4.4 Regularity of polynomials

As we discussed in the introduction, the notion of regularity plays a major role in our proof. Green and Tao in [GT07] suggested one notion of regularity (we refer to it henceforth as *regularity*) which limited their proof to work only for large fields (i.e. $d < |\mathbb{F}|$). We suggest a stronger notion of regularity (noted henceforth as *strong regularity*). This new notion of strong regularity is essential for obtaining a result for general fields. In the following we review the regularity definitions given by Green and Tao. Then, we present the notion of strong regularity and show that every set of polynomials which approximates a polynomial p can be transformed into a larger set that approximates p and is also strongly regular. We end this section by showing that strong regularity implies almost independence for sets of variables that forms some specific structures. This almost independence is the crux of the proof of Theorem 4.3.

Definition 4.5 (Regularity of polynomials)**.** *Let $\mathcal{F}$ be any growth function. A set of polynomials $\{g_1, ..., g_m\}$ is called $\mathcal{F}$-regular if any linear combination $\alpha_1 g_1(X) + ...\alpha_m g_m(X)$ cannot be expressed as a function of at most $\mathcal{F}(m)$ polynomials of degree $k-1$, where $k = \max\{\deg(g_i) : \alpha_i \neq 0\}$ (i.e. k is the maximal degree of g_i appearing in the linear combination).*

Notice we use a growth function $\mathcal{F}(m)$ instead of a specific number. The reason is that in the application we would not be able to control the number m, and would only care about the relation between the number of polynomials (m) and the strength of the regularity of the set ($\mathcal{F}(m)$).

Green and Tao also define the notion of a refinement of a set of polynomials. Informally, a set $\{g_1, ..., g_m\}$ is a refinement of $\{f_1, ..., f_s\}$ if for any $i \in [s]$, $f_i(x)$ can be computed given the values of $\{g_1(x), ..., g_m(x)\}$.

Definition 4.6 (Refinement)**.** *A set of polynomials $\{g_1, ..., g_m\}$ is a refinement of $\{f_1, ..., f_s\}$ if for any $i \in [s]$ there exists a function $F_i : \mathbb{F}^m \to \mathbb{F}$ s.t.*

$$f_i(X) = F_i(g_1(X), ..., g_m(X))$$

Green and Tao prove that for any growth function $\mathcal{F}$, any set of polynomials $F = \{f_1, ..., f_s\}$ can be refined to a $\mathcal{F}$-regular set $\{g_1, .., g_m\}$, s.t. m depends only on s, $\mathcal{F}$ and the maximal degree in F. Importantly, m is independent of n.

We now discuss the way Green and Tao use the regularity condition, and why it fails to work when $d > |\mathbb{F}|$. We will then introduce our definition for strong regularity, which overcomes this obstacle.

As we discussed in the proof overview, if $\{g_1, ..., g_m\}$ are $\mathcal{F}$-regular for a large enough $\mathcal{F}$, then the joint distribution of

$$\{g_1(X), ..., g_m(X) : X \in \mathbb{F}^n\}$$

is close to uniform. Green and Tao need in fact a strong condition from the polynomials $g_1, ..., g_m$ in the process of their proof. Let $Y_1, ..., Y_{d+1} \in \mathbb{F}^n$ be new independent chunks of

variables. They require that for any $x_0 \in \mathbb{F}^n$,the joint distribution of

$$\{g_i(x_0 + \sum_{i \in I} Y_i) : |I| \leq deg(g_i)\}$$

is also close to uniform. They prove this is true if the field is large ($|\mathbb{F}| > d$). However, over small fields, this doesn't hold in general, as the following example shows.

Example 4.1. *Consider the symmetric polynomial S_4 over $\mathbb{F}_2$, i.e.*

$$S_4(x_1, ..., x_n) = \sum_{i<j<k<l} x_i x_j x_k x_l$$

Consider the fourth derivative of S_4, i.e. the polynomial in $X, Y_1, ..., Y_4$

$$G(X, Y_1, ..., Y_4) = \sum_{I \subseteq [4]} S_4(X + \sum_{i \in I} Y_i)$$

This polynomial corresponds to the 4-th Gowers Norm of S_4, and as was shown in [GT07] and [LMS08], it has bias $1/8$. Thus, the joint distribution of the set

$$\{S_4(x_0 + \sum_{i \in I} Y_i) : |I| \leq \deg(S_4)\}$$

is not close to uniform. This stands in contrast to the fact that $S_4(X)$ is equidistributed over $\mathbb{F}_2$.

Our definition for *strong-regularity* avoids this obstacles by allowing to effectively reduce the degree of a polynomial, if it's high-order derivatives can be calculated from lower-order ones. In fact, for any polynomial g_i we declare an effective degree $\Delta(g_i) \leq \deg(g_i)$. We require that the set

$$\{g_i(X + \sum_{i \in I} Y_i) : \ i \in [m], |I| \leq \Delta(g_i)\}$$

is almost uniform, while for every g_k and K s.t. $|K| > \Delta(g_k)$, $g_k(X + \sum_{iinK} Y_i)$ can be calculated by a function of $\{g_i(X + \sum_{i \in I} Y_i) : \ i \in [m], I \subseteq K, |I| \leq \Delta(g_i)\}$

We now move to formally define our notion of strong regularity, and to show it implies the almost independence/total dependence structure we have just described. We first define the notion of a derivative space.

Definition 4.7 (Derivative space)**.** *For a set of polynomials $F = \{f_1(X), ..., f_s(X)\}$ we define:*

$$Der(F) = \{f_i(X + a) - f_i(X) : i \in [s], \ a \in \mathbb{F}^n\}$$

Similarly, for a set of polynomials in several variable chunks $F = \{f_1(Y_1, ..., Y_k), ..., f_s(Y_1, ..., Y_k)\}$ ($Y_1, ..., Y_k \in \mathbb{F}^n$) we define:

$$Der(F) = \{f_i(Y_1 + a_1, ..., Y_k + a_k) - f_i(Y_1, ..., Y_k) : \\ i \in [s], \ a_1, ..., a_k \in \mathbb{F}^n\}$$

Notice that if the maximal degree of polynomials in F is k, then the maximal degree of polynomials in $Der(F)$ is at most $k-1$. We now formally define strong regularity. We recall that for a set of variables $Y_1, Y_2, ...$, we shorthand $Y_I = \sum_{i \in I} Y_i$.

Definition 4.8 (Strong regularity of polynomials)**.** *Let $\mathcal{F}$ be any growth function. Let $G = \{g_1, ..., g_m\}$ be a set of polynomials and $\Delta : G \to \mathbb{N}$ be a mapping from G to the natural numbers. We say the set G is strongly $\mathcal{F}$-regular with effective degree Δ if:*

1. *For any $i \in [m]$, $1 \leq \Delta(g_i) \leq \deg(g_i)$.*
2. *For any $i \in [m]$ and $r > \Delta(g_i)$, let X and $Y_1, Y_2, ..., Y_r$ be variables in $\mathbb{F}^n$. There exist a function $F_{i,r}$ s.t.*

$$g_i(X + Y_{[r]}) = \\ F_{i,r}\left(g_j(X + Y_J) : j \in [m],\ J \subseteq [r],\ |J| \leq \Delta(g_j)\right)$$

3. *For any $r \geq 0$, let X and $Y_1, ..., Y_r$ be variables in $\mathbb{F}^n$. Let $\{\alpha_{i,I}\}_{i \in [m], I \subseteq [r], |I| \leq \Delta(g_i)}$ be any collection of field elements, not all zero. Let $a(X, Y_1, ..., Y_r)$ stand for the linear combination:*

$$a(X, Y_1, ..., Y_r) = \sum_{i \in [m], I \subseteq [r], |I| \leq \Delta(g_i)} \alpha_{i,I} g_i(X + Y_I)$$

Let $G' \subseteq G$ be the set of all g_i's which appear in a, i.e.:

$$G' = \{g_i \in G : \exists I\ \alpha_{i,I} \neq 0\}$$

There <u>does not</u> exist polynomials $h_1, ..., h_l \in Der(G')$, $l \leq \mathcal{F}(m)$ s.t. $a(X, Y_1, ..., Y_r)$ can be expressed as:

$$H(h_1(X + Y_{I_1}), ..., h_l(X + Y_{I_l}))$$

for $I_1, ..., I_l \subseteq [r]$ and some function $H : \mathbb{F}^l \to \mathbb{F}$.

If the set G satisfies only (1) *and* (2)*, we say G is* pre-strong-regular *(notice that $\mathcal{F}$ appears only in* (3)*).*

We first prove, similar to the proof in [GT07], that any set of polynomials can be refined to a strong $\mathcal{F}$-regular set, where the size of the resulting set depends only on the size of the original set, and the maximal degree of polynomials in it. Also, the refining set is contained in the space of iterated derivatives of the original polynomials.

We now formally define the space of iterated derivatives.

Definition 4.9 (Space of iterated derivatives)**.** *For a polynomial set F, we define its iterated derivative set Der_C to be the set of taking at most C derivatives of F, i.e.*

$$Der_0(F) = F$$

$$Der_C(F) = Der(Der_{C-1}(F)) \cup Der_{C-1}(F)$$

Lemma 4.1 (Strong-Regularity Lemma)**.** *Let $\mathcal{F}$ be any growth function. Let $F = \{f_1, ..., f_s\}$ be a set of polynomials of maximal degree k. There exist a refinement $G = \{g_1, ..., g_m\}$ of F s.t.*

1. *The maximal degree of polynomials in G is also at most k*
2. *The set G is strong $\mathcal{F}$-regular.*
3. *The size m of G is a function of only $\mathcal{F}$, s and k. Importantly, it is independent of n.*
4. *There exists $C = C(\mathcal{F}, s, k)$ s.t. $G \subseteq Der_C(F)$*

Proof. We will start by defining a pre-strong-regular set G from F, and will keep refining it until we reach a strong $\mathcal{F}$-regular set. Our set G will also be in $Der_i(F)$ at the i-th iteration. We will finish by showing that the refinement process must end in a finite number of steps.

We start by defining $\Delta : F \to \mathbb{N}$ by $\Delta(f_i) = \deg(f_i)$, and set the initial value of G to be F. To show that the initial G is pre-strong-regular with effective degree Δ, observe that for any $r > \deg(f_i)$, deriving f_i r-times yields the zero polynomial. Thus, if $Y_1, ..., Y_r$ are variables, we have the identity:

$$f_i(X + Y_{[r]}) = \sum_{I \subsetneq [r]} (-1)^{r-|I|+1} f_i(X + Y_I)$$

Since we can do this for any $r > \deg(f_i)$, we can continue and express $f_i(X + Y_{[r]})$ as a linear combination of $\{f_i(X + Y_I) : I \subseteq [r],\ |I| \leq \deg(f_i)\}$. Thus, G is pre-strong-regular with effective degree Δ.

We will continue to refine G as long as it is not strong $\mathcal{F}$-regular. Assume $G = \{g_1, ..., g_m\}$ at some iteration is not strong-$\mathcal{F}$-regular. By definition, there is some $r \geq 0$ and coefficients $\{\alpha_{i,I}\}_{i \in [m], I \subseteq [r], |I| \leq \Delta(g_i)}$ s.t. the linear combination:

$$a(X, Y_1, ..., Y_r) = \sum_{i \in [m], I \subseteq [r], |I| \leq \Delta(g_i)} \alpha_{i,I} g_i(X + Y_I)$$

can be expressed as a function of $l \leq \mathcal{F}(m)$ polynomials $h_1, ..., h_l \in Der(G')$, where $G' = \{i \in [m] : \exists I\ \alpha_{i,I} \neq 0\}$ is the set of all g_i's participating in the linear combination.

Let g_{i_0} be a polynomial of maximal degree k in G' and let I_0 be a maximal I in respect to inclusion s.t. $\alpha_{i_0,I_0} \neq 0$. Notice that we must have that $|I_0| \leq \Delta(g_{i_0})$. We have:

$$\sum_{i \in [m], I \subseteq [r], |I| \leq \Delta(g_i)} \alpha_{i,I} g_i(X + Y_I) =$$
$$H(h_1(X + Y_{J_1}), ..., h_l(X + Y_{J_l}))$$

for some function $H : \mathbb{F}^l \to \mathbb{F}$.

Notice first that $\deg(h_i) \leq k - 1$ for all $i \in [l]$. Substitute in the expression $Y_i = 0$ for all $i \notin I_0$. We get that $g_{i_0}(X + Y_{I_0})$ can be expressed as a function of $\{g_{i_0}(X + Y_J) : J \subsetneq I_0\}$, $\{g_j(X + Y_J) : j \neq i_0,\ J \subseteq I_0,\ |J| \leq \Delta(g_j)\}$ and $\{h_j(X + Y_J) : J \subseteq I_0,\ |J| \leq \deg(h_j)\}$. Thus, if we add the polynomials $h_1, ..., h_l$ to G (and set $\Delta(h_i) = \deg(h_i)$), we can reduce $\Delta(g_{i_0})$ to

$|I_0| - 1$. If we reduced it to zero, we can remove g_{i_0} entirely from G. The resulting G will be our set for the next iteration.

In order to prove that the refinement process ends after a finite number of iterations (depending on the initial size of F and its maximal degree), notice that at each iteration, the sum of $\Delta(g_i)$ for all $g_i \in G$ with some degree d' reduces by at least 1, where the new polynomials added are all of degree strictly smaller than d', and their number is bounded (as a function of $\mathcal{F}$ and the size of G at the beginning of the iteration). So the total number of iterations is some Ackermann-like function of the initial number of polynomials, their maximal degree and the growth function $\mathcal{F}$. □

4.4.1 Almost independence by strong regularity

We continue by showing that strong regularity induces almost independence/total dependence structure over general sets of variables. The lemmas we derive are the main technical building blocks in the proof of Theorem 4.3.

We start with a lemma correlating applications of g_i on sums below the effective degree Δ to all sums over a set of variables.

Lemma 4.2. *Let $G = \{g_1, ..., g_m\}$ be a strong-regular set with effective degree Δ. Let $x, x' \in \mathbb{F}^n$ be two points s.t. $g_i(x) = g_i(x')$ for all $i \in [m]$. Let $y'_1, ..., y'_k \in \mathbb{F}^n$ be values for some $k \geq 1$, and let $Y_1, ..., Y_k \in \mathbb{F}^n$ be k random variables. Then the following two events are equivalent:*

1. *$A = [g_i(x + Y_I) = g_i(x' + y'_I)$ for all $i \in [m]$ and $I \subseteq [k]]$*
2. *$B = [g_i(x + Y_I) = g_i(x' + y'_I)$ for all $i \in [m]$ and $I \subseteq [k]$ s.t. $1 \leq |I| \leq \Delta(g_i)]$*

Proof. It is obvious that if A holds then also B holds. Assume that B holds, i.e. that

$$g_i(x + Y_I) = g_i(x' + y'_i)$$

for all $i \in [m]$ and $I \subseteq [k]$ s.t. $|I| \leq \Delta(g_i)$. Take some I s.t. $I > \Delta(g_i)$. We need to show that also $g_i(x + Y_I) = g_i(x' + y'_I)$. Since $|I| > \Delta(g_i)$ we know by the strong regularity of G that there is a function $F_{i,I}$ s.t.

$$g_i(X + Y_I) = \\ F_{i,I}\left(g_j(X + Y_J) : \ j \in [m], \ J \subseteq I, \ |J| \leq \Delta(g_j)\right)$$

By first substituting $X = x$ to compute $g(x + Y_I)$, and then substituting $X = x'$ and $Y_j = y'_j$ to compute $g(x' + y'_I)$, and using that both $g_j(x) = g_j(x')$ for all $j \in [m]$ and the assumption that B holds, we get that also $g_i(x + Y_I) = g_i(x' + y'_I)$. □

Our next lemma shows that certain evaluations of the polynomials $g_1, ..., g_m$ on linear combinations of the inputs are almost independent, assuming the linear combinations don't have too many non-zero entries. Remember that we are in the process of proving Theorem 4.3 for degree d by induction. Thus, we assume it to hold for all degrees $d' < d$, and in particular to all linear combinations of $g_1, ..., g_m$.

Lemma 4.3. *Let $\gamma = \gamma(m)$ be an error term. Let $Y_1, ..., Y_k \in \mathbb{F}^n$ be random variables for some $k \geq 1$. Assume $\mathcal{F}$ is large enough (as a function of γ and k). Assume $g_1, ..., g_m$ are strong $\mathcal{F}$-regular with effective degree Δ. For any non-empty $I \subseteq [k]$ let $x_I \in \mathbb{F}^n$ be some point, and $a^{(I)} = (a_1^{(I)}, ..., a_k^{(I)}) \in \mathbb{F}^k$ s.t.*

- $a_i^{(I)} \neq 0$ *for all* $i \in I$
- $a_i^{(I)} = 0$ *for all* $i \notin I$

Then the joint distribution of

$$\left(g_i(x_I + \sum_{i \in I} a_i^{(I)} Y_i) : i \in [m], I \subseteq [k], 1 \leq |I| \leq \Delta(g_i) \right)$$

is γ-close to the uniform distribution on $\mathbb{F}^{\sum_{i=1}^m \sum_{j=1}^{\Delta(g_i)} \binom{k}{j}}$.

We need the following simple lemma for the proof of Lemma 4.3. It states that a random derivative of a biased polynomial is also biased.

Lemma 4.4. *Let $h(Y_1, ..., Y_k)$ be a polynomial with bias δ. Let h' be the derivation of h in variables $Y_1, ..., Y_r$ along the directions $Z_1, ..., Z_r$, $(r \leq k)$ i.e.*

$$h'(Y_1, ..., Y_k, Z_1, ..., Z_r) = \sum_{w \in \{0,1\}^r} (-1)^{|w|} h(Y_1 + w_1 Z_1, ..., Y_r + w_r Z_r, Y_{r+1}, ..., Y_k)$$

where $|w|$ denotes the Hamming weight of w. Then $\text{bias}(h') \geq \delta^{2^r}$.

Proof. We apply Cauchy-Schwarz. It's enough to prove for $k = 2$ and $r = 1$ because we can group variables.

$$\begin{aligned} \text{bias}(h') =& \mathbb{E}_{Y_1, Y_2, Z_1 \in \mathbb{F}^n} [\omega^{h(Y_1, Y_2) - h(Y_1 + Z_1, Y_2)}] = \\ & \mathbb{E}_{Y_2 \in \mathbb{F}^n} [\left(\mathbb{E}_{Y_1 \in \mathbb{F}^n} [\omega^{h(Y_1, Y_2)}] \right)^2] \geq \\ & \left(\mathbb{E}_{Y_1, Y_2 \in \mathbb{F}^n} [\omega^{h(Y_1, Y_2)}] \right)^2 = \delta^2 \end{aligned}$$

□

Proof. (of Lemma 4.3) We start by using the well known fact, that if a distribution over $\mathbb{F}^r$ is not uniform, it must have some biased functional. If the distribution we study is γ-far from uniform, then there must be a linear functional on $\{g_i(x_I + \sum_{i \in I} a_i^{(I)} Y_i) : i \in [m], I \subseteq [k], |I| \leq \Delta(g_i)\}$ with some non-negligible bias depending on γ. We will prove that if we assume that, we reach a contradiction.

Denote by $Y'_I = \sum_{i \in I} a_i^{(I)} Y_i$, and notice it depends on exactly the same set of variables from $Y_1, ..., Y_k$ as Y_I. By our assumption, there exist coefficients $\{\alpha_{i,I}\}$, not all zero, s.t. the polynomial

$$h(Y_1, ..., Y_k) = \sum_{i \in [m], I \subseteq [k], |I| \leq \Delta(g_i)} \alpha_{i,I} g_i(x_I + Y'_I)$$

has bias at least ρ, where ρ is a function of γ, k and m only (and not of n).

Fix I_0 maximal with regards to inclusion s.t. not all α_{i,I_0} are zero. Since we just care about the bias of h under random $Y_1, ..., Y_k$, we can multiply each Y_i by some non-zero coefficient. We thus assume w.l.o.g that $a_i^{(I_0)} = 1$ for all $i \in I_0$. Let $|I_0| = r$. We assume w.l.o.g that $I_0 = \{1, 2, ..., r\}$. Notice that $Y'_{[r]} = Y_{[r]}$. We also shorthand $x = x_{[r]}$.

Let g_{i_0} be a polynomial with maximal degree $d'' \leq d' < d$ s.t. $\alpha_{i_0,I_0} \neq 0$.

We derive now once each of the variables in $Y_1, ..., Y_r$. Let $\{Z_i\}_{i=1..r}$ be new variables in $\mathbb{F}^n$,

and consider:

$$h'(Y_1, ..., Y_k, Z_1, ..., Z_r) = \sum_{w \in \{0,1\}^r} (-1)^{|w|} h(Y_1 + w_1 Z_1, ..., Y_r + w_r Z_r, Y_{r+1}, ..., Y_k)$$

First, by Lemma 4.4, h' has bias at least $\rho' = \rho^{2^k}$.

Now, consider what happens to a term $g_i(x + Y'_I)$ in h after the derivation. If $I \neq [r]$, by the maximality of I_0 there must exist $i' \in [r]$ s.t. $i' \notin I$. Thus, deriving $Y_{i'}$ zeroes out $g_i(x + Y'_I)$.

So, the only terms remaining in h' come from terms in h of the form $g_i(x + Y_{[r]})$. Thus, h' does not depend on Y_i for $i \notin [r]$, and also all the g_i's remaining must have $\Delta(g_i) \geq r$ (because $g_i(x + Y_{[r]})$ appeared in h with non-zero coefficient). Thus we can write:

$$h' = h'(Y_1, ..., Y_r, Z_1, ..., Z_r) = \sum_{i \in [m]} \alpha_{i,[r]} \sum_{w \subseteq [r]} (-1)^{|w|} g_i(x + Y_{[r]} + Z_w)$$

We now make an important observation. Notice that h' depends only on the sum $Y_{[r]}$, and not on the individual $Y_1, ..., Y_r$. So we can substitute $W = x + Y_{[r]}$ and get:

$$h' = h'(W, Z_1, ..., Z_r) = \sum_{i \in [m]} \alpha_{i,[r]} \sum_{w \subseteq [r]} (-1)^{|w|} g_i(W + Z_w)$$

We have assumed that G is strong $\mathcal{F}$-regular. We will show now that if we choose $\mathcal{F}$ large enough, we have already reached a contradiction. Notice the polynomials $g_i(W + Z_w)$ are exactly those which appear in the regularity requirements (where X is replaced here by W, and $Y_1, Y_2, ...$ by $Z_1, Z_2, ...$). Let G' denote the set of g_i's s.t. g_i appear in h' with non-zero coefficient.

We assume by induction that Theorem 4.3 holds for $d'' < d$ and for all n. Since all polynomials $g_i \in G$ have degree at most $d-1$, then also $\deg(h') \leq d-1$, and so we can apply Theorem 4.3 on h'. So, since h' has bias ρ', there must exist polynomials $q_1, ..., q_t \in Der(h')$ s.t.

$$h'(W, Z_1, ..., Z_r) = Q(q_1(W, Z_1, ..., Z_r), ..., q_t(W, Z_1, ..., Z_r))$$

for some function $Q : \mathbb{F}^t \to \mathbb{F}$, s.t. $t = t(\rho', d'')$. Moreover, since every polynomial q_i is of the form $h'(W + a_0, Z_1 + a_1, ..., Z_r + a_r) - h'(W, Z_1, ..., Z_r)$ for some constants $a_0, ..., a_r \in \mathbb{F}^n$, and h' is the sum of $g_i(W + Z_w)$, we can decompose each q_i to a sum of at most 2^r polynomials of the form $g_i(W + Z_w + a) - g_i(W + Z_w) \in Der(G')$ for $w \subseteq \{0, 1\}^r$. Let $q'_1, ..., q'_{t'} \in Der(G')$ denote these decomposed polynomials. We thus have that:

$$h'(W, Z_1, ..., Z_r) = Q'(q'_1(W + Z_{I'_1}), ..., q'_{t'}(W + Z_{I'_{t'}}))$$

for some function $Q' : \mathbb{F}^{t'} \to \mathbb{F}$, $t' = 2^r t$ and $I'_1, ..., I'_t \subseteq [r]$. We got that we can compute

$$h'(W, Z_1, ..., Z_r) = \sum_{i \in [m]} \alpha_{i,[r]} \sum_{w \subseteq [r]} (-1)^{|w|} g_i(W + Z_w)$$

as a function of t' polynomials of degree strictly smaller than d''. If we have $\mathcal{F}(m) > t'$ this is a contradiction to the strong $\mathcal{F}$-regularity of $g_1, ..., g_m$.

Summarizing, there can be no linear combination of $\{g_i(x + Y_I) : I \in S, 1 \leq |I| \leq \Delta(g_i)\}$ which has bias more than ρ, and so the distribution is γ-close to uniform. □

A Useful corollary of Lemma 4.3 and Lemma 4.2 is the following.

Corollary 4.1. *Let $x, x' \in \mathbb{F}^n$ be two points s.t. $g_i(x) = g_i(x')$ for all $i \in [m]$. Let $y'_1, ..., y'_k \in \mathbb{F}^n$ be values for some $k \geq 1$, and let $Y_1, ..., Y_k \in \mathbb{F}^n$ be k random variables. Then*

$$\Pr\left[g_i(x + Y_I) = g_i(x' + y'_I) \ \forall \ i \in [m], \ I \subseteq [k]\right] =$$
$$|\mathbb{F}|^{-\sum_{i=1}^{m} \sum_{j=1}^{\Delta(g_i)} \binom{k}{j}} (1 \pm \gamma)$$

4.5 From approximation to computation: Proof of Theorems 4.2 and 4.3

In this section we prove Theorem 4.2 and Theorem 4.3. We start with the proof of Theorem 4.2 which follows directly from Theorem 4.3. Assume $F(g_1(X), ..., g_c(X))$ δ-approximates $p(X)$. Develop $\omega^{F(z_1,...,z_c)} : \mathbb{F}^c \to \mathbb{C}$ in the Fourier basis. If $F(g_1(X), ..., g_c(X))$ δ-approximates $p(X)$, there must exist some Fourier coefficient which δ'-approximates p , where $\delta' \geq \delta |\mathbb{F}|^{-c}$. That means, there exist $\alpha_1, ..., \alpha_c \in \mathbb{F}$ s.t. the polynomial

$$p'(x) = p(x) - (\alpha_1 g_1(x) + ... + \alpha_c g_c(x))$$

has bias at least δ'. Using Theorem 4.3 we get that there must exist at most c' derivatives of p' which computes it exactly. We can now use them and $\alpha_1 g_1 + ... \alpha_c g_c$ to compute p.

In the remaining of this section we prove Theorem 4.3. The proof will be by induction on the degree d of the polynomial (notice that for $d = 1$ Theorem 4.3 is trivial). Let $p(X)$ stand for a degree d polynomial with bias δ. The proof starts by a lemma of Bogdanov and Viola [BV07], showing that if a polynomial is biased, then it can be well approximated by a function a small number of degree $d - 1$ polynomials. This was also the starting point in the work of Green and Tao:

Lemma 4.5 (Bias imply approximation by few lower degree polynomials)**.** *Let $p(X)$ be a polynomial of degree d with bias δ. For any $\epsilon > 0$ there exist polynomials $f_1(X), ..., f_s(X)$ of degree at most $d-1$ and a function $F : \mathbb{F}^s \to \mathbb{F}$ s.t.*

$$\Pr_{X \in \mathbb{F}^n}[F(f_1(X), ..., f_s(X)) \neq p(X)] < \epsilon$$

The number s of the polynomials depends only on δ and ϵ. Moreover, $f_1, ..., f_s \in Der(p)$.

The following lemma is the technical heart of the paper.

Lemma 4.6 (Approximation by few lower degree polynomials imply computation by few lower degree polynomials)**.** *Let $p(X)$ be a polynomial of degree d, $f_1, ..., f_s$ polynomials of degree $d-1$, ($s = O(1)$) and $H : \mathbb{F}^s \to \mathbb{F}$ a function s.t. the composition $H(f_1(X), ..., f_s(X))$ ϵ_d-approximates p, where $\epsilon_d = 2^{-\Omega(d)}$ Then there exist s' polynomials $f'_1, ..., f'_{s'}$ and a function $H' : \mathbb{F}^{s'} \to \mathbb{F}$ s.t.*

$$H'(f'_1(X), ..., f'_{s'}(X)) \equiv p(X)$$

Moreover, $s' = s'(d, s)$ (i.e. independent of n) and each f'_i if of the form $p(X + a) - p(X)$ or $f_j(X + a)$ for $a \in \mathbb{F}^n$.

Thus, to complete the proof of Theorem 4.3, it remains to prove Lemma 4.6.

In the following we prove Lemma 4.6. The main technical tool that we will use are Lemmas 4.2 and 4.3. We start the proof of Lemma 4.6 by refining $F = \{f_1, ..., f_s\}$ to a strong-regular set. Let $\mathcal{F}$ be a large enough growth function (to be determined later). By Lemma 4.1 there exists a set $G = \{g_1, ..., g_m\}$ refining F, and an effective degree Δ, s.t. G is strong $\mathcal{F}$-regular with effective degree Δ. Moreover, there exists a $C = C(\mathcal{F}, d, \delta)$ s.t. $G \subseteq Der_C(F)$. We know that G also approximates $p(X)$ at least as well as F does. We will prove that it is in fact computes F completely. We can then decompose each $g_i \in Der_C(F)$ as a sum of at most 2^C elements in $Der(p)$ to conclude the result.

Thus, we need to show that G in fact computes $p(X)$ completely. For $c = (c_1, ..., c_m) \in \mathbb{F}^m$, denote by $R_c \subseteq \mathbb{F}^n$ the region

$$R_c = \{x \in \mathbb{F}^n : \forall i \; g_i(x) = c_i\}$$

To show that G computes $p(X)$ is equivalent to showing that $p(X)$ is constant on any region R_c. Thus, we turn to study the regions R_c.

We first show (Lemma 4.7) that all regions R_c have about the same volume, i.e. that they form an almost uniform division of $\mathbb{F}^n$ to $\mathbb{F}^m$ regions. Since G is a strong regular refitment of F that ϵ_d-approximates p we know that also G ϵ_d-approximates p, i.e. there exists some $H' : \mathbb{F}^m \to \mathbb{F}$ s.t.

$$\Pr_{X \in \mathbb{F}^n}[H'(g_1(X), ..., g_m(X)) \neq p(X)] < \epsilon_d$$

For every region R_c, let η_c be the probability that p is different from G on that region (G is constant on the region).

$$\eta_c = \Pr_{X \in R_c}[p(X) \neq G|_{R_c}]$$

Since the average of η_c is at most ϵ_d, and all regions are almost uniform (Lemma 4.7) there can be at most $\sqrt{\epsilon_d}|\mathbb{F}|^m$ regions on which $\eta_c > \sqrt{\epsilon_d}$. We call these the *bad regions*, and we call the rest of the regions *almost good regions*. Next we show (Lemma 4.8) that the almost good regions are totally good and p is fixed on them. Last, we use the fact that there are only few bad regions and p is fixed on the rest to conclude that p is also fixed on the bad regions (Lemma 4.10). Thus, $p(X)$ is in fact constant on all regions. To complete the proof of Lemma 4.6, it remains to prove Lemmas 4.7, 4.8 and 4.10. The following lemma is a direct implication of Corollary 4.1.

Lemma 4.7 (Regions are uniform)**.** *Let $\gamma = \gamma(m) > 0$ be a small enough error term. If $\mathcal{F}$ is large enough than $|R_c| = |\mathbb{F}|^{n-m}(1 \pm \gamma)$, for all $c \in \mathbb{F}^m$.*

Proof. Let $c \in \mathbb{F}^m$ and assume first that R_c is not empty, i.e. there exist some x s.t. $g_i(x) = c_i$ for all $i \in [m]$. We apply Corollary 4.1 with $k = 1$, $x' = x$ and $y_1 = 0$ and get:

$$\Pr_{Y_1}[g_i(x + Y_1) = g_i(x),\ \forall\ i \in [m]] = |\mathbb{F}|^{-m}(1 \pm \gamma)$$

Substituting $Y = x + Y_1$ proves the result for R_c.

To show that there can be no empty regions, assume otherwise. Thus, there are at most $|\mathbb{F}|^m - 1$ non-empty cells, and each has volume at most $|\mathbb{F}|^{n-m}(1 + \gamma)$. Thus $(|\mathbb{F}|^m - 1)|\mathbb{F}|^{n-m}(1 + \gamma) \geq |\mathbb{F}|^n$. If $\gamma(m) < |\mathbb{F}|^{-m}$ we get a contradiction. Thus, there are no empty regions, and so all regions have volume $|\mathbb{F}|^{n-m}(1 \pm \gamma)$. □

Lemma 4.8 (Almost good regions are good)**.** *Let R_c be a region s.t $\Pr_{X \in R_c}[p(X) = b] > 1 - 2^{-2(d+1)}$, for some constant $b \in \mathbb{F}$. Then $p(X) = b$ for all $X \in R_c$.*

Before proving the lemma we need the following counting lemma on the number of hypercubes and pairs of hypercubes inside a region, similar to one in [GT07]. However, our technique avoids the need of interpolation.

Lemma 4.9. *Let $\gamma = \gamma(m) > 0$ be small enough error term, and assume $\mathcal{F}$ is large enough. For any point $R = R_c$ and a point $x \in R$ we have:*

1. *Let $Y_1, ..., Y_{d+1}$ be variables in $\mathbb{F}^n$. Then:*

$$\Pr_{Y_1,...,Y_{d+1} \in \mathbb{F}^n}[x + Y_I \in R,\ \forall I \subseteq [d+1]] = |\mathbb{F}|^{-\sum_{i=1}^{m}\sum_{j=1}^{\Delta(g_i)}\binom{d+1}{j}}(1 \pm \gamma)$$

2. *Let $Y_1, ..., Y_{d+1}, Z_1, ..., Z_{d+1}$ be variables in $\mathbb{F}^n$. For any non-empty $I_0 \in [d+1]$:*

$$\Pr\left[x + Y_I \in R, x + Z_I \in R,\ \forall I \subseteq [d+1] | Y_{I_0} = Z_{I_0}\right] \leq |\mathbb{F}|^m \left(|\mathbb{F}|^{-\sum_{i=1}^{m}\sum_{j=1}^{\Delta(g_i)}\binom{d+1}{j}}\right)^2 (1 + \gamma)$$

Proof of Lemma 4.9. In the following we show that the two conditions of the lemma hold.

1. This is a direct application of Corollary 4.1 for $k = d+1$, $x' = x$ and $y_1, ..., y_k = 0$.

2. Assume w.l.o.g that $I_0 = \{1, 2, ..., s\}$ for $1 \leq s \leq d+1$. We start by making a linear transformation on the coordinates to bring Y_{I_0} and Z_{I_0} to a single variable. Let $Y'_i = Y_i$ for $i \neq s$ and $Y'_s = Y_1 + ... + Y_s$, and similarly define $Z'_1, ..., Z'_{d+1}$. We write Y_I in the basis of $Y'_1, ..., Y'_{d+1}$. Divide $I = I_s \cup I_{\bar{s}}$ where $I_s = I \cap [s]$ and $I_{\bar{s}} = I \setminus I_s$. We have:

   - If $s \notin I$, $Y_I = \sum_{i \in I} Y'_i$
   - If $s \in I$, $Y_I = Y'_s - \sum_{i \in [s] \setminus I_s} Y'_i + \sum_{i \in I_{\bar{s}}} Y'_i$

   Consider for every I the set T_I of indices of Y'_i which appear in the expansion of Y_I. Notice that for any $T \subseteq [d+1]$ there is exactly one I s.t. $T_I = T$. In particular, in order that $g_i(x + Y_I) = g_i(x)$ for all I, we must have in particular that:

   - For any $I \subseteq [d+1]$ s.t. $s \notin I$ and $|I| \leq \Delta(g_i)$,
   $$g_i(x + Y'_I) = g_i(x)$$
   - For any $I \subseteq [d+1]$ s.t. $s \in I$ and $|I| \leq \Delta(g_i)$,
   $$g_i(x + Y'_s - Y'_{I \cap [s-1]} + Y'_{I \cap \{s+1, ..., d+1\}}) = g_i(x)$$

   Similarly for the Z''s, using the fact that the event $Y_{I_0} = Z_{I_0}$ translates to $Z'_s = Y'_s$:

   - For any $I \subseteq [d+1]$ s.t. $s \notin I$ and $|I| \leq \Delta(g_i)$,
   $$g_i(x + Z'_I) = g_i(x)$$
   - For any $I \subseteq [d+1]$ s.t. $s \in I$ and $|I| \leq \Delta(g_i)$,
   $$g_i(x + Y'_s - Z'_{I \cap [s-1]} + Z'_{I \cap \{s+1, ..., d+1\}}) = g_i(x)$$

   The probability of this event is an upper bound on our required probability. Since our variables
   $$Y'_1, ..., Y'_{d+1}, Z'_1, .., Z'_{s-1}, Z'_{s+1}, ..., Z'_{d+1}$$
   are uniform and independent, we can apply Lemma 4.3 to show that its probability is the required upper bound. The number of subsets of size $j > 1$ in the above events is $\binom{d+1}{j}$ for the event on the Y''s, and also $\binom{d+1}{j}$ for the event on $Z'_1, ..., Z'_{s-1}, Y'_s, Z'_{s+1}, ..., Z'_{d+1}$. For $j = 1$ however we have intersection (Y'_s is appearing twice), and so the number of events is $2\binom{d+1}{1} - 1$. Thus,by Lemma 4.3 the probability of the total event is:
   $$|\mathbb{F}|^m \left(|\mathbb{F}|^{-\sum_{i=1}^{m} \sum_{j=1}^{\Delta(g_i)} \binom{d+1}{j}} \right)^2 (1 \pm \gamma)$$
   which upper bounds the required probability.

□

We now prove Lemma 4.8 using Lemma 4.9. Our proof is similar to the one of [GT07].

Proof of Lemma 4.8. Let $B \subseteq R$ be the set of all "bad" points $x \in R$ on which $p(x) \neq b$. By our assumption, $|B| < 2^{-2(d+1)}|R|$. Assume B is non-empty, and choose some $x \in B$. Let $Y_1, ..., Y_{d+1}$ be random variables in $\mathbb{F}^n$. Fix small enough $\gamma = \gamma(m)$. By Lemma 4.9 (1),

$$p_R = \Pr[x + Y_I \in R,\ \forall I \subseteq [d+1]] \geq |\mathbb{F}|^{-\sum_{i=1}^{m}\sum_{j=1}^{\Delta(g_i)}\binom{d+1}{j}}(1-\gamma)$$

We now wish to bound the event that when all $X + Y_I$ are in R, some $X + Y_I$ is in B, and then union bound over all possible I.

We start by applying Cauchy-Schwarz to transform the problem to counting pairs of hypercubes. Fix some non-empty $I_0 \subseteq [d+1]$, and let

$$p_B = \Pr[x + Y_I \in R\ \forall I \subseteq [d+1]\ \wedge\ x + Y_{I_0} \in B] = \sum_{x_0 \in B} \Pr[x + Y_I \in R\ \forall I \subseteq [d+1]\ \wedge\ x + Y_{I_0} = x_0]$$

We need to upper bound p_B.

$$p_B^2 = \left(\sum_{x_0 \in B} \Pr[x + Y_I \in R\ \forall I \subseteq [d+1]\ \wedge\ x + Y_{I_0} = x_0]\right)^2 \leq |B| \sum_{x_0 \in B} \Pr[x + Y_I \in R\ \forall I \subseteq [d+1]\ \wedge\ x + Y_{I_0} = x_0]^2$$

Introducing new variables $Z_1, ..., Z_{d+1}$ in $\mathbb{F}^n$, we have.

$$p_B^2 \leq |B| \Pr[x + Y_I \in R\ \forall I \subseteq [d+1] \wedge x + Z_I \in R\ \forall I \subseteq [d+1] \wedge x + Y_{I_0} = x + Z_{I_0}]$$

Thus we get that:

$$p_B^2 \leq |B||\mathbb{F}|^{-n} \Pr[x + Y_I \in R,\ x + Z_I \in R\ \forall I \subseteq [d+1]| x + Y_{I_0} = x + Z_{I_0}]$$

By claim (2) in Lemma 4.9 we get that this probability is at most

$$|B||\mathbb{F}|^{m-n} p_R^2 (1+\gamma)$$

By Lemma 4.7, $|R| = |\mathbb{F}|^n \Pr_{X \in \mathbb{F}^n}[X \in R] = |\mathbb{F}|^{n-m}(1 \pm \gamma)$. Thus, we have that:

$$p_B^2 \leq \frac{|B|}{|R|} p_R^2 (1 \pm 2\gamma) \leq 2^{-2(d+1)} p_R^2$$

and thus $\frac{p_B}{p_R} \leq 2^{-(d+1)}(1 \pm 2\gamma)$. We can now union bound over all non-empty $I_0 \subseteq [d+1]$. The probability that there is some I_0 for which $x + Y_{I_0} \in B$ is at most

$$(2^{d+1} - 1)(2^{-(d+1)} + \gamma) < 1$$

for small enough γ.

Thus, there must exist $y_1, ..., y_{d+1} \in \mathbb{F}^n$ s.t.

$$x + y_I \in R \setminus B$$

for all non-empty $I \subseteq [d+1]$. Equivalently, $p(x + y_I) = b$ for all such I's. However, since $p(X)_{y_1,...,y_{d+1}} \equiv 0$,

$$p(x) = \sum_{I \subseteq [d+1], |I|>0} (-1)^{|I|+1} p(x + y_I)$$

and so if all $p(x + y_I) = b$, then also $p(x) = b$, hence $x \notin B$. So we have proved that B is empty, i.e. p is constant on R. □

We finish the proof of Lemma 4.6 by proving that if $p(X)$ is constant over almost all regions, then it must be constant over any region.

Lemma 4.10 (If almost all regions are totally good, all are totally good)**.** *Assume that the fraction of regions on which p is constant is at least $1 - 2^{-(d+2)}$. Then p is constant over any region.*

Proof. Let R be any region, and $x, x' \in R$ two points in R. We need to show that $p(x) = p(x')$. Choose $y'_1, ..., y'_{d+1} \in \mathbb{F}^n$ randomly. The probability that $x' + y'_I$ falls in a bad region for any non-empty $I \subseteq [d+1]$ is $2^{-(d+2)}$ (since regions are almost uniform, see Lemma 4.7). Thus, applying union bound over all non-empty $I \subseteq [d+1]$ we get that $\{x' + y'_I\}$ fall in good regions for all non-empty I with probability at least $1/2$. Fix some $y'_1, ..., y'_{d+1}$ fulfilling this requirement.

Let $Y_1, ..., Y_{d+1} \in \mathbb{F}^n$ be random variables. Since $g_i(x) = g_i(x')$ for all $i \in [m]$ we can apply Corollary 4.1:

$$\Pr\left[g_i(x + Y_I) = g_i(x' + y'_I) \ \forall \ i \in [m], \ I \subseteq [d+1]\right] = |\mathbb{F}|^{-\sum_{i=1}^{m} \sum_{j=1}^{\Delta(g_i)} \binom{d+1}{j}} (1 \pm \gamma)$$

In particular, for small enough γ we get that

$$\Pr\left[g_i(x + Y_I) = g_i(x' + y'_I) \ \forall \ i \in [m], \ I \subseteq [d+1]\right] > 0$$

Let $y_1, ..., y_{d+1}$ be such assignment to $Y_1, ..., Y_{d+1}$. We thus have that for all non-empty $I \subseteq [d+1]$ and for all $i \in [m]$, $g_i(x + y_I) = g_i(x' + y'_I)$. Since the region of $x' + y'_I$ is good for all non-empty I, we get that for all non-empty $I \subseteq [d+1]$, $p(x + y_I) = p(x' + y'_I)$. We now use the fact that p is a degree d polynomial. If we derive p $d+1$-times in any direction, we will always get zero. We thus have that for $x, y_1, ..., y_{d+1} \in \mathbb{F}^n$: $\sum_{I \subseteq [d+1]} (-1)^{|I|} p(x + y_I) = 0$. Since the same identity is true for $x', y'_1, ..., y'_{d+1}$, we get that $p(x) = p(x')$. □

Chapter 5

Random degree d polynomials are far from d-1 polynomials

We study the problem of how well a typical multivariate polynomial can be approximated by lower degree polynomials over $\mathbb{F}$. We prove that almost all degree d polynomials have only an exponentially small correlation with all polynomials of degree at most $d-1$, for all degrees d up to $\Theta(n)$. That is, a random degree d polynomial does not admit a good approximation of lower degree. In order to prove this, we prove far tail estimates on the distribution of the bias of a random low degree polynomial. Recently, several results regarding the weight distribution of Reed–Muller codes were obtained. Our results can be interpreted as a new large deviation bound on the weight distribution of Reed–Muller codes.

Joint work with Ido Ben-Eliezer and Rani Hod.

5.1 Introduction

Two functions $f, g : \mathbb{F}^n \to \mathbb{F}$ are said to be ϵ-*correlated* if

$$\Pr\left[f(x) = g(x)\right] \geq \frac{1+\epsilon}{2}.$$

A function $f : \mathbb{F}^n \to \mathbb{F}$ is said to be ϵ-*correlated* with a set of functions $F \subseteq \mathbb{F}^n \to \mathbb{F}$ if it is ϵ-correlated with at least one function $g \in F$.

We are interested in functions that have a low correlation with the set of degree $d-1$ polynomials; namely, functions that cannot be approximated by any polynomial of total degree at most $d-1$. How *complex* must such a function be? We use the most natural measure for complexity in these settings, which is the degree of the function when considered as a polynomial.

A simple probabilistic argument shows that for any constant $\delta < 1$ and for $d < \delta n$, a random function has an exponentially small correlation with degree $d-1$ polynomials. However, a random function is complex since, with high probability, its degree is at least $n-2$. In this work, we study how well a random degree d polynomial can be approximated by any

lower degree polynomial, and show that with very high probability a random polynomial of degree d cannot be approximated by polynomials of lower degree in a strong sense. Thus, if we want to find functions that are uncorrelated with degree $d-1$ polynomials, considering degree d polynomials is enough.

5.1.1 Motivation

The correlation of a typical degree d polynomial with the set of lower degree polynomials is a natural question in arithmetic complexity. More generally, the study of the correlation of functions with the set of low degree polynomials is interesting from both coding theory and complexity theory points of view.

Complexity Theory. Approximation of functions by low degree polynomials is one of the main tools used in proving lower bounds for constant depth circuits. For example, Razborov and Smolensky [Raz87, Smo87] provided an explicit function $\textsc{Mod}_3$ that cannot be computed by a constant depth circuit with a subexponential number of And, Or and Xor gates. The proof combines two arguments:

1. Any constant depth circuit of subexponential size has a very high correlation (that is, $1-o(1)$) with some polynomial of degree n^{ϵ};
2. Such a low degree polynomial has a correlation of at most $2/3$ with $\textsc{Mod}_3$. (In fact, this is true for any polynomial of degree at most $\epsilon\sqrt{n}$ for some constant ϵ.)

The best known constructions of explicit functions that cannot be approximated by low degree polynomials (see, e. g., [BSK08, BNS, Raz87, Smo87, VW08]) fall into two categories:

- For large degree bounds ($d < n^{\Omega(1)}$), there exists a symmetric function with a correlation of at most $O(1/\sqrt{n})$ with degree $O(\sqrt{n})$ polynomials;
- For small degree bounds ($d < \log n$) there are explicit functions having a correlation of at most $\exp(-n/c^d)$ with degree d polynomials for some constants c (best known is $c = 2$.)

Certain applications, e. g., pseudorandom generator constructions via the Nisan–Wigderson construction [NW94], require a function having an exponentially small correlation with low degree polynomials. This is only known for degrees up to $\log n$, while for larger degrees the best known bound is polynomial in n. Finding explicit functions with a better correlation is an ongoing quest with limited success. For more details, see a survey by Viola [Vio09].

Coding Theory. The Reed–Muller code $\mathrm{RM}(n,d)$ is a linear code in which codewords correspond to polynomials (over $\mathbb{F}$) in n variables of total degree at most d. This family of codes is one of the most studied objects in coding theory (see, e.g., [MS83]). Nevertheless, determining the weight distribution of these codes (for $d \geq 3$) is a long standing open problem. Interpreted in this language, our main lemma gives a new tail estimate on the weight distribution of Reed–Muller codes.

5.1.2 Our results

We show that, with very high probability, a random degree d polynomial has an exponentially small correlation with polynomials of lower degree, i.e. of degree at most $d-1$. We prove this for degrees ranging from a constant up to $\delta_{\max} n$, where $0 < \delta_{\max} < 1$ is an absolute constant. All results hold for large enough n.

Theorem 5.1 (Main Theorem)**.** *There exist constants $0 < \delta_{\max} < 1$ and $c, c' > 0$ such that the following holds. for every $d \leq \delta_{\max} n$ let f be a random n-variate polynomial of degree d. Then the probability that f has a correlation $2^{-cn/d}$ with polynomials of degree at most $d-1$ is at most $2^{-c'\binom{n}{\leq d}}$, where $\binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i}$.*

The main theorem is an easy corollary of the following lemma, which is the main technical contribution of the paper. We define the *bias* of a function $f : \mathbb{F}^n \to \mathbb{F}$ to be

$$bias\left(f\right) = \mathbb{E}_x\left[(-1)^{f(x)}\right] = \Pr\left[f\left(x\right) = 0\right] - \Pr\left[f\left(x\right) = 1\right].$$

Lemma 5.1 (Main Lemma)**.** *Fix $\epsilon > 0$ and let f be a random degree d polynomial for $d \leq (1-\epsilon)\, n$. Then,*

$$\Pr\left[\left|bias\left(f\right)\right| > 2^{-c_1 n/d}\right] \leq 2^{-c_2\binom{n}{\leq d}},$$

where $0 < c_1, c_2 < 1$ are constants depending only on ϵ.

Note that Lemma 5.1 holds for degrees up to $(1-\epsilon)\, n$, while we were only able to prove Theorem 5.1 for degrees up to $\delta_{\max} n$. The proof of Lemma 5.1 appears in Section 5.2.1.

The following proposition (proved in Section 5.3) shows that the estimate in Lemma 5.1 is somewhat tight for degrees up to $n/2$.

Proposition 5.1. *Fix $\epsilon > 0$ and let f be a random degree d polynomial for $d \leq (1/2-\epsilon)\, n$. Then,*

$$\Pr\left[\left|bias\left(f\right)\right| > 2^{-c_1' n/d}\right] \geq 2^{-c_2'\binom{n}{\leq d}},$$

where $0 < c_1', c_2' < 1$ are constants depending only on ϵ.

Our proof of Lemma 5.1 uses the following tight lower bound on the dimension of truncated Reed–Muller codes, which has appeared in [KS05, Theorem 1.5]. For the sake of self containment, we present an alternative proof of Lemma 5.2. Our proof, unlike the original, has an algorithmic flavor.

Lemma 5.2. *Let $x_1, \ldots, x_{2^m}$ be 2^m distinct points in $\mathbb{F}^n$. Consider the linear space of degree d polynomials restricted to these points; that is, the space*

$$\left\{\left(p\left(x_1\right), \ldots, p\left(x_{2^m}\right)\right) : p \in \mathrm{RM}\left(n, d\right)\right\}.$$

The linear dimension of this space is at least $\binom{m}{\leq d}$.

5.1.3 Related Work

The weight distribution of Reed–Muller codes is completely known for $d = 2$ (see, for example, [CHSL97]) and some partial results are known also for $d = 3$. In the general case, there are estimates (see, e.g., [KT70, KTA76]) on the number of codewords with weight between w and $2.5w$, where $w = 2^{-d}$ is the minimal weight of the code. Kaufman and Lovett [KLP10] proved bounds for larger weights, and following Gopalan et al. [GKZ08], they used it to prove new bounds for the *list-decoding* of Reed–Muller codes.

The case of multilinear polynomials was considered by Alon et al. [ABEK08], who proved a tail estimate similar to Lemma 5.1 and used it to prove bounds on the size of distributions that fool low degree polynomials. Namely, they prove that for any distribution $\mathcal{D}$ that fools degree d polynomials with error ϵ,

$$|support(\mathcal{D})| \geq \Omega\left(\frac{(n/2d)^d}{\epsilon^2 \log(1/\epsilon)}\right).$$

Substituting our Lemma 5.1 for [ABEK08, Lemma 1] yields

$$|support(\mathcal{D})| \geq \Omega\left(\frac{\binom{n}{d}}{\epsilon^2 \log(1/\epsilon)}\right),$$

improving the lower bound for the case of polynomials over $\mathbb{F}^n$ by a factor of roughly $(2e)^d$.

The Gowers Norm is a measure related to the approximability of functions by low degree polynomials. It was introduced by Gowers [Gow01] in his seminal work providing a new proof for Szemerdi's Theorem. Using the Gowers Norm machinery, it is easy to prove that a random polynomial of degree $d < \log n$ has a small correlation with lower degree polynomials. However, this approach fails for degrees exceeding $\log n$. In constrast, note that our result holds for degrees up to $\delta_{\max} n$.

Green and Tao [GT07] study the structure of biased multivariate polynomials. They prove that if their degree is at most the size of the field (which in our case is 2), then they must have structure — they can be expressed as a function of a constant number of lower degree polynomials. Kaufman and Lovett [KL08] strengthen this structure theorem for polynomials of every constant degree, removing the field size restriction.

5.2 Proof of the Main Theorem

First we show that Theorem 5.1 follows directly from Lemma 5.1 by a simple counting argument.

Let f be a random degree d polynomial for $d \leq \delta_{\max} n$, where $\delta_{\max}$ will be determined later. For every polynomial g of degree at most $d-1$, $f - g$ is also a random degree d polynomial. By the union bound for all possible choices of g,

$$\Pr_f\left[\exists g \in \mathrm{RM}(n, d-1) : |bias(f-g)| \geq 2^{-c_1 n/d}\right] \leq 2^{\binom{n}{\leq d-1} - c_2 \binom{n}{\leq d}}$$

Choosing $\delta_{\max}$ to be a small enough constant, we get that there is a constant $c' > 0$ such that $c_2\binom{n}{\leq d} - \binom{n}{\leq d-1} \geq c'\binom{n}{\leq d}$ for all $d \leq \delta_{\max} n$ (see, for example, [Juk01, Exercise 1.14]).

We now move on to prove Lemma 5.1. To keep the proof's flow, the rest of this section is organized as follows. Subsection 5.2.1 presents the gist of the proof, employing several technical claims. The proofs of these claims are given in Subsection 5.2.2. Lemma 5.2, which is used by one of the claims but is interesting also out of the context of this proof, is proved in Subsection 5.2.3.

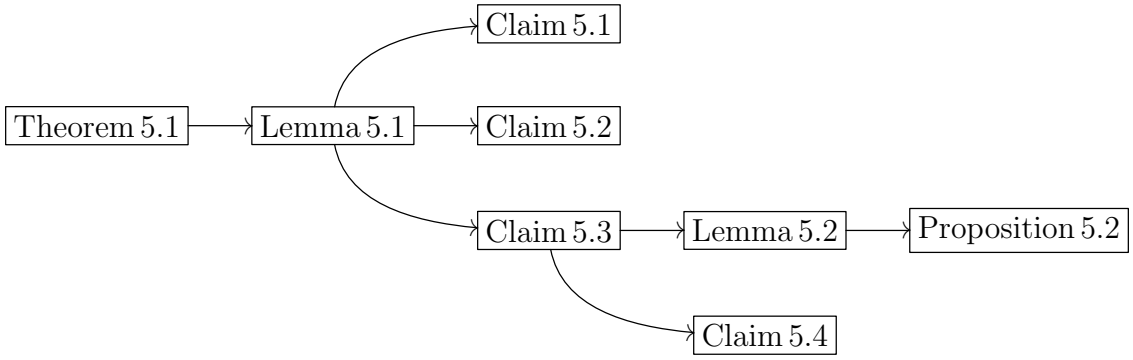

Figure 5.1: Proof tree for Theorem 5.1

5.2.1 Proof of Lemma 5.1

We need to prove that a random degree d polynomial has a very small bias with very high probability. Denote by $\mathrm{RM}(n,d)^{\perp}$ the dual code of $\mathrm{RM}(n,d)$. We start by correlating the moments of the bias of a random degree d polynomial to short words in $\mathrm{RM}(n,d)^{\perp}$.

Claim 5.1. *Fix $t \in \mathbb{N}$ and let $p \in \mathrm{RM}(n,d)$ and $x_1, \ldots, x_t \in \mathbb{F}^n$ be chosen independently and equiprobably. Then,*

$$\mathbb{E}\left[bias(p)^t\right] = \Pr\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right],$$

where e_x for $x \in \mathbb{F}^n$ is the unit vector in $\mathbb{F}^{2^n}$, having 1 in position x and 0 elsewhere.

We proceed by introducing the following definitions. Fix d. For $x \in \mathbb{F}^n$ let $\mathrm{eval}_d(x)$ denote its *d-evaluation*; that is, a (row) vector in $\mathbb{F}^{\binom{n}{\leq d}}$ whose coordinates are the evaluation of all monomials of degree up to d at the point x. Formally,

$$\mathrm{eval}_d(x) = \left(\prod_{i \in I} x(i)\right)_{I \subset [n], |I| \leq d}.$$

For points $x_1, \ldots, x_t \in \mathbb{F}^n$ let $\mathcal{M}_d(x_1, \ldots, x_t)$ denote their *d-evaluation matrix*; this is a $t \times \binom{n}{\leq d}$ matrix whose ith row is the d-evaluation of x_i. We denote the rank of $\mathcal{M}_d(x_1, \ldots, x_t)$ by $\mathrm{rank}_d(x_1, \ldots, x_t)$. As this value is independent of the order of $x_1, \ldots, x_t$, we may refer without ambiguity to the *d-rank* of a set $S \subseteq \mathbb{F}^n$ by $\mathrm{rank}_d(S)$.

According to Claim 5.1, in order to bound the moments of the bias of a random polynomial we need to study the probability that a random word of length about[1] t is in $\mathrm{RM}(n,d)^{\perp}$.

Let $A = \mathcal{M}_d(x_1, \ldots, x_t)$. Note that $e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}$ if and only if $e_{x_1} + \cdots + e_{x_t}$ is orthogonal to all degree d polynomials, namely, if

$$p(x_1) + \cdots + p(x_t) = 0 \tag{5.1}$$

for any degree d polynomial p. It is sufficient to satisfy (5.1) only on the monomial basis of the degree d polynomials; that is, verify that each column in A sums to zero. Therefore, $e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}$ if and only if the sum of the rows of A is zero.

We turn to upper bound the probability that the rows of A sum to the zero vector for random $x_1, \ldots, x_t \in \mathbb{F}^n$. Instead of requiring that *every* column of A sums to zero, we require this only for columns corresponding to "special" monomials. For this we divide the n variables into two sets: L of size $l = \lfloor n/d \rfloor$ and R of size $r = n - l$. The special monomials that interest us are exactly those that contain exactly one variable from the left side L (and thus up to $d-1$ variables from R).

For $i = 1, \ldots, t$ denote by y_i ($\in \mathbb{F}^r$) the restriction of $x_i \in \mathbb{F}^n$ to the variables in R. The following claim bounds the probability that the sum of A's rows is zero in terms of $\alpha = l/n \approx 1/d$ and the $(d-1)$-rank of $y_1, \ldots, y_t$.

Claim 5.2.

$$\Pr_{\{x_i\}}\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right] \leq \mathbb{E}_{\{y_i\}}\left[2^{-\mathrm{rank}_{d-1}(y_1,\ldots,y_t)\alpha n}\right].$$

To finish the proof, we provide a (general) lower bound on d-ranks of random vectors.

Claim 5.3. *For all fixed $\beta < 1$ and $\delta < 1$, there exist constants $c > 0$ and $\eta > 1$ such that if $x_1, \ldots, x_t \in \mathbb{F}^n$ are chosen uniformly and independently, where $t \geq \eta\binom{n}{\leq d}$ and $d \leq \delta n$, then*

$$\Pr\left[\mathrm{rank}_d(x_1, \ldots, x_t) < \beta\binom{n}{\leq d}\right] \leq 2^{-c\binom{n}{\leq d+1}}.$$

We now put it all together, in order to complete the proof of Lemma 5.1. According to Claim 5.2, we have

$$\Pr_{\{x_i\}}\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right] \leq \mathbb{E}_{\{y_i\}}\left[2^{-\mathrm{rank}_{d-1}(y_1,\ldots,y_t)\alpha n}\right].$$

Applying Claim 5.3 for $d-1$ and r (instead of d and n in the claim statement), and assuming $t \geq \eta\binom{r}{\leq d-1}$, we get that

$$\Pr\left[\mathrm{rank}_{d-1}(y_1, \ldots, y_t) < \beta\binom{r}{\leq d-1}\right] < 2^{-c\binom{r}{\leq d}}.$$

Therefore,

$$\Pr_{\{x_i\}}\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right] \leq 2^{-\beta\binom{r}{\leq d-1}\alpha n} + 2^{-c\binom{r}{\leq d}}.$$

[1] We say "about t" as $x_1, \ldots, x_t$ might not be distinct.

Recalling that $r = n - \lfloor n/d \rfloor$ and $\alpha = 1 - r/n = 1/d + O(1/n)$, we get that for any constant β (and $c = c(\beta)$) there is a constant c' such that

$$\Pr_{\{x_i\}}\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right] \leq 2^{-c'\binom{n}{\leq d}}.$$

This is because $\binom{r}{\leq d-1} = \Theta\left(\binom{n}{\leq d} d/n\right)$ and $\binom{r}{\leq d} = \Theta\left(\binom{n}{\leq d}\right)$.

We thus proved that there is a constant c' such that

$$\mathbb{E}_{f \in \mathrm{RM}(n,d)}\left[bias(f)^t\right] \leq 2^{-c'\binom{n}{\leq d}},$$

for $t = \eta\binom{r}{\leq d-1} = \Theta\left(\binom{n}{\leq d-1}\right)$. Hence, $tn/d \leq c''\binom{n}{\leq d}$ for some constant c''.

For small enough $c_1 > 0$ such that $c_2 = c' - c''c_1 > 0$, by Markov inequality,

$$\Pr\left[|bias(f)| \geq 2^{-c_1 n/d}\right] \leq 2^{tc_1 n/d - c'\binom{n}{\leq d}} \leq 2^{(c''c_1 - c')\binom{n}{\leq d}} \leq 2^{-c_2\binom{n}{\leq d}}.$$

5.2.2 Proofs of technical claims

Proof of Claim 5.1. Write p as

$$p(x) = \sum_{I \subset [n], |I| \leq d} \alpha_I \prod_{i \in I} x(i),$$

where $x(i)$ denotes the ith coordinate of $x \in \mathbb{F}^n$. As p was chosen uniformly, all α_I are uniform and independent over $\mathbb{F}$. Therefore,

$$\begin{aligned}
\mathbb{E}_p\left[(bias(p))^t\right] &= \mathbb{E}_p\left[\prod_{j=1}^{t} \mathrm{bias}(p)\right] \\
&= \mathbb{E}_{\{\alpha_I\}}\left[\prod_{j=1}^{t} \mathbb{E}_{x_j}\left[(-1)^{\sum_I \alpha_I \prod_{i \in I} x_j(i)}\right]\right] \\
&= \mathbb{E}_{\{x_j\}}\left[\prod_{I} \mathbb{E}_{\alpha_I}\left[(-1)^{\alpha_I(\sum_{j=1}^{t} \prod_{i \in I} x_j(i))}\right]\right] \\
&= \mathbb{E}_{\{x_j\}}\left[\prod_{I} \mathbf{1}_{\{\sum_{j=1}^{t} \prod_{i \in I} x_j(i) = 0\}}\right] \\
&= \Pr_{\{x_j\}}\left[\forall I \sum_{j=1}^{t} \prod_{i \in I} x_j(i) = 0\right] \\
&= \Pr_{\{x_j\}}\left[e_{x_1} + \cdots + e_{x_t} \in \mathrm{RM}(n,d)^{\perp}\right].
\end{aligned}$$

□

Proof of Claim 5.2. Let $A' = \mathcal{M}_{d-1}(y_1, \ldots, y_t)$ be the $t \times \binom{r}{\leq d-1}$ sub-matrix of A corresponding to monomials of degree at most $d-1$ in variables from R. Let $\mathcal{E}$ be the event in which

every column of A corresponding to a monomial that contains exactly one variable from L sums to zero.

We observe that this event is equivalent to the event that every column of A' is orthogonal to the set of vectors $\{(x_1(i), \ldots, x_t(i)) : i \in L\}$, as the inner product of every column from A' with vectors from this set corresponds to a monomial of degree at most d.

Fix the variables in R; this determines A'. As the variables in L are independent of those in R, the probability of $\mathcal{E}$ (given A') is

$$\left(2^{-\mathrm{rank}(A')}\right)^{|L|} = 2^{-\mathrm{rank}(A')\alpha n} = 2^{-\mathrm{rank}_{d-1}(y_1,\ldots,y_t)\alpha n}.$$

This holds for every assignment for the variables in L, hence the result follows. □

Proof of Claim 5.3. Let $B = \mathcal{M}_d(x_1, \ldots, x_t)$ be the $t \times \binom{n}{\leq d}$ d-evaluation matrix of the random $x_1, \ldots, x_t \in \mathbb{F}^n$. We need to bound the probability that $\mathrm{rank}(B) < \beta\binom{n}{\leq d}$.

Fix some $b \leq \beta\binom{n}{\leq d}$, and let us consider the event that the first b rows of B span the entire row span of B. Denote by V the linear space spanned by the first b rows of B. Since all rows of B are d-evaluations of some points in $\mathbb{F}^n$, we need to study the maximum number of d-evaluations contained in a linear subspace of dimension b.

Assume there are at least 2^r distinct d-evaluations in V. By Lemma 5.2, $\dim(V) \geq \binom{r}{\leq d}$. Assume further that $\mathrm{rank}(B) < \beta\binom{n}{\leq d}$; we get that

$$\beta\binom{n}{\leq d} > \mathrm{rank}(B) \geq \dim(V) \geq \binom{r}{\leq d}.$$

By Claim 5.4 below, $r \leq n(1 - \gamma/d)$, where γ is a constant depending only on β. In other words, out of the 2^n d-evaluations of all points in $\mathbb{F}^n$, at most $2^{n(1-\gamma/d)}$ fall in V and hence the probability that a random d-evaluation is in V is at most $2^{-\gamma n/d}$.

Assume without loss of generality that the number of rows t is exactly $\eta\binom{n}{\leq d}$ for some $\eta > 1$. The probability that all the remaining rows of B are in V is at most

$$\left(2^{-\gamma n/d}\right)^{t-b} \leq 2^{-(\eta-\beta)\binom{n}{\leq d}\gamma n/d} \leq 2^{-\gamma\rho(\eta-\beta)\binom{n}{\leq d+1}},$$

where the last inequality follows from the fact that there exists a constant $\rho > 0$ such that $(n/d)\binom{n}{\leq d} \geq \rho\binom{n}{\leq d+1}$ for all n and d.

Choosing $\eta > \beta$ (and large enough n), we get that when we union bound over all possible ways to choose at most $\beta\binom{n}{\leq d}$ rows out of $t = \eta\binom{n}{\leq d}$, the probability that any of them spans the rows of B is at most $2^{-c\binom{n}{\leq d+1}}$, where c depends only on β. □

Claim 5.4. *For any $\beta, \delta < 1$, there is a constant $\gamma = \gamma(\beta, \delta)$ such that if $1 \leq d \leq \delta n$ and $k \geq d$ satisfy $\beta\binom{n}{\leq d} \geq \binom{k}{\leq d}$ then $k \leq n(1 - \gamma/d)$.*

Proof. We bound

$$\frac{1}{\beta} \leq \frac{\binom{n}{\leq d}}{\binom{k}{\leq d}} \leq \max_{0\leq i\leq d} \frac{\binom{n}{i}}{\binom{k}{i}} = \frac{\binom{n}{d}}{\binom{k}{d}} = \prod_{i=0}^{d-1} \frac{n-i}{k-i} \leq \left(\frac{n-d}{k-d}\right)^d = \left(1 + \frac{n-k}{k-d}\right)^d.$$

Assuming for the sake of contradiction that $k > n(1-\gamma/d)$ and taking logarithms, we get

$$\ln \frac{1}{\beta} \leq d \ln \left(1 + \frac{n-k}{k-d}\right) \leq \frac{d(n-k)}{k-d} < \frac{\gamma n}{k-d} < \frac{\gamma}{k/n-\delta} < \frac{\gamma}{1-\delta-\gamma/d}\,.$$

This can be made false by picking, e.g., $\gamma = \frac{(1-\delta)\ln(1/\beta)}{1+\ln(1/\beta)}$. □

5.2.3 Proof of Lemma 5.2

Restating the lemma in terms of d-evaluations, we need to show that for every subset $S \subseteq \mathbb{F}^n$ of size 2^m, $\text{rank}_d(S) \geq \binom{m}{\leq d}$. Let $S = \{x_1, \ldots, x_{2^m}\}$ be the set of points. We simplify S by applying a sequence of transformations that do not increase its d-rank until we arrive to the linear space $\mathbb{F}^m \times \{0\}^{n-m}$, whose d-rank is exactly $\binom{m}{\leq d}$.

We now define our basic non-linear transformation Π, mapping the set S to a set $\Pi(S)$ of equal size and not greater d-rank. Informally, Π tries to set the first bit of each element in S to zero, unless this results in an element already in S (and in this case Π keeps the element unchanged). The operator Π was used in other contexts of extremal combinatorics, and is usually referred to as the *compressing* or *shifting* operator (see, e.g., [Alo83, Fra83].)

For $y = (y_1, \ldots, y_{n-1}) \in \mathbb{F}^{n-1}$, denote by $0y$ and $1y$ the elements $(0, y_1, \ldots, y_{n-1})$ and $(1, y_1, \ldots, y_{n-1})$ in $\mathbb{F}^n$, respectively. Extend this notation to sets by writing $0T = \{0y : y \in T\}$, $1T = \{1y : y \in T\}$ for a set $T \subseteq \mathbb{F}^{n-1}$.

We define the following three sets in $\mathbb{F}^{n-1}$.

$$\begin{aligned} T_* &= \{y \in \mathbb{F}^{n-1} :\ 0y \in S \text{ and } 1y \in S\}\,, \\ T_0 &= \{y \in \mathbb{F}^{n-1} :\ 0y \in S \text{ and } 1y \notin S\}\,, \\ T_1 &= \{y \in \mathbb{F}^{n-1} :\ 0y \notin S \text{ and } 1y \in S\}\,. \end{aligned}$$

Writing S as

$$S = 0T_* \cup 1T_* \cup 0T_0 \cup 1T_1\,,$$

we define $\Pi(S)$ to be

$$\Pi(S) = 0T_* \cup 1T_* \cup 0T_0 \cup 0T_1\,;$$

namely, we set to zero the first bit of all the elements in $1T_1$. It is easy to see that $|\Pi(S)| = |S|$ as $\Pi(S)$ introduces no collisions.

Proposition 5.2. $\text{rank}_d(\Pi(S)) \leq \text{rank}_d(S)$.

Proof. It will be easier to prove this using an alternative definition for $\text{rank}_d(S)$.

Let $(x_1, \ldots, x_{2^m})$ be some ordering of S. For a degree d polynomial $p \in \text{RM}(n,d)$, let $v_p \in \mathbb{F}^{2^m}$ be the evaluation of p on the points of S

$$v_p = (p(x_1), p(x_2), \ldots, p(x_{2^r}))\,.$$

Consider the linear space of vectors v_p for all $p \in \text{RM}(n,d)$. The dimension of this space is exactly $\text{rank}_d(S)$, as the monomials used in the definition of d-rank form a basis for the space of polynomials.

But now, instead of the dimension, consider the co-dimension. We call a point x_i, $1 \leq i \leq 2^m$, *dependent* if there are coefficients $\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{F}$ such that for all degree d polynomials

$$p(x_i) = \sum_{j=1}^{i-1} \alpha_j p(x_j) \,.$$

We thus expressed $\mathrm{rank}_d(S)$ as the number of independent points in S, which is the same as the difference between $|S| = 2^m$ and the number of dependent points in S. To prove that $\mathrm{rank}_d(\Pi(S)) \leq \mathrm{rank}_d(S)$, it suffices to show that Π maps dependent points in S to dependent images in $\Pi(S)$. Let us consider an ordering of S in which the elements of $1T_1$ come last. Since all other points in S are mapped to themselves by Π, it is clear that dependent points in S appearing before $1T_1$ are also dependent in $\Pi(S)$. It remains to prove the claim for points in $1T_1$.

Let $t_1 = |T_1|$ and let $y_1, \ldots, y_{t_1}$ be some ordering of T_1. Assume $1y_i \in S$ is dependent and we will show that $0y_i \in \Pi(S)$ is also dependent. By definition, there exist coefficients $\alpha_y, \beta_y, \gamma_y, \delta_y$ such that, for any degree d polynomial,

$$p(1y_i) = \sum_{y \in T_*} \alpha_y p(0y) + \sum_{y \in T_*} \beta_y p(1y) + \sum_{y \in T_0} \gamma_y p(0y) + \sum_{y_j \in T_1 : j < i} \delta_{y_j} p(1y_j) \,.$$

Each polynomial $p \in \mathrm{RM}(n, d)$ can be uniquely decomposed as

$$p(x_1, \ldots, x_n) = x_1 p'(x_2, \ldots, x_n) + p''(x_2, \ldots, x_n) \,,$$

where $p' \in \mathrm{RM}(n-1, d-1)$ and $p'' \in \mathrm{RM}(n-1, d)$. Moreover, for every $y \in \mathbb{F}^{n-1}$, we have that $p(0y) = p''(y)$ and $p(1y) = p'(y) + p''(y)$. Since p' and p'' are independent, we can decompose the dependency of $p(1y_i)$ into its p' and p'' components as follows.

$$p'(y_i) = \sum_{y \in T_*} \beta_y p'(y) + \sum_{y_j \in T_1 : j < i} \delta_{y_j} p'(y_j) \,, \tag{5.2}$$

$$p''(y_i) = \sum_{y \in T_*} (\alpha_y + \beta_y) p''(y) + \sum_{y \in T_0} \gamma_y p''(y) + \sum_{y_j \in T_1 : j < i} \delta_{y_j} p''(y_j) \,. \tag{5.3}$$

We now move to consider $\Pi(S)$. Every $1y_i$ for $y_i \in T_1$ is mapped to $0y_i$, so we should only consider the p'' component for T_1's elements. Also, by the definition of T_* and T_0, for each $y \in T_* \cup T_0$, $0y \in S \cap \Pi(S)$. By (5.3), for any $p \in \mathrm{RM}(n, d)$,

$$p(0y_i) = \sum_{y \in T_*} (\alpha_y + \beta_y) p(0y) + \sum_{y \in T_0} \gamma_y p(0y) + \sum_{y_j \in T_1 : j < i} \delta_{y_j} p(0y_j) \,,$$

that is, $0y_i$ is also dependent in $\Pi(S)$.

Therefore, we have established that $\mathrm{rank}_d(\Pi(S)) \leq \mathrm{rank}_d(S)$. □

We now combine our basic Π with invertible linear transformations to define a wider class of simplifying transformations. For any $u, v \in \mathbb{F}^n$ whose inner product is $\langle u, v \rangle = 1$, we define the mapping $\Pi_{u,v}$ as follows. Informally, $\Pi_{u,v}$ tries to add v to elements x of S for

which $\langle u, x\rangle = 1$, unless this results in an element already in S. In other words, if both x and $x + v$ are in S, then $\Pi_{u,v}(S)$ maps them both to themselves. Otherwise, if just one of them is in S, it maps it to x if $\langle u, x\rangle = 0$, and to $x + v$ if $\langle u, x + v\rangle = 0$. This is well defined as $\langle u, v\rangle = 1$. Note that $\Pi_{e_1,e_1} \equiv \Pi$.

Formally, let A be an $n \times n$ invertible matrix such that $e_1^T A = u$ and $A^{-1}e_1 = v$. We can construct such invertible A since $\langle u, v\rangle = 1$ by setting the first row of A to be u and the remaining rows of A to be a basis for the $(n-1)$-dimensional space normal to v. Define $\Pi_{u,v} = A^{-1}\Pi A$.

Note that this definition is equivalent to the informal one. Assume $\langle u, x\rangle = \langle e_1^T A, x\rangle = (Ax)_1 = 1$. Then Π would try to set $(Ax)_1$ to zero, that is, would add e_1 to Ax. Applying A^{-1}, this is the same as adding $A^{-1}e_1 = v$ to x.

Observe that invertible affine transformations do not change the d-rank of a set, as they act as permutations on the set of degree d polynomials. Combining this with Proposition 5.2, we get that $\Pi_{u,v}$ maintains the size of S without increasing the d-rank.

We now use a sequence of $\Pi_{u,v}$ applications to transform the set S into the linear space $V = \mathbb{F}^m \times \{0\}^{n-m}$ spanned by the first m unit vectors $e_1, \ldots, e_m$. We say that $x \in S$ is *good* if $x \in V$, and is *bad* otherwise. If all the elements of S are good then $S = V$ since all the elements of S are distinct. Otherwise, let $x \in S$ be some bad element and let $x' \in V \setminus S$. Since $x \notin V$, there must be some index $m < i \leq n$ such that $x_i = 1$; set $u = e_i$ and $v = x + x'$.

We show that applying $\Pi_{u,v}$ maps x to x' and does not affect any good elements, thus increasing the number of good elements. First see that $\langle u, v\rangle = v_i = x_i + x'_i = 1 + 0 = 1$ since $x' \in V$ so $\Pi_{u,v}$ is well defined. See also that as $\langle u, x\rangle = x_i = 1$ and $x + v \notin S$, $\Pi_{u,v}$ will add v to x, transforming it to $x' \in V$. Also, any good element y is unchanged by $\Pi_{u,v}$ since $\langle u, y\rangle = y_i = 0$. In total, the number of good elements increased by at least one.

We repeat this until all elements are good, that is, until S is transformed to V, establishing that $\mathrm{rank}_d(S) \geq \mathrm{rank}_d(V)$. To finish the proof, observe that the restriction of polynomials in $\mathrm{RM}(n, d)$ to points in a linear space of dimension m is exactly $\mathrm{RM}(m, d)$. Since $|\mathrm{RM}(m, d)| = \binom{m}{\leq d}$ (see [MS83]), we get that for any set S of size 2^m,

$$\mathrm{rank}_d(S) \geq \binom{m}{\leq d},$$

as required.

5.3 Proof of Proposition 5.1

Let $d < \gamma n$ for a constant $\gamma < 1/2$. We define a set of polynomials with measure of at least $2^{-c_2'\binom{n}{\leq d}}$ such that all polynomials in this set have a bias of at least $2^{-c_1' n/d}$ (for constants c_1', c_2'). That is, we will prove

$$\Pr_{f \in \mathrm{RM}(n,d)}\left[bias(f) > 2^{-c_1' n/d}\right] \geq 2^{-c_2'\binom{n}{\leq d}}.$$

Similar to the proof of Theorem 5.1, we divide the n variables into two sets: L of size $l = \lceil n/d \rceil$ and R of size $r = n - l$. Consider the set of monomials of degree at most d that have exactly one variable in L (and thus have degree at most $d - 1$ in R).

We first show that the number of such monomials is only a constant factor smaller than the number of all monomials of degree at most d. The number of monomials we consider is

$$\binom{l}{1}\binom{r}{\leq d-1} \geq \frac{n}{d}\binom{\lfloor n(1-1/d)\rfloor}{d-1}.$$

There exists a constant $c_\gamma > 0$ such that if $d < \gamma n$ then

$$\binom{\lfloor n(1-1/d)\rfloor}{d-1} \geq c_\gamma \binom{n-1}{d-1} \qquad \text{and also} \qquad \binom{n}{d} \geq c_\gamma \binom{n}{\leq d}.$$

Hence the number of monomials multilinear in L is at least $c_\gamma^2 \binom{n}{\leq d}$.

Let $\mathcal{L}$ be the linear space of polynomials on these monomials, $|\mathcal{L}| \geq 2^{c_\gamma^2 \binom{n}{\leq d}}$. Consider a random polynomial $f \in \mathcal{L}$. Since each monomial of f has exactly one variable in L, we can decompose f as the sum of products of a variable from L and a random degree $d-1$ polynomial from V''. That is, if $L = \{x_1, \ldots, x_l\}$ and $R = \{x_{l+1}, \ldots, x_n\}$, we can write

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{l} x_i g_i(x_{l+1}, \ldots, x_n).$$

We now show f has an expected bias of $2^{-l} \geq 2^{-n/d}$. Consider a partial assignment to the variables $x_1, \ldots, x_l$ of L. If all of them are zero, then $f(0, \ldots, 0, x_{l+1}, \ldots, x_n) \equiv 0$, and hence has bias 1. In all other cases, we are left with a random degree $d-1$ polynomial in the variables from R and as such it has bias 0 (e.g., since the constant term is random). Thus,

$$\begin{aligned} \mathbb{E}_{f\in\mathcal{L}}\left[bias(f)\right] &= 1 \cdot \Pr\left[\forall 1 \leq i \leq l : x_i = 0\right] \\ &+ 0 \cdot \Pr\left[\exists 1 \leq i \leq l : x_i \neq 0\right] = 2^{-l}, \end{aligned}$$

and we get that

$$\Pr\left[bias(f) > 2^{-(l+1)} \,\middle|\, f \in \mathcal{L}\right] > 2^{-(l+1)}.$$

We conclude that there is a constant c_2' such that

$$\Pr\left[bias(f) > 2^{-(n/d+1)}\right] \geq \Pr\left[f \in \mathcal{L}\right] \cdot \Pr\left[bias(f) > 2^{-(n/d+1)} \,\middle|\, f \in \mathcal{L}\right] \geq 2^{-c_2'\binom{n}{\leq d}}.$$

Chapter 6

Representation of boolean functions as polynomials in different characteristics

Every Boolean function on n variables can be expressed as a unique multivariate polynomial modulo p for every prime p. In this work, we study how the degree of a function in one characteristic affects its complexity in other characteristics. We establish the following general principle: *functions with low degree modulo p must have high complexity in every other characteristic q.* More precisely, we show the following results about Boolean functions $f : \{0,1\}^n \to \{0,1\}$ which depend on all n variables, and distinct primes p, q:

- If f has degree $o(\log n)$ modulo p, then it must have degree $\Omega(n^{1-o(1)})$ modulo q. Thus a Boolean function has degree $o(\log n)$ in at most one characteristic. This result is essentially tight as there exist functions that have degree $\log n$ in every characteristic.
- If f has degree $d = o(\log n)$ modulo p, then it cannot be computed correctly on more than $1 - p^{-O(d)}$ fraction of the hypercube by polynomials of degree $n^{\frac{1}{2}-\epsilon}$ modulo q.

As a corollary of the above results it follows that if f has degree $o(\log n)$ modulo p, then it requires super-polynomial size $\mathrm{AC}_0[q]$ circuits. This gives a lower bound for a broad and natural class of functions.

Joint work with Parikshit Gopalan and Amir Shpilka.

6.1 Introduction

Representations of Boolean functions as polynomials in various characteristics have been studied intensively in Computer science [NS92, Pat92, Bei93, BBR94]. This algebraic view of Boolean functions has found numerous applications to diverse areas including circuit lower bounds [Raz87, Smo87, BRS, ABFR94], computational learning [KM93, LMN93, KS01, MOS03] and explicit combinatorial constructions [Gro00, Gro02, Gop06b, Efr09]. As a purely algebraic model of computation, polynomial representations lead to some natural complexity

measures such as exact degree, approximation degree and sparsity needed to represent a function. In this work, we are primarily concerned with the polynomial degree of a function, defined as follows:

Definition 6.1. *For a Boolean function $f : \{0,1\}^n \to \{0,1\}$, the degree of f in characteristic k, denoted $\deg_k(f)$, is the degree of the unique multilinear polynomial $P(X_1, \ldots, X_n) \in R[X_1, \ldots, X_n]$ such that $P(x) = f(x)$ for every $x \in \{0,1\}^n$, where $R = \mathbb{Z}/k\mathbb{Z}$.*

We say that the polynomial P represents f over R. The existence and uniqueness of such a representing polynomial follows from the Möbius inversion formula (see 6.2). Of particular importance in complexity theory are the cases $k = 0$ ($R = \mathbb{Z}$) and $k = p$ ($R = \mathbb{F}_p$) for some prime p; these will also be our primary focus, though we will also consider the case of composite m. We denote $\deg_0(f)$ simply by $\deg(f)$; it also equals the degree of the Fourier polynomial for the function $(-1)^{f(x)}$. Let us note a basic relation between these various degrees, namely that for every f and k, we have

$$\deg_k(f) \leq \deg(f) \ .$$

This is because the polynomial representing f over $\mathbb{Z}/k\mathbb{Z}$ can be obtained from the representation over $\mathbb{Z}$ by taking each coefficient modulo k. The gap between these quantities can be arbitrarily large; consider the function $\mathrm{Par}(x) = \sum_i x_i \bmod 2$. It is easy to show that $\deg(\mathrm{Par}) = n$ whereas $\deg_2(\mathrm{Par}) = 1$. Indeed, it is not hard to show that $\deg_p(\mathrm{Par}) = n$ for every prime $p \neq 2$.

In this paper, we show that this is an instance of a more general principle:

A function on all n variables which has low degree in characteristic p is bound to have high degree in every other prime characteristic $q \neq p$.

Moreover, we prove that any function f where $\deg_p(f) = o(\log n)$ is hard to approximate by low-degree polynomials modulo q, and hence requires large $\mathrm{AC}_0[q]$ circuits.

6.1.1 Our Results

When we refer to Boolean functions on n variables, we only consider functions where all n variables are influential. This rules out trivial counterexamples like k-juntas that have low degree in all characteristics. The following is our main theorem:

Theorem 6.1. *(Main) Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function which depends on all n variables. Let $p \neq q$ be distinct primes. Then*

$$\deg_q(f) \geq \frac{n}{\lceil \log_2 p \rceil \deg_p(f) p^{2\deg_p(f)}} \ .$$

This gives a lower bound of $\Omega(n^{1-o(1)})$ on $\deg_q(f)$ as long as $\deg_p(f) = o(\log n)$. This bound is close to the best possible, as there exist functions on all n variables (such as the addressing function [NS92]) where $\deg(f) \leq \log n$ and hence $\deg_p(f) \leq \log n$ for all characteristics p. Thus, one cannot get nontrivial lower bounds on $\deg_q(f)$ once $\deg_p(f)$ exceeds $\log n$.

Nisan and Szegedy showed that any function on n variables must have degree at least $\deg(f) \geq \log n - O(\log \log n)$ [NS92]. An interesting consequence of 6.1 is the following analog of the Nisan-Szegedy bound for non-prime power moduli.

Corollary 6.1. *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function which depends on all n variables. Suppose m is not a prime power, and p is its smallest prime divisor. We have*

$$\deg_m(f) \geq \frac{1}{2}\log_p n - \log_p \log_p n - \frac{1}{2}\log_p \lceil \log_2 p \rceil \ .$$

This corollary is interesting as it illuminates a sharp difference between degrees over composite numbers and over primes. A simple way to construct Boolean functions of degree $O(1)$ over $\mathbb{F}_p$ is to take any constant degree polynomial $P(x_1, \ldots, x_n) \in \mathbb{F}_p[x_1, \ldots, x_n]$ and raise it to the power $p-1$. This construction fails for composite m since there is no analog of Fermat's little theorem. 6.1 shows that indeed any polynomial modulo m computing a Boolean function requires degree $\Omega(\log n)$, as it does over the reals.

While 6.1 immediately implies a lower bound for $\deg(f)$, one can obtain the following stronger bound by a simple modification of the Nisan-Szegedy proof:

Lemma 6.1. *Let p be a prime and $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function which depends on all n variables. Then*

$$\deg(f) \geq \frac{n}{2^{\deg_p(f)}} \ .$$

We prove this lemma in 6.2.1.

The results above show a very basic relation between the degrees of Boolean functions over different characteristics. A natural question to ask is what happens if we relax the requirement and only consider polynomials over $\mathbb{F}_q$ that approximate a low degree polynomial over $\mathbb{F}_p$. However, similarly to the case of degree 1 polynomials that was studied in [Smo87], we prove that low degree polynomials modulo p are hard to even approximate by polynomials in other characteristics.

Theorem 6.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be a function depending on all n variables with $\deg_p(f) = d$. Then, for any $q \neq p$ and any $\mathbb{F}_q$ polynomial $Q(x_1, \ldots, x_n) : \mathbb{F}_q^n \to \{0,1\}$, satisfying $\deg_q(Q) = o\left(\sqrt{\frac{n}{dp^{3d}}}\right)$, it holds that*

$$\Pr_{x \in \{0,1\}^n}[f(x) = Q(x)] \leq 1 - \epsilon p^{-d} \ ,$$

where ϵ depends only on p, q.

We note that both the error bound of $1 - p^{-O(d)}$ and the degree bound of $o(\sqrt{n})$ are close to optimal; there are polynomials of degree d over $\mathbb{F}_p$ that are 0 on the boolean hypercube with probability $1 - 2^{-d}$, hence they have trivial approximations over $\mathbb{F}_q$. Secondly, the Mod_p function (and indeed every symmetric function) can be $1-\epsilon$ approximated by polynomials of degree $c(\epsilon)\sqrt{n}$ over $\mathbb{F}_q$ [BGL06], despite being hard to approximate for polynomials of lower degree.

As a corollary of 6.2 we get that if a Boolean function has low degree modulo p, then the function requires large $\mathrm{AC}_0[q]$ circuits for any prime $q \neq p$. Several of the known lower bounds for $\mathrm{AC}_0[q]$ are for functions like Par and the Mod_{p^k} function where $p \neq q$ that are easily seen to be low-degree polynomials in some characteristic. Our result generalizes this to give a very general class of hard functions for $\mathrm{AC}_0[q]$, namely all functions that have degree $o(\log n)$ modulo $p \neq q$.

Theorem 6.3. *Let p, q be distinct primes. Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function which depends on all n variables with $\deg_p(f) = o(\log_p n)$. Then any $\mathrm{AC}_0[q]$ circuit of depth t computing f requires size at least $\exp(n^{(1-o(1))/2t})$.*

It is not hard to see that most known lower bounds for $AC_0[q]$ circuits follow from the theorem above. For example, the lower bound for Mod_{p^k} of [Smo87] follows from the observation that $\deg_p(\mathsf{Mod}_{p^k}) \leq p^k$ (see e.g. [BGL06]). Additionally, it gives several new lower bounds, for instance it shows that every quadratic form on n variables over $\mathbb{F}_2$ requires large $\mathrm{AC}_0[q]$ circuits, for $q \neq 2$. Though we note that 6.3 does not imply Razborov's lower bound for $\mathsf{Majority}$.

Summarizing, Theorems 6.1 and 6.2 show that for a Boolean function, having low degree mod p, or even being close to a low degree polynomial mod p, is a "singular" event, in the sense it can only occur for at most one characteristic p.

6.1.2 Polynomial representations in computer science

The study of polynomial representations of Boolean functions dates at least as far back as the 1960's, when they arose in various contexts including switching theory [Mur71], voting theory [Cho61] and machine learning [MP68]. Representations of Boolean functions over finite fields, especially over $\mathbb{F}_2$ were studied by coding theorists in the context of Reed-Muller codes, see [MS83, Chapters 13-14] and the references therein. The codewords of the code $\mathrm{RM}_2(d, n)$ are all Boolean functions $f : \{0,1\}^n \to \{0,1\}$ where $\deg_2(f) \leq d$, while received words are arbitrary functions f.

Polynomial representations have proved especially useful in circuit complexity [Bei93] where a natural lower bound technique is to relate concrete complexity measures (such as circuit-size) which we wish to bound, to purely algebraic complexity measures. Examples of this paradigm include the Razborov-Smolensky lower bounds for $\mathrm{AC}_0[p]$ [Raz87, Smo87], which relates the circuit size to the polynomial degree needed to approximate f over $\mathbb{F}_p$, and the work of Beigel et al. [BRS] and Aspnes et al. [ABFR94] which relate AC_0 circuit size with approximations by real polynomials.

Polynomial representations are among the most powerful tools in computational learning. The best learning algorithms for many basic concept classes, including but not limited to decision trees [KM93], DNF formulae [KS01], AC_0 circuits [LMN93, JCJS02], juntas [MOS03] and halfspaces [ARKS, KKMS05] all proceed by showing that the concept class to be learned has some *nice* polynomial representation. In particular, the algorithm for learning juntas of [MOS03] exploits a connection between $\deg_2(f)$ and the sparsity of its Fourier polynomial.

Finally, polynomial representations of Boolean functions have found applications to constructing combinatorial objects such as set systems [Gro00, Gro02], Ramsey graphs

[Gro00, Gop06b] and locally decodable codes [Efr09]. These results require low-degree *weak* representations of simple Boolean functions like the Or function but modulo composites.

Definition 6.2. *The polynomial $P(x_1, \ldots, X_n) \in \mathbb{Z}[X_1, \ldots, X_n]/m\mathbb{Z}$ weakly represents $f : \{0,1\}^n \to \{0,1\}$ over $\mathbb{Z}/m\mathbb{Z}$ if $f(x) \neq f(y) \Rightarrow P(x) \neq P(y)$ ($P(x)$ may take values in $\mathbb{Z}/m\mathbb{Z}$).*

Such representations have been well studied in complexity theory (see [BBR94, BGL06] and the references therein), but embarrassingly simple questions like the degree required to represent the Or function mod 6 remain open, there is a gap of $O(\sqrt{n})$ [BBR94] versus $\Omega(\log n)$ [TB98] between upper and lower bounds. Better upper bounds would lead to improved constructions of all the above combinatorial objects. In [Gop06b], Gopalan proposes viewing this as a question about the degree of two related functions in distinct characteristics:

Problem 6.1. *[Gop06b] If two functions $f, g : \{0,1\}^n \to \{0,1\}$ satisfy $f(x) \vee g(x) = \mathsf{Or}(x)$, how small can $\max(\deg_2(f), \deg_3(g))$ be?*

Questions like this emphasize the importance of the natural and basic question of understanding the behavior of $\deg_p$ for various characteristics p.

6.1.3 Techniques

Our proofs are conceptually very simple, we reduce the degree d case to the linear case and then appeal to known lower bounds. This reduction is carried out via a degree reduction lemma (6.4) that shows that for any degree d polynomial $P(x)$ over $\mathbb{F}_p$ on n variables, there exist a constant t and a linear combination of the form

$$P'(x) = \sum_{i \leq t} \lambda_i P(x + a_i) \quad \lambda_i \in \mathbb{F}_p, \ a_i \in \mathbb{F}_p^n$$

so that by fixing some variables in P' to constants, we get a linear polynomial in many variables. This lemma is proved using discrete derivatives, a notion that has proved very useful lately in complexity theory [BV07, Lov08, Vio08].

With this lemma in hand, one would like to proceed as follows: suppose $P(x)$ and $Q(x)$ represent the same function f over $\mathbb{F}_p$ and $\mathbb{F}_q$, and that $P(x)$ has low degree (say a constant). The polynomial $P'(x)$ is tightly related to the Mod_p function, which is known to require high degree in characteristic q. We would like to claim that the degree of $P'(x)$ over $\mathbb{F}_q$ is a small multiple of $\deg(Q)$, which would then imply that $\deg(Q)$ must be large. Implementing this scheme runs into an obstacle: P' is a function that maps $\mathbb{F}_p^n \to \mathbb{F}_p$, further the values a_i are from $\mathbb{F}_p^n$, thus while $P(x) = Q(x)$ for $x \in \{0,1\}^n$, it is unclear how $Q(x)$ can help us evaluate $P(x + a_i)$.

Most of the technical work in this paper goes towards circumventing this obstacle. The special case of $p = 2$ is easier to handle, as since $\{0,1\} \subset \mathbb{F}_q$ one can mimic operations modulo 2 in characteristic $\mathbb{F}_q$ without a large overhead. we present the case of characteristic 2 separately in 6.4. For $p > 2$, we show that one can still mimic differentiation modulo p in characteristic q without a large blowup in the degree, however the argument is more complicated. We present the general case in 6.5.

6.2 Preliminaries

Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. We will only consider Boolean functions that depend on all n variables, meaning that they cannot be written as $f(x_1, \ldots, x_n) = g(x_{i_1}, \ldots, x_{i_k})$ for some $k < n$. We start by establishing the correspondence between functions and polynomials. We state the correspondence in the general setting of any commutative ring R containing $\{0,1\}$, but we will only be interested in the cases where R is either $\mathbb{Z}$, $\mathbb{Z}/m\mathbb{Z}$ for some integer m or a finite field $\mathbb{F}_q$. We say that a polynomial $P(x_1, \ldots, x_n) \in R[x_1, \ldots, x_n]$ computes the function f if $P(x) = f(x)$ for all $x \in \{0,1\}^n$. While there could be many polynomials that satisfy this condition, if we insist that the polynomial be multilinear (every variable occurs with degree at most 1), then the polynomial is unique. This can be seen via the Möbius inversion formula, which gives a unique multilinear polynomial $P(x_1, \ldots, x_n) \in R[x_1, \ldots, x_n]$ satisfying $P(x) = f(x)$ for every function $f : \{0,1\}^n \to R$:

$$P(x) = \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i$$
$$\text{where } \; c_S = \sum_{x \leq x(S)} (-1)^{|S| - \mathrm{wt}(x)} f(x)$$

where $x(S)$ denotes the indicator vector of the set S, $x \leq x(S)$ denotes that $x_i \leq x(S)_i$ for every coordinate i and $\mathrm{wt}(x)$ denotes the Hamming weight of the vector x. If f is Boolean, the Möbius inversion shows that the representing polynomial depends only on the characteristic of R.

We state some basic facts about $\deg_k(f)$, proofs of which can be found in [Gop06a]. The multilinear polynomial computing f over $\mathbb{Z}/m\mathbb{Z}$ can be obtained by reducing each coefficient of the polynomial computing f over $\mathbb{Z}$ modulo m, which gives the following:

Fact 6.1. *For any $f : \{0,1\}^n \to \{0,1\}$, we have $\deg_m(f) \leq \deg(f)$ for all m. Similarly if $m_1 | m$, then $\deg_{m_1}(f) \leq \deg_m(f)$.*

A consequence of this inequality is that $\deg_m(f) \leq \deg_{m^k}(f)$. The following folklore lemma shows that they are always within a factor $2k$ of each other.

Fact 6.2. *For any $f : \{0,1\}^n \to \{0,1\}$, and integers m, k:*

$$\deg_m(f) \leq \deg_{m^k}(f) \leq (2k-1) \deg_m(f) \; .$$

If $m = m_1 m_2$ where $(m_1, m_2) = 1$, then the multilinear polynomial $P(x) \in \mathbb{Z}[x]/m\mathbb{Z}$ is obtained by combining the coefficients of $P_1(x) \in \mathbb{Z}[x]/m_1\mathbb{Z}$ and $P_2(x) = \mathbb{Z}[x]/m_2\mathbb{Z}$ by the Chinese Remainder Theorem. Hence

Fact 6.3. *Let $m = m_1 m_2$ where $(m_1, m_2) = 1$. Then*

$$\deg_m(f) = \max(\deg_{m_1}(f), \deg_{m_2}(f)) \; .$$

Thus if we know $\deg_p(f)$ for all primes p that divide m, we can use 6.2 and 6.3 to estimate $\deg_m(f)$ up to a constant factor which is independent of n but depends on m.

We define the function $\mathsf{Mod}_m(x)$ to be 1 whenever $\sum_i x_i$ is divisible by m. The degree of such functions in any characteristic can be computed using the following observation:

Fact 6.4. *For any integer k, and primes $p \neq q$, we have*

$$\deg_p(\mathsf{Mod}_{p^k}) = p^k, \quad \deg_q(\mathsf{Mod}_{p^k}) = \Omega(n) \ .$$

Finally, we use two lemmas from the work of Razborov and Smolensky showing that if a Boolean function f can be computed by a small $\mathrm{AC}_0[p]$ circuit, then f can be well approximated by low degree polynomials over $\mathbb{F}_p$. The first is their low-degree approximation lemma for $\mathrm{AC}_0[p]$ circuits.

Lemma 6.2 ([Raz87, Smo87])**.** *For a prime p, let f be a Boolean function on n variables that is computed by an $\mathrm{AC}_0[p]$ circuit of size s and depth t. For every $\delta > 0$, there exists a polynomial $P \in \mathbb{F}_p[x_1, \ldots, x_n]$ of degree $\deg(P) \leq (cp \log(s/\delta))^t$ such that $P(\{0,1\}^n) \subset \{0,1\}$ and*

$$\Pr_{x \in \{0,1\}^n}[P(x) = f(x)] \geq 1 - \delta$$

for some absolute constant c.

The second lemma shows that the Mod_p function does not have such an approximation over $\mathbb{F}_q$.

Lemma 6.3 ([Raz87, Smo87])**.** *For every two primes $p \neq q$, there exist constants $c, \epsilon > 0$ depending only on p, q such that for any polynomial $Q(x)$ over $\mathbb{F}_q$ of degree at most $c\sqrt{n}$,*

$$\Pr_{x \in \{0,1\}^n}[Q(x) = \mathsf{Mod}_p(x)] < 1 - \epsilon \ .$$

We do not care about exact constants in this paper, unless otherwise specified. Hence, to simplify notation we denote constants by c, where we specify whether these are absolute constants or depending on some other parameters (i.e. ϵ, p, q). In all cases constants do not depend on the number of variables n.

6.2.1 Proof of 6.1

For completeness we give the simple proof of 6.1. The proof follows the Nisan-Szegedy argument, which gives upper and lower bounds on the average sensitivity of the Boolean function in terms of $\deg(f)$. We observe that the lower bound holds in any characteristic (but the upper bound holds only for characteristic 0).

Proof of 6.1. Let us define $\mathbb{I}\mathrm{nf}_i(f) = \Pr_{x \in \{0,1\}^n}[f(x) \neq f(x \oplus e_i)]$ where $x \oplus e_i$ denotes x with the i^{th} bit flipped. A simple application of the Schwartz-Zippel lemma shows that

$$\mathbb{I}\mathrm{nf}_i(f) \geq \frac{1}{2^{\deg_p(f)}} \quad \text{hence} \quad \sum_{i \leq n} \mathbb{I}\mathrm{nf}_i(f) \geq \frac{n}{2^{\deg_p(f)}} \ .$$

But by Corollary 1 in [NS92],

$$\sum_{i \leq n} \mathbb{I}\mathrm{nf}_i(f) \leq \deg(f)$$

which gives the required bound. □

6.3 Degree Reduction

A crucial tool in our proofs is the following *Degree reduction lemma* that reduces degree d polynomials in n variables to polynomials with many linear terms. For a polynomial P define the set $L(P)$ to be those variables x_i appearing as linear terms in P but not in any of its higher degree monomials.

Lemma 6.4 (Degree Reduction Lemma)**.** *Let $P(x)$ be a polynomial of degree d over $\mathbb{F}_p$, depending on all n variables, such that the individual degree of each variable is at most $p-1$. Then there exist $t \leq p^{\lceil \frac{d-1}{p-1} \rceil}$, $a_1, \ldots, a_t \in \mathbb{F}_p^n$, and $\lambda_1, \ldots, \lambda_t \in \mathbb{F}_p$ such that the polynomial*

$$Q(x) = \sum_{i \leq t} \lambda_i P(x + a_i)$$

satisfies

$$|L(Q)| \geq \frac{n}{dp^{\lceil \frac{d-1}{p-1} \rceil}} .$$

The reminder of this section is dedicated to the proof of 6.4. The main idea used is that if $P(x)$ is a homogeneous degree d polynomial, then taking $d-1$ directional derivatives of P along random directions will yield with high probability a polynomial with many linear variables. In the non-homogenous case, we have to choose how many times to differentiate carefully, since for example if the polynomial is $X_1X_2 + X_3 + X_4 \cdots + X_n$, then most of the variables will disappear after differentiating just once. To get a large linear form from this polynomial however, we can simply set $X_1 = X_2 = 0$. Our final degree reduction procedure combines these two strategies, we first differentiate and then set some variables to 0 to get a large linear form.

Finally, for technical reasons, we differentiate multiple times along each direction rather than choosing multiple directions. While this makes the proof of the degree reduction more involved, it allows us to get a better dependence on the degree. Roughly speaking, we can show that $\deg_q(f) \geq \frac{n}{p^{\deg_p(f)}}$, whereas differentiating once along multiple directions would yield bounds of the form $\deg_q(f) \geq \frac{n}{2^{p \deg_p(f)}}$ with our proof technique.

We define the *monomial degree* of a variable x_i in a polynomial $P(x)$ to be the maximal degree of a monomial of P containing x_i, and denote it by $\deg_i(P)$. Note that the monomial degree of x_i is different from its individual degree, which is the highest power of x_i that occurs in P. The main tool we use to prove the lemma is the notion of directional derivatives of a polynomial. Given a polynomial P, we define the first derivative along y, denoted $P_{(y,1)}$, as

$$P_{(y,1)}(x) = P(x + y) - P(x) .$$

We define the ℓ^{th} derivative along y for $\ell \geq 1$ inductively as

$$P_{(y,\ell)}(x) = P_{(y,\ell-1)}(x + y) - P_{(y,\ell-1)}(x)$$

when $\ell \geq 1$. It is easy to verify that

$$P_{(y,\ell)}(x) = \sum_{0 \leq j \leq \ell} (-1)^{\ell - j} \binom{\ell}{j} P(x + jy) .$$

We define multiple derivatives in multiple directions, which we denote by $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}(x)$. To derive a formula for those derivatives we define the following quantity for all ℓ, c:

$$\mu(\ell, c) = \sum_{0 \le j \le \ell} (-1)^{\ell - j} \binom{\ell}{j} j^c \ .$$

The following combinatorial identities are well-known; we prove them for completeness:

Fact 6.5. *Let $\ell \le p-1$. Then*

$$\begin{aligned} &\mu(\ell, c) = 0 \text{ for } c \in \{0, \ldots, \ell - 1\} \ , \\ &\mu(\ell, \ell) \not\equiv 0 \mod p \ . \end{aligned}$$

Proof. We prove the first identity by induction on c. The case $c = 0$ is elementary. To prove it for $c \ge 1$, we consider the following identity over $\mathbb{Z}$

$$(X-1)^\ell = \sum_{0 \le j \le \ell} (-1)^{\ell - j} \binom{\ell}{j} X^j \ . \tag{6.1}$$

Differentiating both sides $c \le \ell - 1$ times and then setting $X = 1$ gives

$$\begin{aligned} 0 &= \sum_{0 \le j \le \ell} (-1)^{\ell - j} \binom{\ell}{j} j(j-1)\cdots(j-c+1) \\ &= \mu(\ell, c) + \sum_{1 \le i \le c-1} \lambda(i)\mu(\ell, i) \ , \end{aligned}$$

where the $\lambda(i)$-s are some integers. Using the induction hypothesis for $i \le c - 1$ gives $\mu(\ell, c) = 0$. To prove $\mu(\ell, \ell) \not\equiv 0 \mod p$ we differentiate Equation 6.1 ℓ times to get

$$\begin{aligned} \ell! &= \sum_{0 \le j \le \ell} (-1)^{\ell - j} \binom{\ell}{j} j(j-1)\cdots(j-\ell+1) \\ &= \mu(\ell, \ell) + \sum_{1 \le c \le \ell - 1} \lambda(c)\mu(\ell, c) \\ &= \mu(\ell, \ell) \ . \end{aligned}$$

Since we assume that $\ell \le p - 1$ it follows that $\mu(\ell, \ell) = \ell! \not\equiv 0 \bmod p$. □

We abbreviate the monomial $\prod_{i=1}^n x_i^{d_i}$ by x^d where $d = (d_1, \cdots, d_n)$ is the degree vector. We use $|d| = \sum_i d_i$ to denote its total degree. Given vectors d, e we say $e \le d$ if $e_i \le d_i$ for

all i, and use the notation $\binom{d}{e} = \prod_i \binom{d_i}{e_i}$. We have

$$\begin{aligned} x^d_{(y,\ell)} &= \sum_{j=0}^{\ell} (-1)^{\ell-j} \binom{\ell}{j} (x+jy)^d \\ &= \sum_{j=0}^{\ell} (-1)^{\ell-j} \binom{\ell}{j} \sum_{e \le d} \binom{d}{e} x^{d-e} (jy)^e \\ &= \sum_{e \le d} \binom{d}{e} x^{d-e} y^e \sum_{j=0}^{\ell} (-1)^{\ell-j} \binom{\ell}{j} j^{|e|} \\ &= \sum_{e \le d} \binom{d}{e} x^{d-e} y^e \mu(\ell, |e|) \\ &= \sum_{\substack{e \le d \\ |e| \ge \ell}} \binom{d}{e} x^{d-e} y^e \mu(\ell, |e|) \end{aligned}$$

where we use $\mu(\ell, |e|) = 0$ for $|e| \le \ell - 1$. Thus, differentiating ℓ times along y reduces the degree in x by at least ℓ, as one would expect.

By repeating this calculation, we can compute an expression for derivatives in multiple directions. Given vectors $d, e^{(1)}, \ldots, e^{(k)}$ we use the notation $\binom{d}{e^{(1)}, \ldots, e^{(k)}}$ for the product of multinomials $\prod_{l \in [n]} \binom{d_l}{e_l^{(1)}, \ldots, e_l^{(k)}}$. We have

$$\begin{aligned} &x^d_{(y^{(1)}, \ell^{(1)}), \ldots, (y^{(k)}, \ell^{(k)})} = \\ &\sum_{e^{(1)} + \cdots + e^{(k)} \le d} \binom{d}{e^{(1)}, \ldots, e^{(k)}} x^{d - (e^{(1)} + \cdots + e^{(k)})} \cdot \prod_{j=1}^{k} \mu(\ell^{(j)}, |e^{(j)}|) (y^{(j)})^{e^{(j)}} = \\ &\sum_{|e^{(1)}| \ge \ell^{(1)}, \ldots, |e^{(k)}| \ge \ell^{(k)}} \binom{d}{e^{(1)}, \ldots, e^{(k)}} x^{d - (e^{(1)} + \cdots + e^{(k)})} \cdot \prod_{j=1}^{k} \mu(\ell^{(j)}, |e^{(j)}|) (y^{(j)})^{e^{(j)}} \ . \end{aligned}$$

By linearity, we can compute the derivative of any polynomial $P(x) = \sum_d c_d x^d$.

$$\begin{aligned} &P_{(y^{(1)}, \ell^{(1)}), \ldots, (y^{(k)}, \ell^{(k)})}(x) = \sum_d c_d x^d_{(y^{(1)}, \ell^{(1)}), \ldots, (y^{(k)}, \ell^{(k)})} = \\ &\sum_d c_d \sum_{|e^{(1)}| \ge \ell^{(1)}, \ldots, |e^{(k)}| \ge \ell^{(k)}} \binom{d}{e^{(1)}, \ldots, e^{(k)}} x^{d - (\sum_j e^{(j)})} \cdot \prod_{j=1}^{k} \mu(\ell^{(j)}, |e^{(j)}|) (y^{(j)})^{e^{(j)}} = \\ &\sum_f x^f \left(\sum_{|e^{(1)}| \ge \ell^{(1)}, \ldots, |e^{(k)}| \ge \ell^{(k)}} c_{f + \sum_j e^{(j)}} \binom{f + \sum_j e^{(j)}}{e^{(1)}, \ldots, e^{(k)}} \cdot \prod_{j=1}^{k} \mu(\ell^{(j)}, |e^{(j)}|) (y^{(j)})^{e^{(j)}} \right) \end{aligned} \tag{6.2}$$

where in the last line we use the change of variable $f = d - \sum_j e^{(j)}$. Recall that we define $\deg_i(P)$ to be the largest degree monomial containing the variable x_i. It follows that the

monomial degree of x_i drops by at least $\min(\sum_j \ell^{(j)}, \deg_i(P))$ (note that the degree cannot drop below zero):

$$\deg_i(P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}) \leq \deg_i(P) - \min(\sum_j \ell^{(j)}, \deg_i(P)) \ .$$

Lemma 6.5. *Let*

$$\deg_i(P) = (k-1)(p-1) + \ell + 1 \quad \textit{where} \quad \ell + 1 \leq p-1 \ ,$$
$$\ell^{(1)} = \cdots = \ell^{(k-1)} = p-1 \text{ and } \ell^{(k)} = \ell \ .$$

Then the coefficient of x_i in $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}(x)$ is a non-zero polynomial in $y^{(1)}, \ldots, y^{(k)}$.

Proof. Observe that $\sum_j \ell^{(j)} = \deg_i(P) - 1$, so

$$\deg_i(P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}) \leq \deg_i(P) - \sum_j \ell^{(j)} = 1 \ .$$

Our goal is to show that it is in fact 1. Consider the vector f where $f_i = 1$ and $f_j = 0$ for all $j \neq i$. By Equation 6.2, the coefficient of x^f in $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}(x)$ is given by

$$c'_f = \sum_{|e^{(1)}| \geq \ell^{(1)}, \ldots, |e^{(k)}| \geq \ell^{(k)}} c_{f+\sum_j e^{(j)}} \binom{f + \sum_j e^{(j)}}{e^{(1)}, \ldots, e^{(k)}} \cdot \prod_{j=1}^{k} \mu(\ell^{(j)}, |e^{(j)}|)(y^{(j)})^{e^{(j)}} \ . \tag{6.3}$$

We shall now find $e^{(1)}, \ldots, e^{(k)}$ so that the following conditions hold:

$$c_{f+\sum_j e^{(j)}} \neq 0, \quad \binom{f + \sum_j e^{(j)}}{e^{(1)}, \ldots, e^{(k)}} \neq 0 \tag{6.4}$$
$$|e^{(1)}| = \cdots = |e^{(k-1)}| = p-1, |e^{(k)}| = \ell \ . \tag{6.5}$$

Indeed, Equation 6.5 ensures that $\mu(\ell_j, |e^{(j)}|) \neq 0$. By Equation 6.4 each solution $(e^{(1)}, \cdots, e^{(k)})$ will contribute a non-zero multiple of the monomial $\prod_{j=1}^{k}(y^{(j)})^{e^{(j)}}$ to c'_f. Notice that distinct solutions contribute distinct monomials to the right hand side of 6.3. Hence, the claim will follow if we show that there is at least one choice of $e^{(1)}, \ldots, e^{(k)}$ satisfying Equations 6.4,6.5.

Fix a monomial x^d, where $|d| = \deg_i(P)$ and $c_d \neq 0$, containing the variable x_i. Now $|d - f| = (k-1)(p-1) + \ell$. It is easy to define $e^{(1)}, \ldots, e^{(k)}$ so that

$$|e^{(1)}| = \cdots = |e^{(k-1)}| = p-1, |e^{(k)}| = \ell$$

and

$$\sum_j (e^{(j)})_l + f_l = d_l \quad \forall \ l \in [n] \ .$$

Note that

$$\binom{f+\sum_j e^{(j)}}{e^{(1)},\ldots,e^{(k)}} = \prod_{l\in[n]} \binom{f_l+\sum_j (e^{(j)})_l}{(e^{(1)})_l,\ldots,(e^{(k)})_l} .$$

As

$$\sum_j (e^{(j)})_l \le f_l + \sum_j (e^{(j)})_l = d_l \le p-1 ,$$

each binomial coefficient in the product is non-zero mod p. This gives a solution satisfying both Equations 6.4 and 6.5. □

Let $\delta_p(d)$ denote the minimum probability that a nonzero degree d polynomial over $\mathbb{F}_p$ evaluates to zero on a random input. It is well-known (see e.g. [MS83]) that if $d = a(p-1)+b$ where $a \ge 0$ and $b \le p-1$, then

$$\delta_p(d) = \frac{1}{p^a}\left(1-\frac{b}{p}\right) \ge p^{-\lceil \frac{d}{p-1} \rceil} .$$

Lemma 6.6. *Let $P(x) \in \mathbb{F}_p[x]$ be a degree d polynomial that depends on all n variables. Then there exist $k \le \left\lceil \frac{d-1}{p-1} \right\rceil$, directions $y^{(1)},\ldots,y^{(k)} \in \mathbb{F}_p^n$ and integers $\ell^{(1)},\ldots,\ell^{(k)} \le p-1$ such that*

$$|L(P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})})| \ge \frac{n}{dp^{\lceil \frac{d-1}{p-1} \rceil}} .$$

Proof. There exists some $d' \le d$ so that $\deg_i(P) = d'$ for at least $\frac{n}{d}$ variables, call this set of variables G. If $d' = 1$, then the claim trivially holds, so assume $d' > 1$. Let $d'-1 = (k-1)(p-1)+\ell$ for $\ell \le p-2$ and set $\ell^{(1)} = \cdots = \ell^{(k-1)} = p-1, \ell^{(k)} = \ell$. 6.5 implies that, for every $x_i \in G$, the coefficient $c_i(y^{(1)},\ldots,y^{(k)})$ of x_i in $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}$ is a non-zero polynomial of degree at most $d'-1 \le d-1$ in $y^{(1)},\ldots,y^{(k)}$. Thus, there exists a setting for $y_1,\ldots,y_k$ where at least

$$\delta_p(d-1)|G| \ge \frac{n}{dp^{\lceil \frac{d-1}{p-1} \rceil}}$$

of the c_is are non-zero. Since variables in G have degree 1 in $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}$, there are no higher degree terms which contain them, so these variables all lie in $L(P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})})$. □

To complete the proof of 6.4, we observe that $P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}$ can be written as

$$P_{(y^{(1)},\ell^{(1)}),\ldots,(y^{(k)},\ell^{(k)})}(x) = \sum_{i\le t} \lambda_i P(x+a_i)$$

$$\text{where } \; t \le \prod_{j=1}^{k} (\ell^{(j)}+1) \le p^{\lceil \frac{d-1}{p-1} \rceil} .$$

6.4 The case of characteristic 2

Let $P(x)$ be a low degree polynomial over $\mathbb{F}_2$. We prove in this section that P must have high degree over characteristics $q \neq 2$. Since we will be working with operations over different fields, we will use $+$ to denote summation modulo q, and $\oplus$ for summation modulo 2. We start with some simple claims:

Claim 6.1. *Let $f(x) = \oplus_{i=1}^{n} x_i$ be the parity function on n bits. Then for $q \neq 2$, $\deg_q(f) = n$.*

Proof. The unique multilinear polynomial over $\mathbb{F}_q$ computing f is

$$H^{\oplus}(x) = \frac{1}{2}(1 - \prod_{i=1}^{n}(1 - 2x_i))$$

□

Lemma 6.7. *Let $a_1, \ldots, a_k \in \mathbb{F}_2^n$. Define $g : \{0,1\}^n \to \{0,1\}$ by $g(x) = \oplus_{i=1}^{k} f(x \oplus a_i)$. Then*

$$\deg_q(g) \leq k \deg_q(f) .$$

Proof. For any $a \in \mathbb{F}_2^n$, consider $f_a(x) = f(x \oplus a)$. Clearly, $g(x) = \oplus_{i=1}^{k} f_{a_i}(x)$. We claim that $\deg_q(f_a) = \deg_q(f)$. Let $Q(x)$ be a polynomial over $\mathbb{F}_q$ which computes f over $\{0,1\}^n$. Define a new polynomial $Q_a(x) = Q(x \oplus a)$ by replacing x_i with $1 - x_i$ whenever $a_i = 1$, and keeping x_i whenever $a_i = 0$. Clearly Q_a computes $f_a(x)$ over $\{0,1\}^n$, and $\deg_q(Q_a) = \deg_q(Q)$.

Composing the polynomial $H^{\oplus}$ over $\mathbb{F}_q$ that computes $\oplus$ on $\{0,1\}^k$ with the Q_a-s, we get a polynomial of degree at most $k \deg_q(f)$ that represents g over $\mathbb{F}_q$. Hence, $\deg_q(g) \leq k \deg_q(f)$. □

We now restate and prove 6.1 in the $p = 2$ case, showing that any Boolean function with small degree over $\mathbb{F}_2$ must have high degree over $\mathbb{F}_q$ for a prime $q \neq 2$.

Theorem 6.4 (6.1, $p = 2$ case)**.** *For any $f : \{0,1\}^n \to \{0,1\}$, and prime $q \neq 2$:*

$$\deg_q(f) \geq \frac{n}{\deg_2(f) 4^{\deg_2(f)}} .$$

Proof. Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function such that $\deg_2(f) = d$. Let $P(x)$ be the degree d polynomial over $\mathbb{F}_2$ computing f. We will prove that the multilinear polynomial $Q(x)$ over $\mathbb{F}_q$ computing f has high degree.

By 6.4, there exist $a_1, \ldots, a_k \in \mathbb{F}_2^n$, for $k \leq 2^d$, such that if $\tilde{P}(x) = \oplus_{i=1}^{k} P(x \oplus a_i)$, then $|L(\tilde{P})| \geq \frac{n}{d2^d}$. Let us denote the set $L(\tilde{P})$ by S. Let $\tilde{P}_S$ be the restriction of $\tilde{P}$ to the variables in S obtained by fixing the remaining variables to zero. Clearly, $\tilde{P}_S(x)$ is either Par on the set S or its negation. Assume w.l.o.g it is the former.

Now consider the polynomial Q. Since $Q(x) = f(x)$ for all $x \in \{0,1\}^n$, then the polynomial $\tilde{Q}$ defined as $\tilde{Q}(x) = H^{\oplus}(Q(x \oplus a_1), \ldots, Q(x \oplus a_k))$ satisfies that $\tilde{Q}(x) = \tilde{P}(x)$ for all $x \in \{0,1\}^n$. So if we let $\tilde{Q}_S$ be the restriction of $\tilde{Q}$ to the variables in S, then $\tilde{Q}_S(x) = \tilde{P}_S(x)$ for all $x \in \{0,1\}^n$.

Now, since $\tilde{P}_S$ is the parity function over $|S|$ bits, 6.1 implies that $\deg(\tilde{Q}_S) = |S| \geq \frac{n}{d2^d}$. On the other hand, by 6.7 we have that $\deg(\tilde{Q}_S) \leq \deg(\tilde{Q}) \leq k \deg_q(f)$. Therefore we conclude that

$$\deg_q(f) \geq \frac{n}{kd2^d} \geq \frac{n}{d4^d} .$$

□

We now generalize this result and show that f cannot be approximated by low degree polynomials over $\mathbb{F}_q$. We need the following claim, which is proven using the union bound.

Claim 6.2. *Let $f' : \{0,1\}^n \to \{0,1\}$ be such that $\Pr_{x\in\{0,1\}^n}[f'(x) = f(x)] \geq 1 - \epsilon$. Let $a_1, \ldots, a_k \in \mathbb{F}_2^n$. Then*

$$\Pr_{x\in\{0,1\}^n}[\oplus_{i=1}^k f'(x \oplus a_i) = \oplus_{i=1}^k f(x \oplus a_i)] \geq 1 - k\epsilon .$$

We now restate and prove 6.2 in the $p = 2$ case.

Theorem 6.5 (6.2, $p = 2$ case)**.** *For a prime $q \neq 2$ let $c, \epsilon > 0$ be given by 6.3. Let $f : \{0,1\}^n \to \{0,1\}$ be of degree $\deg_2(f) = d$. If $h : \mathbb{F}_q^n \to \mathbb{F}_q$ satisfies*

$$\Pr_{x\in\{0,1\}^n}[h(x) = f(x)] \geq 1 - 2^{-d}\epsilon ,$$

then

$$\deg_q(h) \geq c\sqrt{\frac{n}{d8^d}} .$$

Proof. Using 6.4, choose $k \leq 2^d$ and $a_1, \ldots, a_k \in \mathbb{F}_2^n$ so that there exists a set of variables S of size $|S| \geq \frac{n}{d2^d}$ such that the function $\tilde{f}(x) = \oplus_{i=1}^k f(x \oplus a_i)$ is either Par or its negation when restricted to the variables in S. Similarly, define $\tilde{h}(x) = \oplus_{i=1}^k h(x \oplus a_k)$. By 6.2 we get that

$$\Pr_{x\in\{0,1\}^n}[\tilde{f}(x) = \tilde{h}(x)] \geq 1 - k2^{-d}\epsilon \geq 1 - \epsilon .$$

For every assignment $b \in \{0,1\}^{[n]\setminus S}$ to the variables outside S, define $\tilde{f}_{S,b}(x)$ as the restriction of $\tilde{f}$ to the variables in S, obtained by assigning values to the variables outside S according to b. Let $\tilde{h}_{S,b}$. We claim there exists some b such that

$$\Pr_{x\in\{0,1\}^S}[\tilde{f}_{S,b}(x) = \tilde{h}_{S,b}(x)] \geq 1 - \epsilon .$$

Indeed, this is true as for a randomly chosen b,

$$\mathbb{E}_{b\in\{0,1\}^{[n]\setminus S}}\left[\Pr_{x\in\{0,1\}^S}[\tilde{f}_{S,b}(x) = \tilde{h}_{S,b}(x)]\right] = \Pr_{x\in\{0,1\}^n}[\tilde{f}(x) = \tilde{h}(x)] \geq 1 - \epsilon .$$

We also have $\deg_q(\tilde{h}_{S,b}) \leq \deg_q(\tilde{h}) \leq 2^d \deg_q(h)$, where the last inequality uses 6.7. Now, $\tilde{f}_{S,b}(x)$ is either Par or its negation (assume w.l.o.g the former) over $|S|$ variables. Since $\tilde{h}_{S,b}$ approximates Par over $|S|$ variables with probability at least $1 - \epsilon$, 6.3 implies $\deg_q(\tilde{h}_{S,b}) \geq c\sqrt{|S|}$. Thus

$$2^d \deg_q(h) \geq \deg(\tilde{h}_{S,b}) \geq c\sqrt{\frac{n}{d2^d}}$$

which proves the theorem. □

Combining 6.5 with the Razborov-Smolensky bound, we conclude that any $\mathrm{AC}_0[q]$ circuit that computes a low $\mathbb{F}_2$-degree Boolean function on n variables must be of exponential size.

Theorem 6.6 (6.3, $p = 2$ case)**.** *For any prime $q \neq 2$, there exist constants c_1, c_2 so that any $\mathrm{AC}_0[q]$ circuit of depth t computing a function $f : \{0,1\}^n \to \{0,1\}$ on n variables with $\deg_2(f) = d$ requires size $c_1 2^{-d} \exp((c_2 \frac{n}{d8^d})^{\frac{1}{2t}})$.*

Proof. Assume there is an $\mathrm{AC}_0[q]$ circuit of size s and depth t computing f. Let ϵ be the constant in 6.3. Applying 6.2 with $\delta = 2^{-d}\epsilon$, there is some absolute constant c' and an $\mathbb{F}_q$ polynomial Q of degree $\deg(Q) \leq \left(c' \log \frac{s}{2^{-d}\epsilon}\right)^t$ such that

$$\Pr_{x \in \{0,1\}^n}[Q(x) = f(x)] \geq 1 - 2^{-d}\epsilon \, .$$

By 6.5 we get that $\deg(Q) \geq c\sqrt{\frac{n}{d8^d}}$ for some constant c. Hence,

$$s \geq c_1 2^{-d} \exp\left(\left(c_2 \frac{n}{d8^d}\right)^{\frac{1}{2t}}\right) \, ,$$

for absolute constants c_1, c_2. □

6.5 The case of general characteristic

Since we will be working with operations over different fields, we will denote by $+_p, +_q$ summation modulo p, q respectively, and by $+$ summation where the context is clear.

In this section we work with polynomials that represent a Boolean function over different characteristics. Suppose f is a Boolean function with low degree over $\mathbb{F}_p$. Our goal is to show that some suitable derivative of f is a linear function. We will then try to relate the degree of this derivative over $\mathbb{F}_q$ to $\deg_q(f)$. This scheme becomes harder to implement, since in differentiating a polynomial over $\mathbb{F}_p^n$, we need to take linear combinations of various points in $\mathbb{F}_p^n$. There is no natural way to associate $\mathbb{F}_p^n$ with a subset of $\mathbb{F}_q^n$ for $p > 2$. To overcome this difficulty, we define a suitable embedding of $\mathbb{F}_p^n$ to $\mathbb{F}_q^n$. While the proof is now technically harder, the basic idea stays the same.

Let $f(x)$ be a Boolean function. We start by defining a polynomial extending f to a function $F : \mathbb{F}_p^n \to \{0,1\}$. Given a vector $x \in \mathbb{F}_p^n$, we define $x^{p-1} = (x_1^{p-1}, \ldots, x_n^{p-1}) \in \{0,1\}^n$, which is the indicator of whether x is non-zero on each coordinate. Define the function $F : \mathbb{F}_p^n \to \{0,1\}$ by $F(x) = f(x^{p-1})$. $F(x)$ can be expressed as a polynomial of degree $(p-1)\deg_p(f)$ by considering the multilinear representation of f over $\mathbb{F}_p$ and replacing each variable x_i with x_i^{p-1}; henceforth we shall think of F as this polynomial. Our goal will be to show that if f has low degree over $\mathbb{F}_q$, then so does F and any function of the form $F(x +_p a_1) +_p \ldots +_p F(x +_p a_k)$. Since these are functions on $\mathbb{F}_p^n$, we need to define the notion of computing functions on $\mathbb{F}_p^n$ by polynomials over $\mathbb{F}_q$. Set $b = \lceil \log_2 p \rceil$. We identify the lexicographically first p bit strings in $\{0,1\}^b$ with the set $\{0, \ldots, p-1\}$. We then identify $\mathbb{F}_p^n$ with a subset of $\mathbb{F}_q^{nb}$ by identifying $x = (x_1, \ldots, x_n) \in \mathbb{F}_p^n$ with $(x_{1,1}, \ldots, x_{1,b}, \ldots, x_{n,1}, \ldots, x_{n,b}) \in \mathbb{F}_q^{nb}$, where the value of x_i determines the values of

$(x_{i,1}, \ldots, x_{i,b})$. Notice that in fact we map $\mathbb{F}_p^n$ into $\{0,1\}^{nb} \subset \mathbb{F}_q^{nb}$. Given $x \in \mathbb{F}_p^n$, we use $\bar{x} \in \{0,1\}^{nb}$ to denote the vector in $\{0,1\}^{nb} \subset \mathbb{F}_q^{nb}$ that represents it. We use $\bar{x_i}$ to denote the vector $(x_{i,1}, \ldots, x_{i,b})$ representing x_i. We say that a polynomial $G(x) \in \mathbb{F}_q[x_{1,1}, \ldots, x_{n,b}]$ computes $F : \mathbb{F}_p^n \to \{0,1\}$ if $F(x) = G(\bar{x})$ for every $x \in \mathbb{F}_p^n$. We now show that if f has low degree in $\mathbb{F}_q$, then $F(x +_p a)$ can also be computed by a low degree polynomial over $\mathbb{F}_q$.

Lemma 6.8. *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. Let $F(x)$ be a polynomial over $\mathbb{F}_p$ defined by $F(x) = f(x^{p-1})$. Then, for every $a \in \mathbb{F}_p^n$ there is a polynomial $G_a(x) \in \mathbb{F}_q[x_{1,1}, \ldots, x_{n,b}]$ over $\mathbb{F}_q$ of degree at most $b \cdot \deg_q(f)$ computing $F(x +_p a)$.*

Proof. For $a = (a_1, \ldots, a_n) \in \mathbb{F}_p^n$ and $i \in [n]$ let $A_i(\bar{x_i}) \in \mathbb{F}_q[\bar{x_i}]$ be such that $\deg(A_i) \leq b$ and

$$A_i(\bar{x_i}) = \begin{cases} 0 & \text{if } x_i +_p a_i = 0 \bmod p \\ 1 & \text{otherwise} \end{cases}$$

Recall that $\bar{x_i}$ is a 0/1 vector of length b, therefore we can define A_i to be a multilinear polynomial by only considering its values on $\{0,1\}^b$. When the input to A_i is not a vector of the form $\bar{x_i}$ we allow it to output an arbitrary value in $\mathbb{F}_q$. As A_i is multilinear its degree is clearly at most b. By definition it follows that $(A_1(\bar{x_1}), \ldots, A_n(\bar{x_n})) = (x +_p a)^{p-1}$. Let $g : \mathbb{F}_q^n \to \mathbb{F}_q$ be a polynomial of degree $\deg_q(f)$ representing f over $\mathbb{F}_q$. Define the polynomial $G_a(\bar{x}) : \mathbb{F}_q^{bn} \to \mathbb{F}_q$ as

$$G_a(\bar{x}) = g(A_1(\bar{x_1}), \ldots, A_n(\bar{x_n})) \ .$$

We have:

$$G_a(\bar{x}) = g(A_1(\bar{x_1}), \ldots, A_n(\bar{x_n})) = g((x +_p a)^{p-1}) = f((x +_p a)^{p-1}) = F(x +_p a)$$

as required, and $\deg(G_a) \leq b \deg(g) = b \deg_q(f)$. □

As in the proof of 6.7 we shall need to compute Boolean predicates, on expressions of the form $F(x +_p a_1) +_p \ldots +_p F(x +_p a_k)$, by low degree polynomials over $\mathbb{F}_q$.

Corollary 6.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function and $F(x)$ be a polynomial over $\mathbb{F}_p$ defined by $F(x) = f(x^{p-1})$. Let $a_1, \ldots, a_k \in \mathbb{F}_p^n$, $\lambda_1, \ldots, \lambda_n \in \mathbb{F}_p$ and $t : \mathbb{F}_p \to \{0,1\}$ be any Boolean valued predicate on $\mathbb{F}_p$. Define the function $T : \mathbb{F}_p^n \to \{0,1\}$ as*

$$T(x) = t\left(\sum_{i \leq k} \lambda_i F(x +_p a_i)\right) \ .$$

Then, T can be computed by a polynomial over $\mathbb{F}_q$ of degree at most $kb \deg_q(f)$.

Proof. By 6.8, each function $F(x +_p a_i) = f((x +_p a_i)^{p-1})$ can be computed by a polynomial $G_i(\bar{x})$ over $\mathbb{F}_q$ of degree at most $b \deg_q(f)$. The function $T(x)$ is a function of $G_1(\bar{x}), \ldots, G_k(\bar{x}) \in \{0,1\}$, and thus can be computed by $H(G_1(\bar{x}), \ldots, G_k(\bar{x}))$, where $H(z_1, \ldots, z_k)$ is a multilinear polynomial over $\mathbb{F}_q$ computing the function $t(\lambda_1 z_1 +_p \ldots +_p \lambda_k z_k) : \{0,1\}^k \to \{0,1\}$. Thus, T can be computed by a polynomial over $\mathbb{F}_q$ of degree at most $kb \deg_q(f)$. □

We now prove 6.1 in the case of general p.

Proof of 6.1 for general p. Let $d = \deg_p(f)$, and consider $F(x) = f(x^{p-1})$ which has degree $(p-1)d$. Invoking 6.4 for $F(x)$ which has degree $(p-1)d$, we conclude that there exist $k \leq p^d$ points $a_1, \ldots, a_k \in \mathbb{F}_p^n$ such that $G(x) = \sum_{i=1}^{k} \lambda_i F(x +_p a_i)$ satisfies $|L(G)| \geq n/(dp^d)$. Let $S = L(G)$ and rename the variables in S as $x_1, \ldots, x_s$, where $s = |S|$. Let G_S be the restriction of G to the variables in S (by setting the other variables to zero). We get that for some $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_p \setminus \{0\}$ and $\alpha_0 \in \mathbb{F}_p$,

$$G_S(x) = \sum_{i=1}^{s} \alpha_i x_i + \alpha_0 \ .$$

Let ω be a p^{th} root of unity in the appropriate extension field $\mathbb{F} = \mathbb{F}_{q^r}$ of $\mathbb{F}_q$. We consider the function $h : \{0,1\}^s \to \mathbb{F}$, which, by abuse of notations, is given by $h(x) = \omega^{\sum_{1 \leq i \leq s} \alpha_i x_i +_p \alpha_0}$. Indeed, we think of the expression $\sum_{1 \leq i \leq s} \alpha_i x_i +_p \alpha_0$ as taking values in $\{0, 1, \ldots, p-1\}$ and then raise ω to the appropriate power. The unique multilinear polynomial $H(x)$ over $\mathbb{F}$ computing h on $\{0,1\}^s$ has degree $\deg_{\mathbb{F}}(H) = s \geq \frac{n}{dp^d}$ and is given by

$$H(x) = \omega^{\alpha_0} \prod_{i=1}^{s} (1 + (\omega^{\alpha_i} - 1)x_i) \ .$$

We now upper-bound $\deg(H)$ in terms of $\deg_q(f)$. First, for $i \in \{0, \ldots, p-1\}$ let $t_i : \mathbb{F}_p \to \{0,1\}$ be the predicate indicating whether $x \equiv i \bmod p$. Consider the polynomial $T_i : \mathbb{F}_p^n \to \{0,1\}$ defined by $T_i = t_i(G_S(x))$. Since $G_S(x)$ is obtained by setting some of the variables in $\sum_i \lambda_i F(x +_p a_i)$ to zero, 6.2 gives $\deg_q(T_i) = \deg_q(t_i(G_S(x))) \leq kb \deg_q(f)$. Notice that as $H(x)$ is unique, it also equal to the multlinearization of the polynomial

$$\tilde{H}(x) = \sum_{i=0}^{p-1} \omega^i T_i(x) \ .$$

It follows that

$$s = \deg_{\mathbb{F}}(H) \leq \max_i \deg_q(T_i(x)) = \max_i \deg_q(t_i(G_S(x))) \leq kb \deg_q(f) \ .$$

Therefore,

$$\deg_q(f) \geq \frac{s}{bk} \geq \frac{n}{\lceil \log_2 p \rceil dp^{2d}} \ .$$

□

We use 6.1 to prove 6.1.

Proof of 6.1. Let p be the smallest prime divisor of m and let $q \neq p$ be another prime divisor. Note that by 6.3, we have $\deg_m(f) \geq \max(\deg_p(f), \deg_q(f))$ so it suffices to show that one of $\deg_p(f)$ or $\deg_q(f)$ exceeds the claimed bound.

So assume that $\deg_p(f) \leq \frac{1}{2} \log_p n - \log_p \log_p n - \frac{1}{2} \log_p \lceil \log_2 p \rceil$. By 6.1, we get

$$\deg_q(f) \geq \frac{n}{\lceil \log_2 p \rceil \deg_p(f) p^{2 \deg_p(f)}} \geq \log_p n$$

where the last inequality is a simple calculation. This proves the desired bound. □

Next we prove 6.2 showing that functions with low degree over $\mathbb{F}_p$ are hard to approximate over $\mathbb{F}_q$. First we state the theorem precisely.

Theorem 6.7 (6.2 for general p)**.** *For any primes $p \neq q$ there exist constants $c, \epsilon > 0$ depending only on p, q such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function depending on all n variables with $\deg_p(f) = d$. Let $h : \mathbb{F}_q^n \to \{0,1\}$ be any function satisfying*

$$\Pr_{x \in \{0,1\}^n}[h(x) = f(x)] \geq 1 - p^{-d}\epsilon \; .$$

Then

$$\deg_q(h) \geq c\sqrt{\frac{n}{dp^{3d}}} \; .$$

We start with some technical claims.

Claim 6.3. *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function, such that $\deg_p(f) = d$. For $v \in \{0,1\}^n$ define $F_v : \mathbb{F}_p^n \to \{0,1\}$ as*

$$F_v(x) = f(x^{p-1} \oplus v)$$

*where for $y, v \in \{0,1\}^n$, $y \oplus v \in \{0,1\}^n$ denotes their coordinatewise-*Xor*. Then F_v is a polynomial over $\mathbb{F}_p$ of degree at most $(p-1)d$.*

To prove this claim, we construct the polynomial for F_v from the multilinear polynomial for f by replacing x_i with x^{p-1} or $1 - x^{p-1}$ depending on whether or not $v_i = 0$. As this argument appeared several times before we omit the details.

Claim 6.4. *Let $f(x)$ and $g(x)$ be two Boolean functions such that*

$$\Pr_{x \in \{0,1\}^n}[f(x) = g(x)] \geq 1 - \epsilon \; .$$

Then there exists $v \in \{0,1\}^n$ such that if we define $F_v(x) = f(x^{p-1} \oplus v)$ and $G_v = g(x^{p-1} \oplus v)$ then

$$\Pr_{x \in \mathbb{F}_p^n}[F_v(x) = G_v(x)] \geq 1 - \epsilon \; .$$

Proof. Consider the following expression over a uniform choice of $v \in \{0,1\}^n$

$$\mathbb{E}_v[\Pr_{x \in \mathbb{F}_p^n}[F_v(x) = G_v(x)]] = \Pr_{x \in \{0,1\}^n}[f(x) = g(x)] \geq 1 - \epsilon \; .$$

Thus the inequality holds for some $v \in \{0,1\}^n$, □

We also need the following analogue of 6.2:

Claim 6.5. *Let $F(x)$ and $H(x)$ be functions such that $\Pr_{x \in \mathbb{F}_p^n}[F(x) = H(x)] \geq 1 - \epsilon$. Let $a_1, \ldots, a_k \in \mathbb{F}_p^n$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{F}_p$. Then:*

$$\Pr_{x \in \mathbb{F}_p^n}[\sum_i \lambda_i F(x +_p a_i) = \sum_i \lambda_i H(x +_p a_i)] \geq 1 - k\epsilon \; .$$

We now prove 6.7.

Proof of 6.2 in the case of general p. Let $f(x)$ be a Boolean function of small degree d over $\mathbb{F}_p$. Let $h(x) : \mathbb{F}_q^n \to \{0,1\}$ be such that $\Pr_{x\in\{0,1\}^n}[f(x) = h(x)] \geq 1 - p^{-d}\epsilon$, for a small enough $\epsilon > 0$. We will prove that $\deg_q(h)$ is large. The proof will proceed by a series of transformations on the pair of functions, such that the pairs generated will remain close, f will be transformed into the Mod_p function, whereas h will be transformed into a function whose degree over $\mathbb{F}_q$ is bounded in terms of $\deg_q(h)$. By 6.3, it must then follow that $\deg_q(h)$ is large. From this point on we shall 'forget' that h is defined over $\mathbb{F}_q^n$ and only consider its values on $\{0,1\}^n \subset \mathbb{F}_q^n$. In other words, we shall think of h as a Boolean function.

The first step is to extend f, h to functions mapping $\mathbb{F}_p^n$ to $\{0,1\}$. Let $F_v(x) = f(x^{p-1} \oplus v)$ and $H_v(x) = h(x^{p-1} \oplus v)$ be mappings from $\mathbb{F}_p^n$ to $\{0,1\}$. By 6.4, there exists $v \in \{0,1\}^n$ such that

$$\Pr_{x\in\mathbb{F}_p^n}[F_v(x) = H_v(x)] \geq \Pr_{x\in\{0,1\}^n}[f(x) = h(x)] \geq 1 - p^{-d}\epsilon \; .$$

In addition, the degree of F_v over $\mathbb{F}_p$ is at most $(p-1)d$. The next step is to apply the degree reduction lemma to F_v. By 6.4, there is some k where

$$k \leq p^{\lceil \frac{\deg(F_v)-1}{p-1} \rceil} \leq p^d$$

vectors $a_1, \ldots, a_k \in \mathbb{F}_p^n$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{F}_p$, such that for $G_f(x) = \sum_{i\leq k} \lambda_i F_v(x +_p a_i)$ (the sum is addition modulo p) it holds that the set $S = L(G_f)$ has size $s \geq \frac{n}{dp^d}$. Let $G_h : \mathbb{F}_p^n \to \mathbb{F}_p$ be defined as

$$G_h(x) = \sum_{i\leq k} \lambda_i H_v(x +_p a_i) \; . \tag{6.6}$$

6.5 implies that

$$\Pr_{x\in\mathbb{F}_p^n}[G_f(x) = G_h(x)] \geq 1 - kp^{-d}\epsilon \geq 1 - \epsilon \; .$$

As in the proof of 6.5, there exists an assignment $u \in \mathbb{F}_p^{[n]\setminus S}$ to the variables outside S so that the agreement between G_f and G_h is at least as large. To ease notation, we denote these restrictions also as $G_f(x)$ and $G_h(x)$ (instead of $G_{f_{S,u}}(x)$ and $G_{h_{S,u}}(x)$). Note that $G_f(x) = \sum_{i\leq k} \alpha_i x_i +_p \alpha_0$ where for $1 \leq i \leq s$ $\alpha_i \in \mathbb{F}_p \setminus \{0\}$, $\alpha_0 \in \mathbb{F}_p$ and the summation is modulo p. By replacing each x_i in G_f and G_h by $\alpha_i^{-1} x_i$, we get new functions $G'_f, G'_h : \mathbb{F}_p^s \to \mathbb{F}_p$ such that $G'_f(x) = \sum_i x_i +_p \alpha_0$ and

$$\Pr_{x\in\mathbb{F}_p^s}[G'_h(x) = \sum_i x_i +_p \alpha_0] = \Pr_{x\in\mathbb{F}_p^s}[G'_h(x) = G'_f(x)] \geq 1 - \epsilon \; .$$

The final step is to convert G'_h to a Boolean function approximating the Mod_p function on s variables. Towards this, for each $w \in \mathbb{F}_p^s$, we define $h_w : \{0,1\}^s \to \mathbb{F}_p$ by $h_w(y) = G'_h(y +_p w)$. Note that since $y +_p w$ is distributed uniformly at random over $\mathbb{F}_p^s$ we have that

$$\begin{aligned} &\Pr_{w\in\mathbb{F}_p^s}[\Pr_{y\in\{0,1\}^s}[h_w(y) = \sum_i y_i +_p \sum_i w_i +_p \alpha_0]] \\ &= \Pr_{x\in\mathbb{F}_p^s}[G'_h(x) = \sum_i x_i +_p \alpha_0] \geq 1 - \epsilon \; . \end{aligned}$$

Thus there exists w so that

$$\Pr_{y\in\{0,1\}^s}[h_w(y) = \sum_{i\le s} y_i +_p \alpha] \ge 1-\epsilon$$
$$\text{where } \alpha = \alpha_0 +_p \sum_i w_i \in \mathbb{F}_p \ .$$

Define $t : \mathbb{F}_p \to \{0,1\}$ by $t(z) = 1$ iff $z \equiv \alpha \bmod p$ and $t(z) = 0$ otherwise. Finally, let $\tilde{h}(y) = t(h_w(y))$. Notice that $t(\sum_{i\le s} y_i +_p \alpha) = 1$ iff $\sum_{i\le s} y_i \equiv 0 \bmod p$. In other words, $t(\sum_{i\le s} y_i +_p \alpha) = \mathsf{Mod}_p(y)$. We thus have

$$\Pr_{y\in\{0,1\}^s}[\tilde{h}(y) = \mathsf{Mod}_p(y)] \ge \Pr_{y\in\{0,1\}^s}[h_w(y) = \sum_{i\le s} y_i +_p \alpha] \ge 1-\epsilon \ .$$

Set $\epsilon > 0$ to be the constant guaranteed by 6.2. By 6.2, there exist a constant $c' > 0$ (where both c', ϵ depend only on p, q) such that $\deg_q(\tilde{h}) \ge c'\sqrt{s}$. Our goal now is to relate $\deg_q(h)$ to $\deg_q(\tilde{h})$. We make the following observations:

1. We have $h_w(y) = G'_h(y +_p w)$.
2. $G'_h(x)$ is obtained from $G_h(x)$ by setting variables outside S to constants and replacing each $x_i \in S$ by $\alpha^{-1}x_i$.
3. By Equation 6.6, $G_h(x)$ is a linear combination over $\mathbb{F}_p$ of values of the form $H_v(x+_p a_i)$.
4. Each $H_v(x +_p a_i)$ can be computed by a polynomial $Q_i(\bar{x})$ over $\mathbb{F}_q$ of degree at most $\lceil \log_q p \rceil \cdot \deg_q(h)$ by an argument similar to 6.8.

Thus, we can write $\tilde{h}(y)$ as some predicate $t' : \{0,1\}^k \to \{0,1\}$ applied to a tuple of polynomial $Q_1, \ldots, Q_k$ with $\deg_q(Q_i) \le \lceil \log_q p \rceil \deg_q(h)$, and hence $\deg_q(\tilde{h}) \le kb \deg_q(h)$. We conclude that

$$\deg_q(h) \ge \frac{c'\sqrt{s}}{k\lceil \log_q p \rceil} = \frac{c'}{\lceil \log_q p \rceil}\sqrt{\frac{n}{dp^{3d}}} \ .$$

Hence we proved the theorem with the constant $c = \frac{c'}{\lceil \log_q p \rceil}$. □

As a corollary we obtain a lower bound for the size of $AC_0[q]$ circuits computing functions with low degree over $\mathbb{F}_p$.

Theorem 6.8 (6.3, restated)**.** *Let p, q be distinct primes. Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function depending on all n variables with $\deg_p(f) = d$. Then any $\mathrm{AC}_0[q]$ circuit of depth t computing f requires size at least*

$$c_1 p^{-d} \exp\left(c_2 \left(c_3 \frac{n}{dp^{3d}}\right)^{\frac{1}{2t}}\right),$$

where c_1, c_2, c_3 are constants depending only on p, q. In particular, for $d = o(\log_p n)$, the lower bound is $\exp(n^{1/2t - o(1)})$.

Proof. Assume there is an $\mathrm{AC}_0[q]$ circuit of size s and depth t computing f. Let ϵ be the constant in 6.3. Applying 6.2 with $\delta = p^{-d}\epsilon$ we get that there is some absolute constant c' and an $\mathbb{F}_q$ polynomial $Q : \mathbb{F}_q^n \to \{0,1\}$ of degree $\deg(Q) \leq \left(c' p \log \frac{s}{p^{-d}\epsilon}\right)^t$ such that $\Pr_{x\in\{0,1\}^n}[Q(x) = f(x)] \geq 1 - p^{-d}\epsilon$. By 6.7 $\deg(Q) \geq c''\sqrt{\frac{n}{dp^{3d}}}$ for some constant c'' depending only on p, q. Hence, for $c_1 = \epsilon, c_2 = c'p, c_3 = c''$ we get that

$$s \geq c_1 p^{-d} \exp\left(c_2 \left(c_3 \frac{n}{dp^{3d}}\right)^{\frac{1}{2t}}\right),$$

as claimed. □

6.6 Open problems

Our work raises some natural questions regarding the relations between $\deg_m(f)$ for various characteristics, some of which we list below:

1. For any integer m, we have $\deg(f) \geq \deg_m(f)$. What is the largest separation possible between these quantities when m is not a prime power? For such m, is $\deg(f)$ polynomial in $\deg_m(f)$? We can restate these questions as follows: Can $\deg(f)$ be bounded as a function of $\deg_p(f)$ and $\deg_q(f)$ for distinct primes p and q?

   Note that the gap between $\deg(f)$ and $\deg_m(f)$ can be unbounded when m is a prime-power. If m is not a prime power, 6.1 gives an analog of the $\Omega(\log n)$ Nisan-Szegedy lower bound for composite moduli. Thus trivially, $\deg(f)$ is at most exponential in $\deg_m(f)$.

2. The following question was posed by Troy Lee: Given a set S of vectors in $\{0,1\}^n$, define $\mathrm{Rank}_p(S)$ to be the rank of the set S over $\mathbb{F}_p$ and $\mathrm{Rank}(S)$ to be the rank over $\mathbb{R}$. Are there non-trivial relations between these ranks? For example, assume that both $\mathrm{Rank}_2(S)$ and $\mathrm{Rank}_3(S)$ are small, say $\mathrm{poly}(\log n)$. What can be said about $\mathrm{Rank}(S)$? Note that if we consider only $\mathrm{Rank}_2(S)$ then the Hadamard matrix is an example of a full rank matrix over $\mathbb{R}$ that has rank $\log n$ over $\mathbb{F}_2$.

Part III

Pseudorandom generators for low-degree polynomials

Chapter 7

Pseudorandom generators for low-degree polynomials over finite fields

We give an explicit construction of a pseudorandom generator against low-degree polynomials over finite fields. Pseudorandom generators against linear polynomials, known as *small-bias generators*, were first introduced by Naor and Naor (STOC 1990). We show that the sum of 2^d independent small-bias generators with error $\epsilon^{2^{O(d)}}$ is a pseudorandom generator against degree-d polynomials with error ϵ. This gives a generator with seed length $2^{O(d)} \log(n/\epsilon)$ against degree-d polynomails. Our construction follows the breakthrough result of Bogdanov and Viola (FOCS 2007). Their work shows that the sum of d small-bias generators is a pseudo-random generator against degree-d polynomials, assuming a conjecture in additive combinatorics, known as *the inverse conjecture for the Gowers norm*. However, this conjecture was proven only for $d = 2, 3$. The main advantage of this work is that it does not rely on any unproven conjectures.

Subsequently, the inverse conjecture for the Gowers norm was shown to be false for $d \geq 4$ by Green and Tao (2008) and independently by the author, Roy Meshulam, and Alex Samorodnitsky (STOC 2008). A revised version of the conjecture was proved by Bergelson, Tao, and Ziegler (2009). Additionally, Viola (CCC 2008) showed the original construction of Bogdanov and Viola to hold unconditionally.

7.1 Introduction

We are interested in explicitly constructing pseudorandom generators (PRG) against low-degree polynomials over small finite fields. A pseudorandom generator against a family $\mathbb{T}$ of tests is a function G mapping a small domain into a (much) larger one, such that any test $T \in \mathbb{T}$ cannot distinguish, with noticeable probability, a random element in the large domain from an application of G to a random element in the small domain. We say a PRG requires R random bits if the size of the small domain is 2^R.

In our case, $\mathbb{F}$ is a finite field and a test is a polynomial $p(x_1, \ldots, x_n)$ over $\mathbb{F}$. The image of the PRG is a small subset of $\mathbb{F}^n$, and it is pseudorandom against $p(x_1, \ldots, x_n)$ if the

distribution of the outcome of p, when applied to a random element in the small subset, is close to the distribution of the outcome of p, when applied to a uniform element in $\mathbb{F}^n$. We say the PRG has error ϵ against p if the statistical distance between the two distributions is at most ϵ. We are interested in PRGs that are pseudorandom against all degree-d polynomials with error ϵ, and use as few random bits as possible.

The case of pseudorandom generators against linear polynomials, usually called *small-bias generators* (or *epsilon-biased generators*, a term we do not use in this paper to avoid confusion), was first studied (over $\mathbb{F} = \mathbb{F}_2$) by Naor and Naor [NN93] and later by Alon, Goldreich, Håstad and Peralta [AGHP90]. They and others gave explicit constructions, which were later generalized to arbitrary finite fields. These constructions have a seed length which is optimal up to a constant multiplicative factor. The construction of small-bias generators is a major tool in derandomization, PCPs and lower bounds (see [BSSVW03] and the references within for details regarding small-bias generators).

The generalization of the problem to constant-degree polynomials was first studied by Luby, Velickovic, and Wigderson [LVW93]. Their results apply, in fact, to the more general model of constant depth circuits. In the context of constant degree polynomials, they give an explicit construction of PRG requiring $\exp(O(\sqrt{\log n/\epsilon}))$ random bits.

Bogdanov [Bog05] gave a construction of a PRG in large fields. The minimum field size required for his construction is polynomial in the degree, the required error and the log of the number of variables. In these settings, his construction is optimal up to polynomial factors. The proof of his result uses techniques and results from algebraic geometry and computational algebra.

Recently, Bogdanov and Viola [BV07] presented a novel approach for constructing a PRG for low-degree polynomials over small fields. Their construction is the sum of d independent small-bias generators. They showed that, if a conjecture in additive combinatorics called the *inverse conjecture for the Gowers norm* holds, then their construction is a PRG for degree-d polynomials. At the time, the inverse conjecture for the Gowers norm was known to hold only for degrees 2 and 3, and was conjectured to hold for all constant degrees. Thus, their construction was known to be correct only for quadratic and cubic polynomials.

Our work was inspired by the work of Bogdanov and Viola, with the goal of making their construction unconditional, i.e., not relying on any unproven conjectures. We prove that the sum of 2^d independent small-bias generators is pseudorandom against degree-d polynomials, without relying on any unproven conjectures. Our main theorem is:

Theorem 7.1. *There exists a global constant $c > 0$ such that the following holds. Let G be a small-bias generator with error $\epsilon^{2^{cd}}$. Then the sum of 2^d independent copies of G is pseudorandom against degree-d polynomials with error ϵ. In particular, this gives a pseudorandom generator for degree-d polynomials with error ϵ using $2^{cd}\log(|\mathbb{F}|n/\epsilon)$ random bits for the seed.*

7.1.1 Overview of proof method

This work is inspired by the recent result of Bogdanov and Viola [BV07]. We begin by providing a high level description of it, since several ideas used in [BV07] are also used in our work.

The analysis of [BV07] crucially depended on the inverse conjecture for the Gowers norm. Although we do not use this conjecture in our proof, we now briefly present and discuss it. The Gowers norm, first defined by Gowers in his new proof for Szemerédi's theorem [Gow01], is a norm measuring the local correlation of a function to low-degree polynomials. Let $\mathbb{F} = \mathbb{F}_q$ be a prime finite field, and assume $f(\mathbf{x}) : \mathbb{F}^n \to \mathbb{F}$ is a function. The directional derivative of f in direction $\mathbf{y} \in \mathbb{F}^n$ is defined to be

$$f_{\mathbf{y}}(\mathbf{x}) = f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}).$$

Notice that if f is a degree-d polynomial, then $f_{\mathbf{y}}$ is a polynomial of degree at most $d-1$, hence the term derivative relates to the more common definition of analytical derivative. We define also iterated derivatives: $f_{\mathbf{y}_1,\ldots,\mathbf{y}_k}(x)$ is defined recursively, by taking the k derivatives in directions $y_1, \ldots, y_k$. Opening brackets, this gives

$$f_{\mathbf{y}_1,\ldots,\mathbf{y}_k}(\mathbf{x}) = \sum_{S \subset \{1,\ldots,k\}} (-1)^{k-|S|} f(\mathbf{x} + \sum_{i \in S} \mathbf{y}_i).$$

The d-th Gowers norm of f is defined as

$$U_d(f) = \left(\mathbb{E}_{\mathbf{x},\mathbf{y}_1,\ldots\mathbf{y}_d \in \mathbb{F}_q^n} \left[\omega_q^{f_{\mathbf{y}_1,\ldots,\mathbf{y}_d}(\mathbf{x})} \right] \right)^{\frac{1}{2^d}},$$

where $\omega_q = e^{\frac{2\pi i}{q}}$ is a root of unity of order q. It was proved to be a norm on functions (for $d \geq 2$) by Gowers [Gow01].

Assume f is a degree-$(d-1)$ polynomial. Taking d derivatives results in the zero polynomial, so $f_{\mathbf{y}_1,\ldots,\mathbf{y}_d} \equiv 0$ for any choice of $\mathbf{y}_1, \ldots, \mathbf{y}_d$ and consequently $U_d(f) = 1$. It is relatively easy to see that the converse also holds, that is, $U_d(f) = 1$ iff f is a polynomial of degree at most $d-1$. Alon et al. [NAR03] proved a robust version of this equivalence: the d-th Gowers norm of f is very close to 1 iff the function f is very close to a degree-$(d-1)$ polynomial.

The inverse conjecture for the Gowers norm studies the realm of functions with only a noticeable Gowers norm, that is $U_d(f) \geq \delta$ for some $\delta > 0$. Gowers [Gow01] showed that if f is only somewhat close to a degree-$(d-1)$ polynomial, that is $\Pr_{\mathbf{x}}[f(\mathbf{x}) = p(\mathbf{x})] \geq 1/q + \epsilon$ for some degree-$(d-1)$ polynomial $p(\mathbf{x})$, then f has a noticeable d-th Gowers norm, $U_d(f) \geq \epsilon'$, where $\epsilon' = \Omega(\epsilon)$.

The converse of this claim is known as the inverse conjecture for the Gowers norm: if $U_d(f) \geq \epsilon$, then there exists a degree-$(d-1)$ polynomial p such that $\Pr_{\mathbf{x}}[f(\mathbf{x}) = p(\mathbf{x})] \geq 1/q + \epsilon'$, for some $\epsilon' > 0$ depending on ϵ. The case of $d = 2$ can be proven using standard Fourier analysis tools [BCH+95]. The case of $d = 3$ was proven by Green and Tao [GT08] and independently by Samorodnitsy [Sam07]. Both works conjectured this to hold for any constant degree.

Returning to the argument of [BV07], Bogdanov and Viola analyze the Gowers norm of a degree-d polynomial $p(\mathbf{x})$, and present a win-win argument, depending on whether the Gowers norm is either small or large. In the first case, when the Gowers norm is small, they show that the sum of d small-bias generators is pseudorandom against $p(\mathbf{x})$, by relating the distribution of $p(\mathbf{x}_1 + \cdots + \mathbf{x}_d)$ to the Gowers norm of p. In the latter case, when the Gowers norm is large, and assuming the inverse conjecture for the Gowers norm holds, $p(\mathbf{x})$

is correlated to some degree-$(d-1)$ polynomial $q(\mathbf{x})$. They use $q(\mathbf{x})$ in order to construct a circuit that computes $p(\mathbf{x})$ for almost all values of x. The inputs to this circuit are all degree-$(d-1)$ polynomials; thus they show that a PRG for degree-$(d-1)$ polynomials with small enough error is also pseudorandom against $p(\mathbf{x})$.

Our construction follows similar lines; however, instead of analyzing the Gowers norm of $p(\mathbf{x})$, we analyze its Fourier coefficients. We also divide our treatment into two cases: when p has some large Fourier coefficient, and when all the Fourier coefficients of p are small.

In the first case, when $p(\mathbf{x})$ has no large Fourier coefficients, we consider inputs to p of the form $\mathbf{x}+\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are independent. We consider the polynomial

$$\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}'') = p(\mathbf{x}'+\mathbf{y}') - p(\mathbf{x}'+\mathbf{y}'') - p(\mathbf{x}''+\mathbf{y}') + p(\mathbf{x}''+\mathbf{y}'')\,.$$

We prove that it is enough to be pseudorandom against Δp in order to be pseudorandom against $p(\mathbf{x}+\mathbf{y})$, and also that it is sufficient to have $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{y}$, $\mathbf{y}'$ and $\mathbf{y}''$ come from a PRG that is pseudorandom against degree-$(d-1)$ polynomials. The reason is that Δp contains no degree-d terms in just one of $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{y}'$ or $\mathbf{y}''$. In the second case, when there is some large Fourier coefficient, we know that $p(\mathbf{x})$ is correlated to some linear function. Similarly to the second case in [BV07], we also show in that case, or more generally when $p(\mathbf{x})$ is correlated to some lower degree polynomial, a PRG for degree-$(d-1)$ polynomials with small enough error is also pseudorandom against $p(\mathbf{x})$. However, our proof technique is more direct than the one used in [BV07], which results in better parameters and simpler analysis.

7.1.2 Subsequent work

This paper is a more polished version of the extended abstract of this work first presented at STOC 2008. Subsequently, there were advances on two fronts.

First, the inverse conjecture for the Gowers norm was shown to be false for degrees ≥ 4 by Green and Tao [GT07] and independently by Lovett, Meshulam, and Samorodnitsky [LMS08]. A revised inverse conjecture for the Gowers norm was proved by Bergelson, Tao and Ziegler [BTZ09, TZ09].

Additionally, Viola [Vio08] proved the correctness of the construction of [BV07] without using the inverse conjecture for the Gowers norm, or any other unproven conjectures, thus making the original construction of [BV07] unconditionally correct. His proof method also follows similar lines to the works of [BV07] and this work. He considers $p(\mathbf{x}+\mathbf{y})$, where $\mathbf{x}$ comes from a distribution which is pseudorandom against degree-$(d-1)$ polynomials, and $\mathbf{y}$ is a small-bias generator (i.e., pseudorandom against linear polynomials). He also uses a win-win analysis, based on the *bias* of the polynomial p, and proves that indeed the sum $\mathbf{x}+\mathbf{y}$ fools all degree-d polynomials.

The result presented here can thus be seen as an intermediate step in a sequence of works. The proof of Viola uses some of the techniques developed in this work, in addition to some of the original techniques introduced in [BV07] and some clever new ideas.

7.2 Preliminaries

We work over an arbitrary finite field $\mathbb{F}$. Let $U = U_n$ be the uniform distribution over $\mathbb{F}^n$. We fix $e : \mathbb{F} \to \mathbb{C}$ to be any non-trivial additive character. For example, in a prime field $\mathbb{F}_q$ we can have $e(x) = \omega_q^x$ where $\omega_q = 2^{\frac{2\pi i}{q}}$ is a root of unity of order q. When we refer to the degree of a multivariate polynomial, we always mean its total degree. We denote elements of $\mathbb{F}^n$ by $\mathbf{x} = (x_1, \ldots, x_n)$.

Definition 7.1. *A distribution D over $\mathbb{F}^n$ is said to be pseudorandom against a polynomial $p(x_1, \ldots, x_n)$ with error ϵ if*

$$|\mathbb{E}_{\mathbf{x}\in D}\left[e(p(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{x}\in U}\left[e(p(\mathbf{x}))\right]| < \epsilon\,.$$

Definition 7.2. *A distribution D is said to be pseudorandom against degree-d polynomials with error ϵ if for every degree-d polynomial $p(x_1, \ldots, x_n)$, D is pseudorandom against p with error ϵ.*

We study explicit constructions for pseudorandom generators against degree-d polynomials.

Definition 7.3. *A function $G : \{0,1\}^r \to \mathbb{F}^n$ is said to be a pseudorandom-generator (PRG) against degree-d polynomials if the distribution obtained by applying G to a uniform element in $\{0,1\}^r$ is a pseudorandom distribution against degree-d polynomials. The value in $\{0,1\}^r$ is called the* seed *of G, and r is the* seed length *of G.*

The notion of pseudorandomness we use is different from more standard notions of pseudorandomness. However, since we are working over small fields, they are tightly related. For example, the following Lemma from [BV07] connects it with the common notion of pseudorandomness in statistical distance (The proof in [BV07] is stated just for prime fields, but it remains correct over arbitrary fields):

Lemma 7.1 (Lemma 33 in [BV07])**.** *Let D be a distribution that is pseudorandom against degree-d polynomials with error ϵ. Let $p(x_1, \ldots, x_n)$ be a polynomial of degree at most d. Let $p(D)$ be the distribution, taking values in $\mathbb{F}$, obtained by applying p to an input chosen according to D, and similarly $p(U)$ be the distribution of applying p to a uniformly chosen input in $\mathbb{F}^n$. Then the variation (statistical) distance between $p(D)$ and $p(U)$ is bounded by $\frac{1}{2}\epsilon\sqrt{|\mathbb{F}| - 1}$.*

Remark. *Definition 7.2 does not depend on which non-trivial character is used in Definition 7.1; since we require pseudorandomness for all degree-d polynomials, we can multiply polynomials by any non-zero constant, thus effectively achieving pseudorandomness for all non-trivial characters.*

We use the Cauchy-Schwarz inequality over the complex numbers in the following form several times in the proof.

Claim 7.1. *Let Z be a random variable taking values in $\mathbb{C}$, then*

$$|\mathbb{E}[Z]|^2 \leq \mathbb{E}\left[|Z|^2\right]\,.$$

Fourier analysis plays a central role in our proof. In the following we define Fourier coefficients, and discuss several properties of them required in the proof. We refer to the first chapter of [Šte00] for a more in-depth introduction to Fourier analysis.

Definition 7.4. *The Fourier coefficients of a function $f : \mathbb{F}^n \to \mathbb{C}$ are defined to be*

$$\hat{f}_\alpha = \mathbb{E}_{\mathbf{x}\in U}\left[f(\mathbf{x})e(-\langle\alpha,\mathbf{x}\rangle)\right] ,$$

where $\alpha = (\alpha_1,\ldots,\alpha_n) \in \mathbb{F}^n$ and $\langle\alpha,\mathbf{x}\rangle = \alpha_1x_1+\cdots+\alpha_nx_n$ is the inner product of α and $\mathbf{x}$.

The set of functions $\{e(\langle\alpha,\mathbf{x}\rangle) : \alpha \in \mathbb{F}^n\}$ is an orthonormal basis of the Hermitian space of functions $\mathbb{F}^n \to \mathbb{C}$ under the inner product

$$f \cdot g = \frac{1}{|\mathbb{F}^n|}\sum_{\mathbf{x}\in\mathbb{F}^n}\overline{f(\mathbf{x})}g(\mathbf{x}) .$$

Therefore f can be expressed as

$$f(\mathbf{x}) = \sum_{\alpha\in\mathbb{F}^n}\hat{f}_\alpha e(\langle\alpha,\mathbf{x}\rangle) .$$

For a polynomial $p(\mathbf{x}) \in \mathbb{F}[x_1,\ldots,x_n]$ we define $\hat{p}_\alpha$ to be the α Fourier coefficient of the function $e(p(\mathbf{x}))$, i.e.,

$$\hat{p}_\alpha = \mathbb{E}_{\mathbf{x}\in U}\left[e(p(\mathbf{x})-\langle\alpha,\mathbf{x}\rangle)\right] .$$

We will need the following simple fact, which follows from Parseval's identity and the fact that $|e(p(\mathbf{x}))| = 1$ for all $\mathbf{x} \in \mathbb{F}^n$:

Fact 7.1. $\displaystyle\sum_{\alpha\in\mathbb{F}^n}|\hat{p}_\alpha|^2 = 1.$

The basis elements of our analysis are PRGs for degree-1 polynomials. PRGs for this family have been studied extensively, and are usually referred to as small-bias (or epsilon-biased) generators or distributions. Formally we define:

Definition 7.5. *A distribution D is called a* small-bias distribution *over $\mathbb{F}^n$ with error δ if for all linear polynomials $p(\mathbf{x}) = a_1x_1+\cdots+a_nx_n$ we have*

$$\left|\mathbb{E}_{\mathbf{x}\in D}\left[e(p(\mathbf{x}))\right]-\mathbb{E}_{\mathbf{x}\in U}\left[e(p(\mathbf{x}))\right]\right| < \delta . \tag{7.1}$$

Constructions of small-bias distributions were first studied by Naor and Naor over $\mathbb{F}_2$ in [NN93], and optimal up to constant constructions were later given by Alon, Goldreich, Håstad, and Peralta [AGHP90] over general fields. Such constructions can be achieved by explicit pseudorandom generators with seed length $O(\log(|\mathbb{F}|n/\epsilon))$.

7.3 Main theorem

We restate our main theorem with explicit constants:

Theorem 7.2. *Let $G : \{0,1\}^r \to \mathbb{F}^n$ be a small-bias generator over $\mathbb{F}^n$ with error $(\epsilon/10)^{4^d}$. Then the sum of 2^d independent copies of G is pseudorandom against degree-d polynomials with error ϵ. That is, $G' : \{0,1\}^{r \cdot 2^d} \to \mathbb{F}^n$ defined as*

$$G'(\mathbf{x}_1, \ldots, \mathbf{x}_{2^d}) = G(\mathbf{x}_1) + \ldots + G(\mathbf{x}_{2^d})$$

is a PRG against degree-d polynomials with error ϵ.

Our proof is divided into two cases, based on whether p has some large Fourier coefficient, or does not have any large Fourier coefficients. We show that when a degree-d polynomial $p(\mathbf{x})$ has some large Fourier coefficient, then a PRG for degree-$(d-1)$ polynomials, with better error, is also pseudorandom against p. On the other hand, if p has no large Fourier coefficients, it is "pseudorandom" in a sense, and then the sum of two PRGs for degree-$(d-1)$ is pseudorandom against p.

We divide the proof into two technical lemmas, dealing with the cases of whether p has some large Fourier coefficient, or it does not.

Lemma 7.2. *Let $p(x_1, \ldots, x_n)$ be a degree-d polynomial over $\mathbb{F}^n$, such that for all $\alpha \in \mathbb{F}^n$, $|\hat{p}_\alpha| < \epsilon^2/10$. Let D be a distribution that is pseudorandom against degree-$(d-1)$ polynomials with error $\epsilon^4/400$. Then $\mathbf{x}+\mathbf{y}$, where $\mathbf{x}$, $\mathbf{y}$ are independently chosen from D, is pseudorandom against p with error ϵ.*

Lemma 7.3. *Let $p(x_1, \ldots, x_n)$ be a degree-d polynomial over $\mathbb{F}^n$, such that $|\hat{p}_\alpha| \geq \epsilon^2/10$ for some $\alpha \in \mathbb{F}^n$. Let D be a distribution that is pseudorandom against degree-$(d-1)$ polynomials with error $\epsilon^3/10$. Then D is pseudorandom against $p(\mathbf{x})$ with error ϵ.*

Assuming these two lemmas, our main theorem now follows directly, by also using the following simple observation. This observation allows us to add "extra" small-bias distributions without harming our PRG construction.

Observation 7.1. *Let D be a distribution that is pseudorandom against degree-d polynomials with error ϵ. Let D' be any other independent distribution. Then the distribution of $\mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in D$ and $\mathbf{y} \in D'$ is also pseudorandom against degree-d polynomials with error ϵ.*

We now prove Theorem 7.2, assuming Lemmas 7.2 and 7.3 and Observation 7.1:

Proof. We prove, by induction on d, that the sum of 2^d independent small-bias generators with error $(\epsilon/10)^{4^d}$ is pseudorandom against degree-d polynomials with error ϵ. For $d = 1$ this is clear. For $d > 1$, let D' be the distribution of sum of the first 2^{d-1} small-bias generators, which is also the distribution of the sum of the last 2^{d-1} small-bias generators. Observe that by the inductive hypothesis, D' is pseudorandom against degree-$(d-1)$ polynomials with error $(\epsilon/10)^4 < \min(\epsilon^4/400, \epsilon^3/10)$. Let $p(x)$ be any degree-d polynomial. Consider first the case that all the Fourier coefficients of p are at most $\epsilon^2/10$. By Lemma 7.2, we know that the distribution of $\mathbf{x} + \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are chosen independently according to D', is

pseudorandom against p with error ϵ. Alternatively, consider the case that there exists some Fourier coefficient of p of absolute value at least $\epsilon^2/10$. By Lemma 7.3, D' is pseudorandom against p, and by Observation 7.1 so is the distribution of $\mathbf{x}+\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are chosen independently according to D'. □

The remainder of the paper is organized as follows: Lemma 7.2 is proven in Observation 7.4 and Lemma 7.3 in Section 7.5.

7.4 Case I: No large Fourier coefficients

In this section we prove Lemma 7.2. We assume throughout this section that all the Fourier coefficients of $e(p(\mathbf{x}))$ are small, i.e., $|\hat{p}_\alpha| < \epsilon^2/10$ for all $\alpha \in \mathbb{F}^n$.

We start by defining a derivation polynomial.

Definition 7.6. *Let $p(\mathbf{x}) : \mathbb{F}^n \to \mathbb{F}$ be a polynomial. We define its derivation polynomial $\Delta p : (\mathbb{F}^n)^4 \to \mathbb{F}$ as*

$$\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}'') = p(\mathbf{x}'+\mathbf{y}') - p(\mathbf{x}''+\mathbf{y}') - p(\mathbf{x}'+\mathbf{y}'') + p(\mathbf{x}''+\mathbf{y}'')\,.$$

The following lemma is crucial to our analysis, and is a variation of a lemma proven in [BV07]. We relate the distribution of evaluating p on the sum of two independent inputs to that of Δp.

Lemma 7.4. *Let $p : \mathbb{F}^n \to \mathbb{F}$. Let D be a distribution over $\mathbb{F}^n$. Let $\mathbf{x}$, $\mathbf{y}$ be independently chosen from D, then*

$$|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^4 \leq \mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}[e(\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}''))]\,,$$

where $\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}''$ are also independent.

Proof. The proof is essentially applying the Cauchy-Schwarz inequality twice. We start by showing

$$|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^2 \leq \mathbb{E}_{\mathbf{x},\mathbf{y}',\mathbf{y}''\in D}[e(p(\mathbf{x}+\mathbf{y}') - p(\mathbf{x}+\mathbf{y}''))]\,,$$

and then continue to show

$$|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^4 \leq \mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}[e(p(\mathbf{x}'+\mathbf{y}') - p(\mathbf{x}''+\mathbf{y}') - p(\mathbf{x}'+\mathbf{y}'') + p(\mathbf{x}''+\mathbf{y}''))]\,,$$

which is what we want to prove, by the definition of Δp. We prove the first part by applying the Cauchy-Schwarz inequality

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^2 \leq \mathbb{E}_{\mathbf{x}\in D}\,|\mathbb{E}_{\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^2 = \\
&\mathbb{E}_{\mathbf{x}\in D}\left[\mathbb{E}_{\mathbf{y}'\in D}[e(p(\mathbf{x}+\mathbf{y}'))]\;\overline{\mathbb{E}_{\mathbf{y}''\in D}[e(p(\mathbf{x}+\mathbf{y}''))]}\right] = \\
&\mathbb{E}_{\mathbf{x},\mathbf{y}',\mathbf{y}''\in D}[e(p(\mathbf{x}+\mathbf{y}') - p(\mathbf{x}+\mathbf{y}''))]\,.
\end{aligned}$$

We prove the second part by applying the Cauchy-Schwarz inequality again

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]|^4 \leq \\
&|\mathbb{E}_{\mathbf{x},\mathbf{y}',\mathbf{y}''\in D}[e(p(\mathbf{x}+\mathbf{y}')-p(\mathbf{x}+\mathbf{y}''))]|^2 \leq \\
&\mathbb{E}_{\mathbf{y}',\mathbf{y}''\in D}\,|\mathbb{E}_{\mathbf{x}\in D}[e(p(\mathbf{x}+\mathbf{y}')-p(\mathbf{x}+\mathbf{y}''))]|^2 = \\
&\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}\left[e(p(\mathbf{x}'+\mathbf{y}')-p(\mathbf{x}'+\mathbf{y}'')-p(\mathbf{x}''+\mathbf{y}')+p(\mathbf{x}''+\mathbf{y}''))\right].
\end{aligned}$$

□

In particular the following corollary follows:

Corollary 7.1. $\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right] \geq 0\,.$

We analyze the expression $\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right]$, in two cases: when $D=U$ is the uniform distribution and when D is a PRG for degree-$(d-1)$ polynomials. We show that in both cases it is at most $\epsilon/2$. Combining this with Lemma 7.4 yields the required result. We start our analysis in the uniform case.

We begin by showing the (well-known) connection between the average value of Δp and the Fourier coefficients of p, regarding Δp as an affinity-test for p. A similar analysis, carried in more depth, can be found in [BCH+95].

Lemma 7.5.

$$\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(p(\mathbf{x}'+\mathbf{y}')-p(\mathbf{x}'+\mathbf{y}'')-p(\mathbf{x}''+\mathbf{y}')+p(\mathbf{x}''+\mathbf{y}''))\right] = \sum_{\alpha\in\mathbb{F}^n}|\hat{p}_\alpha|^4\,.$$

Proof. We can write $e(p(\mathbf{x}))$ in the Fourier basis as

$$e(p(\mathbf{x})) = \sum_{\alpha\in\mathbb{F}^n}\hat{p}_\alpha e(\langle\alpha,\mathbf{x}\rangle)\,.$$

Notice that

$$e(-p(\mathbf{x})) = \overline{e(p(\mathbf{x}))} = \sum_{\alpha\in\mathbb{F}^n}\overline{\hat{p}_\alpha}e(-\langle\alpha,\mathbf{x}\rangle)\,.$$

We now expand all four terms of p in

$$e(p(\mathbf{x}'+\mathbf{y}')-p(\mathbf{x}'+\mathbf{y}'')-p(\mathbf{x}''+\mathbf{y}')+p(\mathbf{x}''+\mathbf{y}''))\,.$$

This is equal to

$$\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4\in\mathbb{F}^n}\hat{p}_{\alpha_1}e(\langle\alpha_1,\mathbf{x}'+\mathbf{y}'\rangle)\,\overline{\hat{p}_{\alpha_2}}e(-\langle\alpha_2,\mathbf{x}'+\mathbf{y}''\rangle)\,\overline{\hat{p}_{\alpha_3}}e(-\langle\alpha_3,\mathbf{x}''+\mathbf{y}'\rangle)\,\hat{p}_{\alpha_4}e(\langle\alpha_4,\mathbf{x}''+\mathbf{y}''\rangle)\,.$$

Remember that we are interested in the expected value over uniform $\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in\mathbb{F}^n$, i.e., in

$$\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(p(\mathbf{x}'+\mathbf{y}')-p(\mathbf{x}'+\mathbf{y}'')-p(\mathbf{x}''+\mathbf{y}')+p(\mathbf{x}''+\mathbf{y}''))\right].$$

We now use the Fourier expansion and group elements by their related values. After doing so, the above expectation is equal to

$$\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4 \in \mathbb{F}^n} \hat{p}_{\alpha_1}\overline{\hat{p}_{\alpha_2}\hat{p}_{\alpha_3}}\hat{p}_{\alpha_4}\mathbb{E}_{x' \in U}\left[e(\langle \alpha_1 - \alpha_2, \mathbf{x}' \rangle)\right]\mathbb{E}_{x'' \in U}\left[e(\langle \alpha_4 - \alpha_3, \mathbf{x}'' \rangle)\right]$$
$$\mathbb{E}_{y' \in U}\left[e(\langle \alpha_1 - \alpha_3, \mathbf{y}' \rangle)\right]\mathbb{E}_{y'' \in U}\left[e(\langle \alpha_4 - \alpha_2, \mathbf{y}'' \rangle)\right] .$$

The term inside the sum for $\alpha_1, \ldots, \alpha_4$ is zero unless $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha$, and in that case its contribution is $|\hat{p}_\alpha|^4$. This finishes the proof of the lemma. □

We now use this relation between Δp and the Fourier coefficients of p to show that the expected value of Δp is small.

Lemma 7.6. $|\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}'' \in U}\left[e(\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}''))\right]| < \epsilon^4/100$.

Proof. We use Lemma 7.5. We have

$$\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}'' \in U}\left[e(p(\mathbf{x}' + \mathbf{y}') - p(\mathbf{x}' + \mathbf{y}'') - p(\mathbf{x}'' + \mathbf{y}') + p(\mathbf{x}'' + \mathbf{y}''))\right] = \sum_{\alpha \in \mathbb{F}^n} |\hat{p}_\alpha|^4 .$$

We now combine the fact that $\sum_{\alpha \in \mathbb{F}^n} |\hat{p}_\alpha|^2 = 1$ and our assumption that $|\hat{p}_\alpha| < \epsilon^2/10$ for all $\alpha \in \mathbb{F}^n$, to yield the required bound. □

Combining Lemmas 7.4 and 7.6 we get that

$$|\mathbb{E}_{\mathbf{x},\mathbf{y} \in U}[e(p(\mathbf{x} + \mathbf{y}))]| < \left(\frac{\epsilon^4}{100}\right)^{1/4} < \frac{\epsilon}{2} .$$

We now move on to handle the pseudorandom case. We start with the following observation:

Observation 7.2. *The polynomial $\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}'')$ has total degree-d, but has no degree-d terms which have variables from only one of $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{y}'$, $\mathbf{y}''$. Therefore, the total degree of variables from $\mathbf{x}'$ in each term is at most $d - 1$. The same is true for also $\mathbf{x}''$, $\mathbf{y}'$ and $\mathbf{y}''$.*

We now show that if D is a distribution that is pseudorandom against degree-$(d-1)$ polynomials, then it is also pseudorandom against Δp. We use a hybrid argument similar to the one in [BV07].

Lemma 7.7. *Let D be a distribution that is pseudorandom against degree-$(d-1)$ polynomials with error δ. Then*

$$|\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}'' \in D}\left[e(\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}''))\right] - \mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}'' \in U}\left[e(\Delta p(\mathbf{x}', \mathbf{x}'', \mathbf{y}', \mathbf{y}''))\right]| < 4\delta .$$

Proof. We change the inputs $\mathbf{x}', \mathbf{x}'', \mathbf{y}'$ and $\mathbf{y}''$ from U to D, one at a time. We prove that the expected value of $e(\Delta p)$ changes by at most δ in each step, accumulating to a total of at most 4δ. Formally, let H_k $(k = 0, \ldots, 4)$ be the joint distribution of $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{y}'$, $\mathbf{y}''$, when

the first k are taken from D and the last $4-k$ are taken from U. For example, H_1 is the distribution where $\mathbf{x}' \in D$ and $\mathbf{x}''$, $\mathbf{y}'$, $\mathbf{y}'' \in U$, where $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{y}'$, $\mathbf{y}''$ are independent.

We prove that the distance between $e(\Delta p)$ under H_{k-1} and H_k is at most δ, for all $k = 1, 2, 3, 4$. For the sake of clarity, we focus on the proof for $k = 1$. The proof for the other three cases is essentially identical.

For $k = 1$, we want to show that

$$|\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right] - \mathbb{E}_{\mathbf{x}'\in D,\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right]| < \delta\,.$$

The joint distribution of $\mathbf{x}'', \mathbf{y}', \mathbf{y}''$ is identical in both terms, so we have

$$|\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right] - \mathbb{E}_{\mathbf{x}'\in D,\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right]| \leq \mathbb{E}_{\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in U}|\mathbb{E}_{\mathbf{x}'\in U}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right] - \mathbb{E}_{\mathbf{x}'\in D}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right]|\,.$$

Now, for any fixing of values for $\mathbf{x}'' = a$, $\mathbf{y}' = b$, $\mathbf{y}'' = c$, $\Delta p(\mathbf{x}', a, b, c)$ is a polynomial just in $\mathbf{x}'$. Observation 7.2 tells us that it is a polynomial of degree at most $d-1$. Since D is pseudorandom against degree-$(d-1)$ polynomials, the inequality follows for every fixing of $\mathbf{x}''$, $\mathbf{y}'$, $\mathbf{y}''$. Hence, it also follows for the expected value. □

If we take D to be a PRG against degree-$(d-1)$ polynomials with error $\epsilon^4/400$ and combine this with Lemmas 7.4 and 7.6, we get that

$$|\mathbb{E}_{\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''\in D}\left[e(\Delta p(\mathbf{x}',\mathbf{x}'',\mathbf{y}',\mathbf{y}''))\right]| < \frac{\epsilon^4}{100} + 4\frac{\epsilon^4}{400} = \frac{\epsilon^4}{50}\,,$$

and so using Lemma 7.4 we get that

$$|\mathbb{E}_{\mathbf{x},\mathbf{y}\in D}[e(p(\mathbf{x}+\mathbf{y}))]| < \left(\frac{\epsilon^4}{50}\right)^{1/4} < \frac{\epsilon}{2}\,.$$

This finishes the proof of Lemma 7.2.

7.5 Case II: Some large Fourier coefficient exists

In this section we prove Lemma 7.3. We assume throughout this section that p has some large Fourier coefficient. To be precise, there exists some $\alpha \in \mathbb{F}^n$ such that

$$|\hat{p}_\alpha| \geq \frac{\epsilon^2}{10}\,.$$

Let $\ell(\mathbf{x})$ be the corresponding linear function, i.e., $\ell(\mathbf{x}) = \langle \mathbf{x}, \alpha \rangle$. Define

$$\eta = \overline{\hat{p}_\alpha} = \mathbb{E}_{\mathbf{x}\in U}\left[e(\ell(\mathbf{x}) - p(\mathbf{x}))\right]\,.$$

η is a measure for the approximation of $p(\mathbf{x})$ by $\ell(\mathbf{x})$. By our assumption on $\hat{p}_\alpha$, we know that $|\eta| \geq \epsilon^2/10$. For any constant $\mathbf{a} \in \mathbb{F}^n$ define the polynomial

$$q_{\mathbf{a}}(\mathbf{x}) = p(\mathbf{x}) - p(\mathbf{x}+\mathbf{a}) + \ell(\mathbf{x}+\mathbf{a})\,.$$

Notice that $q_{\mathbf{a}}(\mathbf{x})$ has degree at most $d-1$, because $\ell(\mathbf{x}+\mathbf{a})$ is linear (and so of degree less than d), and the degree-d terms in $p(\mathbf{x})$ and $p(\mathbf{x}+\mathbf{a})$ cancel out.

We can think of $q_{\mathbf{a}}(\mathbf{x})$ as using $\ell(\mathbf{x})$, which approximates $p(\mathbf{x})$ non-uniformly, and the derivative of $p(\mathbf{x})$ in a random direction $\mathbf{a}$, to build a random degree-$(d-1)$ polynomial which approximates $p(\mathbf{x})$ uniformly. In order to show this formally, we define

$$\nu_{\mathbf{x}}(\mathbf{a}) = \frac{1}{\eta} e(q_{\mathbf{a}}(\mathbf{x}))\,,$$

and prove that $\nu_{\mathbf{x}}(\mathbf{a})$, taken on a random $\mathbf{a} \in \mathbb{F}^n$ value, is exactly $e(p(\mathbf{x}))$.

Lemma 7.8. *For every $\mathbf{x} \in \mathbb{F}^n$, $\mathbb{E}_{\boldsymbol{a}\in U}[\nu_{\mathbf{x}}(\boldsymbol{a})] = e(p(\mathbf{x}))$.*

Proof. $\mathbb{E}_{\mathbf{a}\in U}[\nu_{\mathbf{x}}(\mathbf{a})] = \frac{1}{\eta} e(p(\mathbf{x}))\mathbb{E}_{\mathbf{a}\in U}[e(\ell(\mathbf{x}+\mathbf{a}) - p(\mathbf{x}+\mathbf{a}))] = e(p(\mathbf{x}))$. □

Effectively, we have shown that $p(\mathbf{x})$ can be approximated uniformly by a (random) degree-$(d-1)$ polynomial $q_{\mathbf{a}}(\mathbf{x})$. We can now use this to show that a distribution that is pseudorandom against degree-$(d-1)$ polynomials is also pseudorandom against p. First, we prove the following lemma:

Lemma 7.9. *Let D be a distribution that is pseudorandom against degree-$(d-1)$ polynomials with error δ. For every $\boldsymbol{a} \in \mathbb{F}^n$*

$$|\mathbb{E}_{\mathbf{x}\in D}[\nu_{\mathbf{x}}(\boldsymbol{a})] - \mathbb{E}_{\mathbf{x}\in U}[\nu_{\mathbf{x}}(\boldsymbol{a})]| < \frac{\delta}{|\eta|}\,.$$

Proof. We have

$$|\mathbb{E}_{\mathbf{x}\in D}[\nu_{\mathbf{x}}(\mathbf{a})] - \mathbb{E}_{\mathbf{x}\in U}[\nu_{\mathbf{x}}(\mathbf{a})]| = \frac{1}{|\eta|}\,|\mathbb{E}_{\mathbf{x}\in D}[e(q_{\mathbf{a}}(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}\in U}[e(q_{\mathbf{a}}(\mathbf{x}))]| < \frac{\delta}{|\eta|}\,,$$

where we use the fact that $q_{\mathbf{a}}$ is a polynomial of degree at most $d-1$ and so D is pseudorandom against $q_{\mathbf{a}}$ with error δ. □

We now conclude by proving Lemma 7.3.

Proof of Lemma 7.3. Let D be a distribution that is pseudorandom against degree-$(d-1)$ polynomials with error $\epsilon^3/10$. Then

$$\begin{aligned}|\mathbb{E}_{\mathbf{x}\in D}[e(p(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}\in U}[e(p(\mathbf{x}))]| &= |\mathbb{E}_{\mathbf{x}\in D}\mathbb{E}_{\mathbf{a}\in U}[\nu_{\mathbf{x}}(\mathbf{a})] - \mathbb{E}_{\mathbf{x}\in U}\mathbb{E}_{\mathbf{a}\in U}[\nu_{\mathbf{x}}(\mathbf{a})]| \\ &\leq \mathbb{E}_{\mathbf{a}\in U}\,|\mathbb{E}_{\mathbf{x}\in D}[\nu_{\mathbf{x}}(\mathbf{a})] - \mathbb{E}_{\mathbf{x}\in U}[\nu_{\mathbf{x}}(\mathbf{a})]| < \frac{\epsilon^3/10}{|\eta|} \leq \epsilon\,.\end{aligned}$$

□

Chapter 8

Explicit lower bound for fooling polynomials by the sum of small-bias generators

Recently, Viola (CCC'08) showed that the sum of d small-bias distributions fools degree-d polynomial tests; that is, every polynomial expression of degree at most d in the bits of the sum has distribution very close to that induced by this expression evaluated on uniformly selected random bits. We show that this is tight by showing an explicit construction of a small-bias generator (with exponentially small bias), and an explicit degree $d+1$ polynomial, that is distributed almost uniformly on random input, but always takes the value zero when evaluated on the sum of d independent copies of this generator.

Joint work with Yoaz Tzur.

8.1 Introduction

Small-bias distributions, first defined by Naor and Naor [NN93], are distributions over $\{0,1\}^n$ which are designed to fool all linear tests; any non-zero linear functional in the bits is distributed almost uniformly. A natural generalization are distributions that fool higher degree polynomials. A result of Viola [Vio08], improving over previous works of Bogdanov and Viola [BV07] and Lovett [Lov08], yields a general method for obtaining such distributions using any small-bias distribution. In order to construct a distribution that fools polynomials of degree at most d, take the bitwise sum of d independent samples of any small-bias distribution.

This result has been shown in [BV07] to be essentially tight with respect to the number of copies needed; using a counting argument, they show that for a fixed bias, any generator with output length ℓ that fools all degree $d+1$ polynomials must have seed length $(d+1)\cdot\log\ell - O(1)$. Thus, for every generator with shorter seed, there *exists* a polynomial expression of degree at most $d+1$ that distinguishes a random output of the generator from truly random bits. For a suitable choice of $\epsilon = o(1)$, the total length of d independent seeds for a standard construction of an ϵ-biased generator is still small enough, giving that in general the sum of

d small-bias generators does not fool polynomials of degree $d+1$.

Alon et al. [ABEK08] showed almost tight lower bounds for the size of the sample space required to fool all degree d polynomials with given error ϵ. Their bounds relate to the size of a general sample space, and not to the specific construction we are interested in, the bitwise sum of d independent samples of a small-bias distribution.

All the previous bounds metinoned are non-explicit, in the sense they prove the existence of a degree d polynomial which is not fooled by a small enough sample space. The main goal of this work is to prove such lower bounds explicitly.

Our main result is an explicit construction of a small-bias generator, and an explicit polynomial of degree $d+1$, such that this polynomial always evaluates to zero on inputs which are sums of d copies of the small-bias generator, and is almost uniform when evaluated over uniform inputs. Furthermore, our small-bias generator construction allows for exponentially small bias, whereas the proof of [BV07] allows only polynomially small bias.

Theorem 8.1. *For every $n, d \in \mathbb{N}$ and $\ell \geq 2d+1$, there exists an explicit small-bias generator $F : \{0,1\}^{2n} \to \{0,1\}^{\ell n}$ with bias $\epsilon \leq \ell/2^n$ and with the following property. Let $G : (\{0,1\}^{2n})^d \to \{0,1\}^{\ell n}$ be the sum of d independent samples of F, that is*

$$G_i(s) = G_i(s_1, \ldots, s_d) = F_i(s_1) \oplus \ldots \oplus F_i(s_n) \qquad (1 \leq i \leq \ell n).$$

Then there exists an explicit polynomial $p(x_1, \ldots, x_{\ell n})$ of degree $d+1$ over $GF(2)$ such that

- *The polynomial p evaluates to zero on any output of G,*

$$p(G_1(s), \ldots, G_{\ell n}(s)) \equiv 0 \qquad \forall s \in (\{0,1\}^{2n})^d$$

- *The polynomial p is almost uniform on uniform inputs,*

$$\frac{1}{2} - \frac{d}{2^n} \leq \Pr_{x \in \{0,1\}^{\ell n}}[p(x_1, \ldots, x_{\ell n}) = 0] \leq \frac{1}{2} + \frac{d}{2^n}$$

The paper is organized as follows. We give some required definitions in Section 8.2. We present the construction of the small-bias generator and the distinguishing polynomial in Section 8.3. Our results extend naturally to larger prime fields. We discuss this in Section 8.4.

8.2 Preliminaries

8.2.1 Definitions

Let $GF(q)$ denote the finite field of size q. We begin with the definition of a small-bias distribution.

Definition 8.1 (Small-bias distribution)**.** *For $\ell \in \mathbb{N}$, $\epsilon > 0$, a distribution D over $\{0,1\}^\ell$ is called* ϵ-biased *if for every nonzero $\alpha \in \{0,1\}^\ell$,*

$$\left| \Pr_{x \sim D}[\langle \alpha \rangle\, x = 0] - \frac{1}{2} \right| \leq \epsilon,$$

where $\langle \alpha \rangle\, x$ denotes the inner product $\sum_i \alpha_i x_i$ (over $GF(2)$).

The generalization to higher degree polynomials was first studied by [LVW93] (in fact, they considered the larger class of depth-2 boolean circuits), and further studied in [Bog05] (although there, only super-constant field sizes were considered).

Definition 8.2 (Fooling polynomials)**.** *For $d, \ell \in \mathbb{N}$, $\epsilon > 0$, a distribution D over $\{0,1\}^\ell$ is said to* ϵ-fool degree d polynomials *if for every ℓ-variate polynomial p over $GF(2)$ of total degree at most d,*

$$\left| \Pr_{x \sim D}[p(x) = 0] - \Pr_{x \sim U}[p(x) = 0] \right| \leq \epsilon,$$

where U denotes the uniform distribution over $\{0,1\}^\ell$.

We also view distributions as the outputs of pseudorandom generators.

Definition 8.3 (Small-bias generator)**.** *For $k, \ell \in \mathbb{N}$, $\epsilon > 0$, a mapping $G : \{0,1\}^k \to \{0,1\}^\ell$ is called an* ϵ-biased generator *of* stretch $\ell(k)$*, if the distribution induced by $G(s)$ for s selected uniformly in $\{0,1\}^k$ is ϵ-biased.*

Definition 8.4 (Pseudorandom generator for polynomials)**.** *For $d, k, \ell \in \mathbb{N}$, $\epsilon > 0$, a mapping $G : \{0,1\}^k \to \{0,1\}^\ell$ is said to* ϵ-fool degree d polynomials *if the distribution induced by $G(s)$ for s selected uniformly in $\{0,1\}^k$, ϵ-fools degree d polynomials. Again, we say that G has* stretch $\ell(k)$*.*

Remark. When discussing pseudorandom generators, it is common to also consider the complexity of the generator itself (as opposed to the complexity of potential distinguishers). This is quite secondary to the current work.

Despite our final interest in distributions over bits, we will also use generators over the larger field $GF(2^n)$ (see Subsections 8.2.2 and 8.3.3 for details). We first define the appropriate measure of distance between distributions.

Definition 8.5 (Statistical distance)**.** *For $\epsilon > 0$, two distributions X, Y are said to be* ϵ-close (in statistical distance) *if for every event E,*

$$\left| \Pr_X[E] - \Pr_Y[E] \right| \leq \epsilon.$$

Conversely, if there exists an event such that $|\Pr_X[E] - \Pr_Y[E]| \geq \epsilon$, then X and Y are said to be ϵ-far (in statistical distance)*.*

We now define generators over $GF(2^n)$.

Definition 8.6 ($GF(2^n)$-linear tests resilience)**.** *For $k, \ell \in \mathbb{N}$, $\epsilon > 0$, a mapping $G : GF(2^n)^k \to GF(2^n)^\ell$ is said to* ϵ-fool $GF(2^n)$-linear tests*, if for every vector $\alpha \in GF(2^n)^\ell$, and for s chosen uniformly in $GF(2^n)^k$, the distribution of the expression $\sum_{i=1}^{\ell} \alpha_i \cdot G_i(s)$, computed in the arithmetic of $GF(2^n)$, is ϵ-close to the uniform distribution over $GF(2^n)$.*

Definition 8.7 ($GF(2^n)$-polynomial tests resilience)**.** *For $d, k, \ell \in \mathbb{N}$, $\epsilon > 0$, a mapping $G : GF(2^n)^k \to GF(2^n)^\ell$ is said to* ϵ-fool $GF(2^n)$-polynomials of degree d *if for every polynomial $p \in GF(2^n)[x_1, \ldots, x_\ell]$, the following two distributions are ϵ-close (in statistical distance):*

- *$p(G(s))$ for s chosen uniformly from $GF(2^n)^k$.*
- *$p(x)$ for x chosen uniformly from $GF(2^n)^\ell$.*

8.2.2 Representation of $GF(2^n)$

We identify n-bit vectors in $\{0,1\}^n$ with elements of $GF(2^n)$ in a standard representation scheme. We will explicitly specify, for each variable, whether it is seen as an element of the field $GF(2^n)$ or of the vector-space $\{0,1\}^n = GF(2)^n$. We will use the linearity properties of this representation scheme.

Fact 8.1 (linearity of the standard representation)**.** *The following three properties hold for the standard representation scheme of $GF(2^n)$ as vectors in $\{0,1\}^n$.*

- *Addition in the field $GF(2^n)$ corresponds to addition of the respective representations in the vector space $\{0,1\}^n$.*
- *Multiplication of elements in the field $GF(2^n)$ corresponds to a bilinear mapping of the respective representations in the vector space $\{0,1\}^n$. That is, for two vectors $x, y \in \{0,1\}^n$, every bit of the vector representing the multiplication of the two $GF(2^n)$-elements represented by x and y can be written as a $GF(2)$-bilinear form in x and y.*
- *Let $\Psi : GF(2^n) \to \{0,1\}$ be any nonzero linear mapping. Any linear mapping $\Phi : GF(2^n) \to \{0,1\}$ can be written uniquely as $\Phi(x) = \Psi(a \cdot x)$ for some $a \in GF(2^n)$ (the multiplication $a \cdot x$ is done using the arithmetic of $GF(2^n)$).*

Fact 8.1 follows from the representation of $GF(2^n)$ as the quotient $GF(2)[x]/\langle c(x)\rangle$ for $\langle c(x)\rangle$ being the ideal generated by some irreducible polynomial $c(x)$ of degree n. For details, see any standard algebra textbook (e.g. [BM65]) or Lemma 15 in [Tzu].

8.3 The construction

We give an explicit small-bias generator and show that the sum of d independent copies of this generator does *not* fool an explicit polynomial of degree $d+1$.

While we are interested in distributions and polynomial tests over bits, we will first construct a distribution and polynomial over $GF(2^n)$. Using the linearity of the representation, we will then obtain a distribution and a polynomial over bits.

8.3.1 The generator

We use the following geometric generator, considered in [Tzu], which is related to a well known construction of a small-bias generator [AGHP90].

Construction 8.1 (The geometric generator)**.** *For $n, \ell \in \mathbb{N}$, define a mapping $F : GF(2^n) \times GF(2^n) \to GF(2^n)^\ell$ by letting the i-th output element for input elements a, b be $F_i(a,b) = a \cdot b^i$, for $i = 0, ..., \ell - 1$ (using the arithmetic of $GF(2^n)$).*

The geometric generator fools linear tests over $GF(2^n)$.

Proposition 8.1. *The geometric generator $\frac{\ell-1}{2^n}$-fools $GF(2^n)$-linear tests.*

We sketch the proof below. For a complete proof, see Proposition 7 in [Tzu].

Proof. Observe that every fixed nonzero $GF(2^n)$-linear combination $\bar{\alpha}$ in the output of the geometric generator $F(a,b)$ is equal to $a \cdot q(b)$ where q is a nonzero polynomial (determined by $\bar{\alpha}$) of degree at most $\ell - 1$ over $GF(2^n)$. If b is not one of the (at most) $\ell - 1$ roots of q, then $a \cdot q(b)$ is distributed uniformly when a is selected uniformly. Thus the expression is distributed $\frac{\ell-1}{2^n}$-close to uniform over $GF(2^n)$. □

For our purposes, any $\ell \geq 2d+1$ would suffice. To get a final bias of ϵ over bits (see subsection 8.3.3), we choose $n = \log \ell + \log \frac{1}{\epsilon}$ (and note that ϵ can be $2^{-\Omega(\ell)}$ for $n = \Omega(\ell)$).

The element-wise sum of d instances of F gives the generator $G : GF(2^n)^{2d} \to GF(2^n)^\ell$ defined as $G_i(a_1, b_1, \ldots, a_d, b_d) = \sum_{j=1}^{d} a_j {b_j}^i$, using the arithmetic of $GF(2^n)$.

8.3.2 The distinguishing polynomial

We now present a polynomial D over $GF(2^n)$ of the first $2d+1$ output elements of G, denoted $g_0, ..., g_{2d}$, that has degree $d+1$, and show that while the output of this polynomial is close to uniform on uniform input, it always takes the value zero when applied to an output of G.

The polynomial $D(g_0, \ldots, g_{2d})$ will be defined as the determinant of the following $(d+1) \times (d+1)$ Hankel matrix:

$$A_g^{(d)} = \begin{pmatrix} g_0 & g_1 & \cdots & g_d \\ g_1 & g_2 & \cdots & g_{d+1} \\ \vdots & \vdots & & \vdots \\ g_d & g_{d+1} & \cdots & g_{2d} \end{pmatrix}$$

(that is, the (i,k)-th entry of $A_g^{(d)}$ is g_{i+k}).

Indeed, this is a polynomial over $GF(2^n)$ of degree $d+1$ in the output blocks of G. We first claim that it is close to uniform when applied to uniform input:

Lemma 8.1. *Let M be a random $m \times m$ Hankel matrix over a finite field $\mathbb{F}$ (i.e., the Hankel matrix defined by $M_{i,k} = y_{i+k}$ for $y_0, \ldots, y_{2m-2}$ chosen uniformly at random from $\mathbb{F}$). Then, the distribution of the determinant of M is $\frac{m-1}{|\mathbb{F}|}$-close to uniform (in statistical distance).*

Proof. We proceed by induction on m. For $m = 1$, $\det(M)$ is exactly the only element of M, chosen uniformly from $\mathbb{F}$. Now fix $m > 1$, and let x be the first (top-left) element of M, and $\bar{y} = (y_2, \ldots, y_{2m-2})$ denote the remaining elements (on the top row and rightmost column). Denote the submatrix resulting from removing the first row and column by M', and note that it only contains the elements $y_2, \ldots, y_{2m-2}$. We develop the determinant of M by the first row, and write $\det(M) = x \cdot \det(M') + f(\bar{y})$, for some function f of $\bar{y}$. By the induction hypothesis, $\det(M')$ is distributed $\frac{m-2}{|\mathbb{F}|}$-close to uniform, so $\Pr[\det(M') = 0] \leq \frac{1}{|\mathbb{F}|} + \frac{m-2}{|\mathbb{F}|} = \frac{m-1}{|\mathbb{F}|}$. For any fixed nonzero value of $\det(M') \neq 0$ and for any fixed value of $\bar{y}$, the function $\det(M)$ is a (nonconstant) affine function of the uniformly chosen x, implying that, conditioned on $\det(M') \neq 0$, the determinant of M is distributed uniformly in $\mathbb{F}$. □

Corollary 8.1. *For $g_0, \ldots, g_{2d}$ chosen uniformly at random from $GF(2^n)$, the distribution of $D(g_0, \ldots, g_{2d})$ is $\frac{d}{2^n}$-close to uniform (in statistical distance).*

On the other hand, the polynomial D is identically zero on the output of G.

Proposition 8.2. *For every seed $\bar{s} = (a_1, b_1, \ldots, a_d, b_d) \in GF(2^n)^{2d}$, the expression $D(G(\bar{s}))$ evaluates to zero.*

Proof. We will show that the matrix $A_g^{(d)}$ is singular for every seed, and thus the polynomial $D(g_0, \ldots, g_{2d}) = \det(A_g^{(d)})$ will always take the value zero when evaluated on an output of G. To show that $A_g^{(d)}$ is singular for any seed, we show that its columns are always linearly dependent. More specifically, we show that for any $b_1, \ldots, b_d \in GF(2^n)$ there exist $\lambda_0, \ldots, \lambda_d \in GF(2^n)$, not all zero, such that for all $0 \leq k \leq d$, it holds that $\sum_{i=0}^{d} \lambda_i g_{i+k} = 0$. Letting $\bar{c}_i = (g_i, \ldots, g_{i+d})^T$ denote the i-th column of $A_g^{(d)}$, this means that $\sum_{i=0}^{d} \lambda_i \bar{c}_i$ is the zero vector in $GF(2^n)^{d+1}$.

Consider the polynomial $\Lambda(x) = \prod_{j=1}^{d}(x - b_j)$, the degree d polynomial with roots $b_1, \ldots b_d$, and set each λ_i to be the coefficient of x^i in $\Lambda(x)$. Note that always $\lambda_d = 1$. Then, using the definition of G (i.e., $g_i = \sum_j a_j {b_j}^i$), we get for every $0 \leq k \leq \ell - d$:

$$\begin{aligned}
\sum_{i=0}^{d} \lambda_i g_{i+k} &= \sum_{i=0}^{d} \lambda_i \sum_{j=1}^{d} a_j {b_j}^{i+k} \\
&= \sum_{j=1}^{d} a_j {b_j}^k \cdot \sum_{i=0}^{d} \lambda_i {b_j}^i \\
&= \sum_{j=1}^{d} a_j {b_j}^k \cdot \Lambda(b_j),
\end{aligned}$$

which is 0 as the b_j's are all roots of $\Lambda(x)$. □

We have thus obtained, using the event $D = 0$ in Definition 8.5, that

Theorem 8.2 (D distinguishes G from random)**.** *For D the determinant of $A_g^{(d)}$, the distributions $D(U_\ell)$ and $D(G(U_{2d}))$ are $(1 - \frac{d+1}{2^n})$-far (in statistical distance), where U_k denotes the uniform distribution over $GF(2^n)^k$.*

8.3.3 A distribution over bits

Let $F' : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^{\ell \cdot n}$ decode the $2n$ input bits to two elements $a, b \in GF(2^n)$, and output the concatenation of the representations of the elements $F_1(a,b) \ldots F_\ell(a,b)$, where $F_i(a,b) \in \{0,1\}^n$ is the i-th output block of the geometric generator of Construction 8.1. We first argue that F' is a small-bias generator.

Claim 8.1. *F' is an $\frac{\ell-1}{2^n}$-biased generator.*

We sketch the proof below. For a complete proof we refer to Corollary 8 in [Tzu].

Proof. Let $\Phi : \{0,1\}^{\ell n} \to \{0,1\}$ be a nonzero linear combination. Identify $x \in \{0,1\}^{\ell n}$ as $(x_1, \ldots, x_\ell) \in GF(2^n)^\ell$. We can decompose Φ as $\Phi(x) = \Phi_1(x_1) + \ldots + \Phi_\ell(x_\ell)$ where each

$\Phi_i : GF(2^n) \to \{0,1\}$ is a linear mapping. Let $\Psi : GF(2^n) \to \{0,1\}$ be some nonzero linear combination. Using Fact 8.1 we can write each Φ_i as $\Phi_i(x_i) = \Psi(a_i x_i)$ for some $a_i \in GF(2^n)$. By linearity we get that $\Phi(x) = \Psi(a_1 x_1 + \ldots + a_\ell x_\ell)$. Since Φ is nonzero we get that not all $a_1, \ldots, a_\ell$ are zero. Applying Φ to F' we get that for any $s \in \{0,1\}^{2n}$

$$\Phi(F'(s)) = \Psi(a_1 F_1(s) + \ldots + a_\ell F_\ell(s)).$$

Using Proposition 8.1 we know that the distribution of $a_1 F_1(s) + \ldots + a_\ell F_\ell(s)$ is $\frac{\ell-1}{2^n}$-close to uniform. Thus, since the output of Ψ is uniform in $\{0,1\}$ for a uniform input, we get that

$$\left| \Pr_{s \in \{0,1\}^{2n}} [\Phi(F'(s)) = 0] - \frac{1}{2} \right| \leq \frac{\ell - 1}{2^n}.$$

□

Analogously, let $G' : \{0,1\}^{2d \cdot n} \to \{0,1\}^{\ell \cdot n}$ decode its $2dn$ input bits as $2d$ elements $a_1, b_1, \ldots, a_d, b_d \in GF(2^n)$, and output the concatenation of the bit strings representing the output elements of $G(a_1, b_1, \ldots, a_d, b_d)$.

Viola's result [Vio08] implies that G' fools polynomials of degree d; we will show an explicit polynomial of degree $d+1$ that distinguishes a random output of G' from a random element of $\{0,1\}^{\ell \cdot n}$. Having shown that the polynomial D, over $GF(2^n)$, acts significantly differently on an output of G than on random input, we will derive the explicit polynomial in the output bits of G'.

Lemma 8.2. *Fix an ℓ-variate polynomial $D : GF(2^n)^\ell \to GF(2^n)$ of degree d, and define the mapping $D' : \{0,1\}^{\ell \cdot n} \to \{0,1\}^n$ to treat its input as the representation of ℓ elements $x_1, \ldots, x_\ell \in GF(2^n)$, and output the vector representing $D(x_1, \ldots, x_\ell)$. Then, each of the n output bits of D' is a polynomial of degree at most d over $GF(2)$ in the $\ell \cdot n$ input bits.*

Proof. We will show the claim for a polynomial consisting of a single monomial; the general claim follows from the fact that addition in the field $GF(2^n)$ is exactly bitwise addition in the vector space $\{0,1\}^n$ (Fact 8.1). We proceed by induction on the degree d. For $d = 0$ the claim is immediate since D is constant. Now fix $d > 0$, and assume without loss of generality that $D(x_1, \ldots, x_\ell) = x_1 \cdot \ldots \cdot x_d$. By Fact 8.1, the representation of $D(x_1, \ldots, x_\ell)$ is a bilinear expression in the bits of the two vectors x_1 and y_1, where y_1 is the vector representing the multiplication $x_2 \cdot \ldots \cdot x_d$. By the induction hypothesis, every bit in y_1 is a polynomial of degree at most $d-1$ in the bits of $x_2, \ldots x_d$, so each bit of a bilinear form in x_1 and y_1 is a polynomial of degree at most d in the bits of $x_1, \ldots, x_d$. □

Finally, by combining Theorem 8.2 with Lemma 8.2, letting D' be the binary version of D (as in Lemma 8.2), and setting D'_1 to the first bit (say) of D', we obtain our main result.

Theorem 8.3 (D'_1 distinguishes G' from random)**.** *The polynomial $D'_1 : \{0,1\}^{\ell \cdot n} \to \{0,1\}$ has degree at most $d+1$ and satisfies*

$$\left| \Pr_{s \in \{0,1\}^{2d \cdot n}} [D'_1(G'(s)) = 0] - \Pr_{x \in \{0,1\}^{\ell \cdot n}} [D'_1(x) = 0] \right| \geq \frac{1}{2} - \frac{d}{2^n}.$$

Proof. By Lemma 8.2, D'_1 indeed has degree at most $d+1$.

By Corollary 8.1, the distribution of $D(x)$ is $\frac{d}{2^n}$-close to uniform over $GF(2^n)$ when x is chosen uniformly from $GF(2^n)^{\ell+1}$, and thus by definition the distribution of $D'(x)$ is $\frac{d}{2^n}$-close to uniform over $\{0,1\}^n$, where x is chosen uniformly from $\{0,1\}^{(\ell+1)\cdot n}$. Specifically, considering the event "first bit is zero" in Definition 8.5, we have

$$\Pr_{x\in\{0,1\}^{(\ell+1)\cdot n}}[D'_1(x)=0] \leq \frac{1}{2}+\frac{d}{2^n}.$$

On the other hand, by Proposition 8.2, $D(G(s))=0$ for every $s\in GF(2^n)^{2d}$, giving that $D'(G'(s))=0^n$ for every $s\in\{0,1\}^{2d\cdot n}$, and specifically

$$\Pr_{s\in\{0,1\}^{2d\cdot n}}[D'_1(G'(s))=0] = 1.$$

The theorem follows. □

8.4 Larger prime fields

Our construction of a small-bias generator and distinguishing polynomial generalize naturally to general prime finite fields, as does the result of [Vio08]. The following is the analogue of Definition 8.1 (see, e.g., [Eve91] or [GW97]).

Definition 8.8. *For $\ell\in\mathbb{N}$, $\epsilon>0$ and a prime q, a distribution X over $GF(q)^\ell$ is called* ϵ-biased *if for every nonzero $\alpha\in GF(q)^\ell$:*

$$\left|\mathbb{E}_{x\sim X}[e^{\langle x\rangle\alpha\cdot 2\pi i/q}]\right| \leq \epsilon,$$

where $\langle\alpha\rangle\, x$ denotes the inner product $\sum_i \alpha_i\cdot x_i$ over $GF(q)$, and the multiplication by $2\pi i/q$ is then done over the complex field $\mathbb{C}$.

Standard arguments give that in this case,

$$\left|\Pr_{x\sim D}[\langle x\rangle\,\alpha=0]-\frac{1}{q}\right| \leq \sqrt{q-1}\cdot\epsilon/2.$$

(see, e.g., Appendix B in [BV07]).

Generalizing the generator and the distinguishing polynomial in the obvious way, we get the following generalization of Theorem 8.1 to general prime finite fields.

Theorem 8.4. *Let $GF(q)$ be a prime finite field. For every $n,d\in\mathbb{N}$ and $\ell\geq 2d+1$, there exists an explicit small-bias generator $F: GF(q)^{2n}\to GF(q)^{\ell n}$ with bias $\epsilon\leq \ell/q^n$ and with the following property. Let $G:(GF(q)^{2n})^d\to GF(q)^{\ell n}$ be the sum of d independent samples of F, that is*

$$G_i(s)=G_i(s_1,\ldots,s_d)=F_i(s_1)+\ldots+F_i(s_n) \pmod q \qquad (1\leq i\leq \ell n).$$

Then there exists an explicit polynomial $p(x_1,\ldots,x_{\ell n})$ of degree $d+1$ over $GF(q)$ such that

- *The polynomial p evaluates to zero on any output of G,*

$$p(G_1(s), \ldots, G_{\ell n}(s)) \equiv 0 \qquad \forall s \in (GF(q)^{2n})^d$$

- *The distribution of p, when applied on uniform inputs, is $\frac{d}{q^n}$-close to the uniform distribution over $\mathbb{F}_q$.*

Chapter 9

Pseudorandom bit-generators for modular sums

We consider the following problem: for given n, M, produce a sequence $X_1, X_2, \ldots, X_n$ of *bits* that fools every linear test modulo M. We present two constructions of generators for such sequences. For every constant prime power M, the first construction has seed length $O_M(\log(n/\epsilon))$, which is optimal up to the hidden constant. (A similar construction was independently discovered by Meka and Zuckerman [MZ09]). The second construction works for every M, n, and has seed length $O(\log n + \log(M/\epsilon)\log(M\log(1/\epsilon)))$.

The problem we study is a generalization of the problem of constructing *small bias* distributions [NN93], which are solutions to the $M = 2$ case. We note that even for the case $M = 3$ the best previously known constructions were generators fooling general bounded-space computations, and required $O(\log^2 n)$ seed length.

For our first construction, we show how to employ recently constructed generators for sequences of elements of $\mathbb{Z}_M$ that fool small-degree polynomials (modulo M). The most interesting technical component of our second construction is a variant of the derandomized graph squaring operation of [RV05]. Our generalization handles a product of two distinct graphs with distinct bounds on their expansion. This is then used to produce pseudorandom-walks where each step is taken on a different regular directed graph (rather than pseudorandom walks on a single regular directed graph as in [RTV06, RV05]).

Joint work with Omer Reignold, Luca Trevisan and Salil Vadhan.

9.1 Introduction

Pseudorandomness is the theory of generating objects that "look random" despite being constructed using little or no randomness. A primary application of pseudorandomness is to address the question: *Are randomized algorithms more powerful than deterministic ones?* That is, how does randomization trade off with other computational resources? Can every randomized algorithm be converted into a deterministic one with only a polynomial slowdown (*i.e.,* does BPP = P) or with only a constant-factor increase in space (*i.e.,* does RL = L)? The study of both these questions has relied on pseudorandom bit generators that

fool algorithms of limited computational powers. In particular, generators that fool space-bounded algorithms [AKS, BNS, Nis92, INW94] were highly instrumental in the study of the RL vs. L problem (e.g. used in the best known derandomization of RL [SZ99]).

While the currently available space-bounded generators are extremely powerful tools, their seed length is still suboptimal. For example, if we want to fool a $\log n$-space algorithm then known generators require $\log^2 n$ truly random bits (the seed) in order to generate up to polynomially many pseudorandom bits. On the other hand, for several interesting special cases we do know generators with almost optimal seed length. The special case which serves as a motivation for our work is that of small-biased generators [NN93]. These generators produce n bits $X_1, X_2, \ldots, X_n$ that fool all linear tests modulo 2. In other words, for each subset T of the bits, the sum $\Sigma_{i \in T} X_i \mod 2$ is uniformly distributed up to bias ϵ. Explicit constructions of ϵ-biased generators are known with seed-length $O(\log(n/\epsilon))$, which is optimal up to the hidden constant [NN93]. Even though linear tests may seem very limited, ϵ-biased generators have turned out to be very versatile and useful derandomization tools [NN93, MNN94, HPS93, Nao92, AM95, AR94, BSSVW03, BV07, Lov08, Vio08].

Given the several applications of distributions that fool linear tests modulo 2, it is natural to consider the question of fooling modular sums for larger moduli. It turns out that the notion of small-biased generators can be generalized to larger fields. Such generators produce a sequence $X_1, X_2, \ldots, X_n$ of elements in a field $\mathbb{F}$ that fool every linear test over $\mathbb{F}$ [Kat89, AIK+90, RSW93, EGL+98, AM95].[1] In this work, instead, we consider a different generalization of ϵ-biased generators where we insist on *bit*-generators. Namely we would like to generate a sequence $X_1, X_2, \ldots, X_n$ of bits that fool every linear test modulo a given number M. For every sequence $a_1, a_2, \ldots, a_n$ of integers in $\mathbb{Z}_M = \{0, 1, \ldots, M-1\}$ we want the sum $\sum_i a_i X_i \mod M$ to have almost the same distribution (up to statistical distance at most ϵ) as in the case where the X_i's are uniform and independent random bits. (Note that this distribution may be far from the uniform distribution over $\mathbb{Z}_M$, particularly when only a few a_i's are nonzero.) It turns out that even for $M = 3$ and even if we limit all the a_i's to be either ones or zeros, the best generators that were known prior to this work are generators that fool general space-bounded computations [Nis92, INW94], and required a seed of length $O(\log^2 n)$. Therefore, obtaining better pseudorandom bit generators that fool modular sums may be considered a necessary step towards improved space-bounded generators. In addition, we consider this notion to be a natural generalization of that of a small-bias generator, which is a central derandomization tool.

Our Results

We give two constructions of pseudorandom bit generators that fool modular sums. Similarly to [MST06], each construction is actually comprised of two generators: one that fools summations $\sum_i a_i X_i$ in which only relatively few coefficients a_i are nonzero (the "low-weight" case) and one that fools summations $\sum_i a_i X_i$ in which many coefficients a_i are nonzero (the

[1] More generally, an *ϵ-bias space* over a finite abelian group G is a distribution D on elements of G such that for every nontrivial character $\chi : G \to \mathbb{C}$, $|\mathbb{E}[\chi(D)]| \leq \epsilon$. The aforementioned results correspond to the special case $G = \mathbb{F}^n$, using the fact that the characters of $\mathbb{F}^n$ are in one-to-one correspondence with linear functions $\mathbb{F}^n \to \mathbb{F}$.

"high weight" case). The motivation is that fooling low-weight sums and fooling high-weight sums are tasks of a different nature. In the high-weight case, if R_i are truly random bits, then $\Sigma_i a_i R_i \mod M$ is almost uniformly distributed in $\mathbb{Z}_M$ (at least when M is prime). Thus, in analyzing our generator, we just need to argue that $\Sigma_i a_i X_i \mod M$ is close to uniform, where $X_1, \ldots, X_n$ is the output of the generator.

On the other hand, in the low-weight case the distribution may be far from uniform and therefore we may need to imitate the behavior of a random sequence of bits more closely.

Thus, in each construction, we shall present two generators: one that is pseudorandom against low-weight sums, and one that is pseudorandom against high-weight sums. We shall then combine them by evaluating them on independently chosen seeds and XORing the two resulting sequences.

Construction Based on Pseudorandom Generators for Polynomials

In our first construction, we handle the case of $M = 3$ and any other fixed prime modulus M (in fact, our construction works also for any fixed prime power). For these cases, our seed length is $O(\log(n/\epsilon))$ as in the case of ϵ-biased generators (but the hidden constant depends exponentially on M).

As mentioned above, for every fixed finite field $\mathbb{F}$, there are nearly-optimal known generators that construct a small-bias distribution $X_1, \ldots, X_n$ of *field elements*, while our goal is to generate *bits*. A natural approach to construct a bit generator would be to sample a sequence of field elements $X_1, \ldots, X_n$ from a small-bias distribution, and output a bit-sequence $g(X_1), \ldots, g(X_n)$ for an appropriate function $g : \mathbb{F} \to \{0, 1\}$. Unfortunately the pseudorandomness of $g(X_1), \ldots, g(X_n)$ against $\mathbb{F}$-linear tests does not seem to follow from the small-bias property of $X_1, \ldots, X_n$. Indeed, when $|\mathbb{F}|$ is odd, then g cannot be balanced, so at best we could hope is for $g(X_1), \ldots, g(X_n)$ to be indistinguishable by linear tests from a sequence of independent *biased* bits. But even this is not achievable in general, if we only assume the pseudorandomness of $X_1, \ldots, X_n$ against $\mathbb{F}$-linear tests(as per the definition of small-bias space).[2]

If, however, we start from a sequence of field elements $X_1, \ldots, X_n$ that fools *polynomials* over $\mathbb{F}$, then we can indeed show that $g(X_1), \ldots, g(X_n)$ is indistinguishable by linear tests from independent biased bits. The reason is that g can be chosen to be itself a polynomial (of degree $d = \Theta(|\mathbb{F}|)$), and thus any $\mathbb{F}$-linear test distinguisher on $g(X_1), \ldots, g(X_n)$ yields a degree d distinguisher on $X_1, \ldots, X_n$. Since we still only have indistinguishability from *biased* coins, we only apply this approach when the coefficient vector has sufficiently high weight so that both biased and unbiased random bits will yield a sum that is almost uniformly distributed over $|\mathbb{F}|$. Specifically, we need at least k non-zero coefficients a_i, where $k = O(M^2 \log 1/\epsilon)$. For fixed M, there are known constructions [BV07, Lov08, Vio08] of pseudorandom generators that fool polynomials of degree d over $\mathbb{F} = \mathbb{Z}_M$, M prime, and which only require seed length $O_{M,d}(\log n/\epsilon)$.

[2] Let $\mathbb{F} = \mathbb{Z}_3$, and $g : \mathbb{Z}_3 \to \{0, 1\}$ be any nonconstant function. Let a be the element of $\mathbb{Z}_3$ such that a is the unique preimage of $g(a)$. Let $(X_1, \ldots, X_n)$ be uniformly distributed over all elements of $\mathbb{Z}_3^n$ where the number of a's is divisible by 3. Then $\sum_i g(X_i) \mod 3$ is constant, but it can be shown that $(X_1, \ldots, X_n)$ is a $2^{-\Omega(n)}$-biased space.

In order to fool low-weight sums, we observe that a bit generator $X_1, \ldots, X_n$ which is ϵ-almost k-wise independent fools, by definition, every sum $\sum_i a_i X_i \bmod M$ of weight at most k, and that such generators are known which require only seed length $O(\log n + k + \log 1/\epsilon)$.

A similar construction was independently discovered by Meka and Zuckerman [MZ09].

Construction Based on the INW Generator

In our second construction, we give a pseudorandom bit generator that fools sums modulo *any* given M (not necessarily prime) with seed length $O(\log n + \log(M/\epsilon)\log(M\log(1/\epsilon)))$. In both the low-weight and high-weight cases, this generator relies on versions of the Impagliazzo–Nisan–Wigderson [INW94] pseudorandom generator for space-bounded computation. Of course, modular sums are a special case of space-bounded computations, and thus we could directly apply the INW generator. But this would require seed length larger than $\log^2 n$. We obtain better bounds by more indirect use of the INW generator inside our construction.

The most interesting technical contribution underlying this construction is a new analysis of the derandomized graph squaring operation of [RV05], which captures the effect of using the INW generator to derandomize random walks on graphs. Here we study the analogue of derandomized squaring for taking products of two distinct Cayley graphs over an abelian group (namely $\mathbb{Z}_M$). The advantage of the new analysis is that it handles graphs that have distinct bounds on their expansion, and works for bounding each eigenvalue separately. This is then used to produce pseudorandom walks where each step is taken on a different abelian Cayley graph (rather than pseudorandom walks on a single graph as in [RTV06, RV05]).

For the purpose of this informal discussion we will assume that M is prime. (The idea for handling composite M's is to analyze each Fourier coefficient of the distribution of the sum separately. We defer further details to Section 9.2.1.)

Low-Weight Case. Let us first consider the case where the number of non-zero a_i's is at most $M' \cdot \log(1/\epsilon)$, for $M' = \mathrm{poly}(M)$.[3] As before, we could use an almost k-wise independent distribution, but then our seed length would depend polynomially on M, while our goal is a polylogarithmic dependency.

First, we use a hash function to split the index set $[n] = \{1, 2, \ldots, n\}$ into $B = O(M')$ disjoint subsets T_j such that with high probability (say, $1 - \epsilon/10$) over the splitting, each set T_j contains at most $k = \log(1/\epsilon)$ indices i such that $a_i \neq 0$. We show that the selection of the hash function that determines the splitting can be done using $O(\log n + (\log M/\epsilon) \cdot \log(M \log 1/\epsilon))$ random bits.

Once we have this partition, it is sufficient to independently sample in each block from an ϵ/B-almost k-wise independent distribution, which requires $s = O(\log n + k + \log(B/\epsilon)) = O(\log n + \log(M/\epsilon))$ random bits per block. Then we argue that it is not necessary for the sampling in different blocks to be independent, and instead they can be sampled using a pseudorandom generator for space-bounded computation [Nis92, INW94]. (This relies on the fact the computation $\sum_i a_i X_i \bmod M$ can be performed in any order over the i's, in

[3] In this preliminary version we did not try to optimize the various constants. In particular, in our analysis $M' = O(M^{24})$. We note that it can be made as small as $O(M^{2+\alpha})$ for any $\alpha > 0$.

particular the order suggested by $\sum_j \sum_{i \in T_j} a_i \cdot X_i \bmod M$.) Using the INW generator, we can do all the sampling using $O(s + \log B \cdot (\log(B/\epsilon) + \log M)) = O(\log n + \log M \cdot \log(M/\epsilon))$ random bits.

High-Weight Case. We now discuss the generator that fools sums with more than $M' \cdot \log 1/\epsilon$ non-zero coefficients a_i, for $M' = \mathrm{poly}(M)$. Here, we can think of the computation $\sum_i a_i X_i \mod M$ as an n-step walk over $\mathbb{Z}_M$ that starts at 0. Unlike standard walks, *each step is taken on a different graph* (over the same set of vertices, namely $\mathbb{Z}_M$). Specifically, step i is taken on the (directed) Cayley graph where every node v has two outgoing edges. The first edge is labeled 0 and goes into v itself (*i.e.,* this edge is a self loop). The second edge is labeled 1 and goes into $v + a_i \mod M$. Following the walk along the labels $X_1, X_2, \ldots, X_n$ arrives at the vertex $\sum_i a_i X_i \mod M$. If the X_i's are uniform (*i.e.,* we are taking a random walk) then the end vertex will be almost uniformly distributed (because the number of steps is larger than $M^2 \cdot \log(1/\epsilon)$). What we are seeking is a pseudorandom walk that is generated using much fewer truly random bits but still converges to the uniform distribution (possibly slower, e.g. using $M' \cdot \log(1/\epsilon)$ steps).

Pseudorandom walk generators were constructed in [RTV06, RV05] for walks on a single regular and connected graph. In our case, we are walking not on a single graph but rather on a sequence of graphs, each of which is indeed regular. It turns out that the pseudorandom generators of [RTV06, RV05] still work for a sequence of graphs rather than a single graph. The more difficult aspect is that in our walk there is no uniform bound on the expansion of the graphs. Indeed, the graphs that correspond to $a_i = 0$ are not connected at all (they consist solely of self loops). In our setting, where the graphs are directed Cayley graphs for the abelian group $\mathbb{Z}_M$, we show how to generate pseudorandom walks on graphs with varying bounds on expansion.

We do this by a generalization of the derandomized graph product of [RV05]. There, expanders are used to generate two steps on a degree-D graph using less than $2 \log D$ random bits, yet the (spectral) expansion of the resulting graph is almost as good as the square of the original graph. We analyze the analogous derandomization of two steps on two distinct (abelian Cayley) graphs for which we may have distinct bounds on their expansion. Moreover, to handle composite M, we show that the expansion can be analyzed in each eigenspace separately. (For example, for $\mathbb{Z}_6 = \mathbb{Z}_2 \times \mathbb{Z}_3$, a sequence of even coefficients a_i will yield a random walk that does not mix in the $\mathbb{Z}_2$ component, but may mix in the $\mathbb{Z}_3$ component, and our pseudorandom generator needs to preserve this property.)

To obtain our pseudorandom walk generator, we first randomly reorder the index set $[n]$ so that the nonzero coefficients are well-spread out, and then derandomize the walk by a recursive application of our aforementioned derandomized product. As discussed in [RV05], the resulting pseudorandom walk generator is the same as the Impagliazzo–Nisan–Wigderson [INW94] generator for space-bounded computation, with a different setting of parameters that enables a much smaller seed length than their analysis requires for general space-bounded algorithms.

Discussion

The natural open problem left by our work is to reduce the seed length further, ideally to $O(\log(nM/\epsilon))$, which can be shown to be possible via a nonconstructive probabilistic argument. For achieving such optimal parameters, the modular reduction is actually insignificant — it is equivalent to construct generators such that for every bounded coefficient vector $(a_1, \ldots, a_n) \in \mathbb{Z}^n$ where each $|a_i| \leq M$, $\sum_i a_i X_i$ is statistically close to $\sum_i a_i R_i$ as distributions on $\mathbb{Z}$, where $(X_1, \ldots, X_n)$ is the output distribution of the generator, and $(R_1, \ldots, R_n)$ is the uniform distribution on $\{0,1\}^n$. [4] As a result, such generators would also "fool" linear threshold functions (halfspaces) whose coefficients are polynomially bounded. Pseudorandom generators and related objects for threshold functions (with no bound on the coefficients) have recently been studied in [RS09, DGJ+09], with the latter achieving seed length $O((\log n) \cdot \log^2(1/\epsilon)/\epsilon^2)$.

9.2 Definitions and Tools

We denote by U_n the uniform distribution over $\{0,1\}^n$. We fix an integer $M \geq 2$ for the rest of the paper. We will be interested in constructing pseudorandom bit generators that fool sums modulo M. We denote by $\mathbb{Z}_M$ the set $\{0, 1, \ldots, M-1\}$ with arithmetic modulo M. Due to space limitations, we defer many of the proofs to the full version of the paper.

Definition 9.1 (Statistical Distance)**.** *The* statistical distance *between two random variables X, Y taking values in $\mathbb{Z}_M$ is* $\mathrm{dist}(X,Y) = \frac{1}{2}\sum_{i=0}^{M-1} |\Pr[X=i] - \Pr[Y=i]|$. *The variables X and Y are said to be* ϵ-close *if their statistical distance is at most ϵ.*

Definition 9.2 (pseudorandom distributions against modular sums)**.** *A random variable $X = (X_1, \ldots, X_n)$ taking values in $\{0,1\}^n$ is* ϵ-pseudorandom against sums modulo M *if for any $a_1, \ldots, a_n \in \mathbb{Z}_M$, the distribution of $a_1X_1 + \cdots + a_nX_n$* modulo M*, is ϵ-close (in statistical distance) to the distribution $a_1R_1 + \cdots + a_nR_n$* modulo M*, where $R_1, \ldots, R_n$ are uniform and independent random bits.*

Definition 9.3 (pseudorandom bit generators against modular sums)**.** *A function $G : \{0,1\}^r \to \{0,1\}^n$ is an* ϵ-pseudorandom bit generator against sums modulo M *if the distribution $G(U_r)$ is ϵ-pseudorandom against sums modulo M.*

Note that ϵ-biased generators is a special case of the definition of pseudorandom bit generators against sums modulo M, for $M = 2$.

Our goal is to build generators that fool sums modulo M, where M can be either prime or composite. Handling prime modulus is somewhat easier, and the approach in the following section allows handling both cases simultaneously. We will show that it is enough to construct pseudorandom generators which fools the bias of a sum modulo M, and under this approach, there is no major difference between primes and composites.

[4] Indeed, given any coefficient vector $(a_1, \ldots, a_n) \in \mathbb{Z}^n$, where each $|a_i| \leq M$, we can apply the generator for modulus $M' = M \cdot n$ so that no modular reduction occurs.

9.2.1 Small Bias Bit Generators

First we define the *bias* of a linear combination with coefficients $a_1, \ldots, a_n \in \mathbb{Z}_M$, given some distribution of $X = (X_1, \ldots, X_n) \in \{0,1\}^n$:

Definition 9.4. *Let $X = (X_1, \ldots, X_n)$ be a distribution over $\{0,1\}^n$, and $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ a coefficient vector. We define the* bias *of $a_1, \ldots, a_n$ according to X to be*

$$\text{bias}_X(a_1, .., a_n) = \mathbb{E}\left[\omega^{\sum a_i X_i}\right]$$

where $\omega = e^{2\pi i/M}$ is a primitive M-th root of unity.

Notice that the bias can in general be a complex number, of absolute value at most 1.

Definition 9.5. *We say a distribution $X = (X_1, \ldots, X_n)$ over n bits is* ϵ-bit-biased against sums modulo M *if for every coefficient vector $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$,*

$$|\text{bias}_X(a_1, \ldots, a_n) - \text{bias}_{U_n}(a_1, \ldots, a_n)| \leq \epsilon$$

Let $G : \{0,1\}^r \to \{0,1\}^n$ be a bit generator. We shorthand $\text{bias}_G(a_1, \ldots, a_n)$ for $\text{bias}_{G(U^r)}(a_1, \ldots, a_n)$.

Definition 9.6. *$G : \{0,1\}^r \to \{0,1\}^n$ is an* ϵ-bit-biased generator against sums modulo M *if the distribution $G(U_r)$ is ϵ-bit-biased against sums modulo M. That is, for every coefficient vector $(a_1, \ldots, a_n)$,*

$$|\text{bias}_G(a_1, \ldots, a_n) - \text{bias}_{U_n}(a_1, \ldots, a_n)| \leq \epsilon$$

The name "bit-biased" in the above definitions is meant to stress the difference from standard ϵ-biased generators modulo M. Here we compare the bias under the generator to the bias under uniformly selected bits (rather than uniformly selected elements in $\mathbb{Z}_M$).

We first reduce the problem of constructing pseudorandom modular generators to that of constructing ϵ-bit-biased modular generators.

Lemma 9.1. *Let $X = (X_1, \ldots, X_n)$ be an ϵ-bit-biased distribution against sums modulo M. Then X is $(\epsilon\sqrt{M})$-pseudorandom against sums modulo M.*

From now on, we focus on constructing ϵ-bit-biased generators. We will need to differentiate two types of linear combinations, based on the number on non-zero terms in them.

Definition 9.7 (Weight of a coefficient vector)**.** *The* weight *of a coefficient vector $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ is the number of non-zero coefficients a_i.*

We will construct two generators: one fooling linear combination with small weights, and the other fooling linear combinations with large weight. Our final generator will be the be the bitwise-XOR of the two, where each is chosen independently. The following lemma shows this will result in an ϵ-bit-biased generator fooling all linear combinations.

Lemma 9.2. *Fix a weight threshold W. Let $X' = (X'_1, \ldots, X'_n)$ be a distribution over $\{0,1\}^n$ such that for any vector coefficient $a_1, \ldots, a_n$ of weight at most W,*

$$|\mathrm{bias}_{X'}(a_1, \ldots, a_n) - \mathrm{bias}_{U_n}(a_1, \ldots, a_n)| \leq \epsilon.$$

Let $X'' = (X''_1, \ldots, X''_n)$ be a distribution over $\{0,1\}^n$ such that for any vector coefficient $a_1, \ldots, a_n$ of weight at least W,

$$|\mathrm{bias}_{X'}(a_1, \ldots, a_n) - \mathrm{bias}_{U_n}(a_1, \ldots, a_n)| \leq \epsilon.$$

Let X be the bitwise-XOR of two independent copies of X' and X'', i.e.

$$X = X' \oplus X'' = (X'_1 \oplus X''_1, \ldots, X'_n \oplus X''_n).$$

Then X is ϵ-bit-biased against sums modulo M.

Convergence of the bias for large weights

The bias of a coefficient vector with respect to the uniform distribution can be large if there are only a few non-zero elements in the vector. However, when the weight is large, the bias is guaranteed to be small.

Lemma 9.3. *Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector of weight w. Then*

$$|\mathrm{bias}_U(a_1, \ldots, a_n)| \leq \left(1 - \frac{1}{M^2}\right)^w$$

In particular, for $w \geq M^2 \log(1/\epsilon)$ the bias is at most $\epsilon/2$.

Notice that the above lemma holds for all coefficient vectors $(a_1, \ldots, a_n)$ and moduli M, even when M is composite and the coefficients are not relatively prime to M. For example, when $M = 6$ and $(a_1, \ldots, a_n) = (2, \ldots, 2)$. In such a case, $\sum_i a_i R_i \mod M$ does not converge to the uniform distribution on $\mathbb{Z}_M^n$, but the above lemma still says that the bias tends to zero.

A similar result holds if we consider the bias of a large weight coefficient vector under a skewed distribution.

Lemma 9.4. *Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector of weight w. Let $Z_1, \ldots, Z_n \in \{0,1\}$ be independently distributed with $\Pr[Z_i = 0] = (1+\alpha)/2$. Then*

$$|\mathrm{bias}_{Z_1,\ldots,Z_n}(a_1, \ldots, a_n)| \leq \left(1 - \Omega\left(\frac{1-\alpha^2}{M^2}\right)\right)^w$$

In particular, for $w \geq cM^2 \log(1/\epsilon)/(1-\alpha^2)$ for a sufficiently large constant c, the bias is at most $\epsilon/2$.

9.2.2 Hashing

We use hashing as one of the ingredients in our construction. A family (multiset) of functions $\mathcal{H} = \{h : [n] \to [k]\}$ is called a family of hash functions, if a randomly chosen function from the family behaves pseudorandomly under some specific meaning. We consider a hash function $H : [n] \to [k]$ to be a random variable depicting a randomly chosen function from the family. We say H can be generated *efficiently and explicitly* using s random bits, if a random function in the family can be sampled by a randomized polynomial-time algorithm using s random bits, and this function can be evaluated using a deterministic polynomial-time algorithm.

Fix $S \subset [n]$. We define the j-th bucket of H with respect to S, to be the set of elements of S mapped by H into j, i.e. $\{s \in S : H(s) = j\} = H^{-1}(j) \cap S$.

We will use the following three constructions of hash functions.

Lemma 9.5. *Assume k is a power of 2. There exists a hash function $H_1 : [n] \to [k]$ such that for every set $S \subset [n]$ of size at most $k \log(1/\epsilon)$, the probability that H_1 has a bucket $H_1^{-1}(j) \cap S$ with more than $100 \log(1/\epsilon)$ elements is at most $\epsilon/100$. Moreover, H_1 can be generated explicitly and efficiently using $O(\log n + \log(k/\epsilon) \log(k \log(1/\epsilon)))$ random bits.*

Lemma 9.6. *Assume k is a power of 2. There exists a hash function $H_2 : [n] \to [k]$ such that for every $S \subset [n]$ of size at least $100k^2$, the probability that H_2 has an empty bucket $H_2^{-1}(j) \cap S$ is at most $1/100$. Moreover, H_2 can be generated explicitly and efficiently using $O(\log n + \log^2 k)$ random bits.*

Lemma 9.7. *There exists a hash function $H_3 : [n] \to [16 \log(1/\epsilon)]$ such that for every $S \subset [n]$ of size at least $800k \log(1/\epsilon)$, the probability that H_3 has at least $\log(1/\epsilon)$ buckets $H_3^{-1}(j) \cap S$ with at most k elements is at most $\epsilon/100$. Moreover, H_3 can be generated explicitly and efficiently using $O(\log n + \log(1/\epsilon) \log(k \log(1/\epsilon)))$ random bits.*

The constructions of the hashes in Lemmas 9.5, 9.6 and 9.7 are based on almost t-wise independence. A sequence of random variables $X_1, \ldots, X_n \in \{0, 1\}$ is said to be *t-wise independent* if any t random variables in it are independent. It is said to be δ-almost t-wise independent if any t random variables in it are δ-close in statistical distance to independent. Explicit constructions of δ-almost t-wise independent distributions are known, with nearly optimal seed length [NN93, AGHP90].

We identify a function $h : [n] \to [\ell]$, where ℓ is a power of 2, by a sequence of $n \log \ell$ bits. We construct the hash functions by choosing the sequence of bits according to an δ-almost t-wise independent distribution, where the values of δ and t differ in the three constructions. The main tool in our analysis is a tail bound on t-wise independent distributions, due to Bellare and Rompel [BR], extended to the case of δ-almost t-wise distributions. We defer further details to the full version of the paper.

9.2.3 Pseudorandom generators for small space

An ingredient in our construction is the small-space pseudorandom generator of Impagliazzo, Nisan, and Wigderson [INW94]. We first define branching programs, which form a non-uniform model of small-space computations.

Definition 9.8 (Branching program)**.** *A* (read-once, oblivious) branching program *of length n, degree d and width w is a layered graph with $n+1$ layers, where each layer contains at most w vertices. From each vertex in the i-th layer ($1 \leq i \leq n$) there are d outgoint edges, numbered $0, 1, \ldots, d-1$. A vertex in the first layer is designated as the* start vertex*. Running the branching program on an input $x_1, \ldots, x_n \in [d]$ is done by following the path according to the inputs, starting at the start vertex. The* output *of the branching program is the vertex reached in the last layer.*

Definition 9.9 (Pseudorandom generator for branching programs)**.** *A* pseudorandom generator for branching programs *of length n, degree d and width w with error ϵ is a function $G : \{0,1\}^r \to [d]^n$, such that for every branching program of length n, degree d and width w, the statistical distance between the output of the branching program when run on uniform element in $[d]^n$, and the output when run on $G(U_r)$, is at most ϵ.*

Lemma 9.8. *[INW94] There exists an explicit pseudorandom generators for branching programs of length n, degree d, width w with error ϵ, which uses $r = O(\log d + (\log n)(\log(n/\epsilon) + \log w))$ truly random bits.*

9.3 Construction using PRG for low-degree polynomials

We present in this section a simple construction for prime powers M, based on pseudorandom generators for low-degree polynomials. This construction is optimal for constant M, achieving a pseudorandom generator with seed length $O_M(\log(1/\epsilon))$ (where the constant depends exponentially on M).

Let $W = \Omega(M^3 \log 1/\epsilon)$. We will construct two generators: one for coefficient vectors of weight at most W, and one for coefficient vectors of weight at least W. Lemma 9.2 shows that the bitwise-XOR of the two generators is a pseudorandom generator for all coefficient vectors.

For small weights, we will use a distribution that is ϵ-almost W-wise independent. Such a distribution trivially fools coefficient vectors of weight at most W. It can be explicitly generated using $O(\log n + W + \log 1/\epsilon) = O_M(\log n/\epsilon)$ random bits [NN93].

For large weights, let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector of weight at least W. Consider first the distribution of $a_1 R_1 + \ldots a_n R_n$ for independent and uniform bits $R_1, \ldots, R_n$. By Lemma 9.3, $|\mathrm{bias}_{U_n}(a_1, \ldots, a_n)| < \epsilon/2$.

Consider now $Z_i \in \{0,1\}$, where $\Pr[Z_i = 0] = c/M$ for some integer $1 \leq c \leq M-1$. By Lemma 9.4,

$$|\mathrm{bias}_{Z_1,\ldots,Z_n \sim (c/M, 1-c/M)}(a_1, \ldots, a_n)| < \epsilon/4,$$

given that $W = \Omega(M^3 \log(1/\epsilon))$ with a large enough hidden constant.

The benefit of using this skewed distribution, is that it can be simulated by low-degree polynomials modulo M. Since we assume M is a prime power, there is a polynomial $g : \mathbb{Z}_M \to \mathbb{Z}_M$ that maps some c elements of $\mathbb{Z}_M$ to 0, and the rest to 1. For example, if $M = p^k$, the polynomial $g(x) = x^{(p-1)p^{k-1}}$ maps elements divisible by p to 0, and the rest to 1. The degree of this g is at most $M-1$.

Let $Z_1, \ldots, Z_n \in \{0,1\}^n$ be generated by $g(Y_1), \ldots, g(Y_n)$, where $Y_1, \ldots, Y_n \in \mathbb{Z}_M$ are uniform and independent. We thus have:

$$|\mathrm{bias}_{Z_1,\ldots,Z_n \sim g(U_{Z_M})^n}(a_1, \ldots, a_n)| < \epsilon/4$$

Note that

$$\mathrm{bias}_{Z_1,\ldots,Z_n \sim g(U_{Z_M})^n}(a_1, \ldots, a_n) = \mathbb{E}_{Y_1,\ldots,Y_n \in \mathbb{Z}_M}[\omega^{a_1 g(Y_1) + \cdots + a_n g(Y_n)}],$$

and that $a_1 g(Y_1) + \cdots + a_n g(Y_n)$ is a polynomial of degree $\deg(g)$ in $Y_1, \ldots, Y_n$. Thus we can derandomize the choice of $Y_1, \ldots, Y_n$ using a a pseudorandom generator for low-degree polynomials [BV07, Lov08, Vio08]. We note the results in these papers are stated for polynomials over prime finite fields, but they hold also for polynomials over $\mathbb{Z}_M$, using small-bias spaces for $\mathbb{Z}_M^n$ [Kat89, AIK$^+$90, RSW93, EGL$^+$98, AM95] as a building block.

Lemma 9.9. *For every $M, n, d \in \mathbb{N}$, there is an explicit generator $G : \{0,1\}^r \to \mathbb{Z}_M^n$ such that for every polynomial $f : \mathbb{Z}_M^n \to \mathbb{Z}_M$ of degree at most d, the distribution of $f(\mathbb{Z}_M^n)$ and $f(G(U_r))$ are ϵ-close in statistical distance. The number of random bits required is $r = O(d2^d \log(M/\epsilon) + d \log(nM))$.*

We use the generator of Lemma 9.9 for error $\epsilon/4$ and degree $d = M - 1$. We thus get an explicit generator whose output distribution $(Y_1', \ldots, Y_n') \in \mathbb{Z}_M^n$, such that:

$$|\mathbb{E}_{(Y_1',\ldots,Y_n')}[\omega^{a_1 g(Y_1') + \ldots + a_n g(Y_n')}] - \mathbb{E}_{Y_1,\ldots,Y_n \in \mathbb{Z}_M^n}[\omega^{a_1 g(Y_1) + \ldots + a_n g(Y_n)}]| < \epsilon/4$$

Thus, if we define our generator G' to output $g(Y_1'), \ldots, g(Y_n')$, we have $Y_1', \ldots, Y_n'$ are the output of G, we get an explicit generator,such that $|\mathrm{bias}_{G'}(a_1, \ldots, a_n)| < \epsilon/2$. Hence, we get that

$$|\mathrm{bias}_{G'}(a_1, \ldots, a_n) - \mathrm{bias}_G(a_1, \ldots, a_n)| < \epsilon$$

The randomness requirement of our generator comes directly from that of G, which is $O(M2^{M-1} \log(M/\epsilon) + M \log(nM)) = O_M(\log(n/\epsilon))$ for constant M.

9.4 Construction Based on Pseudorandom Walk Generators

9.4.1 A generator for small sums

We construct an ϵ-bit-biased generator for weights at most $W = 10^5 M^{24} \log(1/\epsilon)$. Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector of weight at most W.

The construction has three stages:

1. Partitioning the set of indices $[n]$ into W buckets using the hash function H_1. Lemma 9.5 guarantees that with probability at least $1 - \epsilon/100$, each bucket contains at most $O(\log(1/\epsilon))$ non-zero coefficients.

2. For each bucket j, generate the X_i's for i's in the j'th bucket using an almost $O(\log(1/\epsilon))$-wise independent distribution.

3. Use the INW generator given by Lemma 9.8 to generate the W seeds for the $O(\log(1/\epsilon))$-wise independent distributions used for the different buckets.

Lemma 9.10. *The above construction is an ϵ-bit-biased generator against coefficient vectors of weight at most W, using $O(\log n + \log(M/\epsilon)\log(M\log(1/\epsilon)))$ random bits.*

9.4.2 A generator for large sums

In this section we construct an ϵ-bit-biased distribution for coefficient vectors of weight at least $W = 10^5 M^{24}\log(1/\epsilon)$,

Recall that by Lemma 9.3, when the weight is large, the bias under the uniform distribution is small. Thus, to prove that a distribution is ϵ-bit-biased against large weight sums modulo M, it is enough to show that its bias is also small. We construct our ϵ-bit-biased generator in three steps:

- G_1: a generator that has bias at most $1 - 1/M^2$ on every coefficient vector which is not all zeros.

- G_2: a generator that has bias at most 0.91 on every coefficient vector of weight at least $100M^{24}$.

- G_3: a generator that has bias at most $\epsilon/2$ on every coefficient vector of weight at least $10^5 M^{24}\log 1/\epsilon$.

The generator G_3 will be our ϵ-bit-biased generator for large weights. We will sketch the constructions of G_1, G_2 and G_3, deferring full details and proofs to the full version of the paper. The main ingredient in the construction will be a derandomized expander product, which we now define and analyze.

Derandomized expander products

Definition 9.10. *We say an undirected graph H is a $(2^r, 2^d, \lambda)$-expander if H has 2^r vertices, it is regular of degree 2^d and all eigenvalues but the first have absolute value at most λ. We will identify the vertices of H with $\{0,1\}^r$, and the edges exiting each vertex with $\{0,1\}^d$ in some arbitrary way.*

We will need explicit constructions of expanders, which can be obtained from various known constructions.

Lemma 9.11. *For some constant $Q = 2^q$, there exist an efficient sequence H_k of $(Q^k, Q, 1/100)$-expanders.*

Impagliazzo, Nisan, and Wigderson [INW94] compose two pseudorandom generators using an expander as follows:

Definition 9.11. *Let $G', G'' : \{0,1\}^r \to \{0,1\}^t$ be two bit generators. Let H be a $(2^r, 2^d, \lambda)$-expander. We define $G' \otimes_H G'' : \{0,1\}^{r+d} \to \{0,1\}^{2t}$ to be the concatenation $(G'(x), G''(y))$, where x is a random vertex in H, and y is a random neighbor of x in H.*

Our main lemma relates the bias of $G' \otimes_H G''$ to the biases of G' and G'':

Lemma 9.12. *Let $G', G'' : \{0,1\}^r \to \{0,1\}^t$ be two bit generators and let H be a $(2^r, 2^d, \lambda)$-expander. Let $(a_1, \ldots, a_t)$,$(b_1, \ldots, b_t)$ be two coefficient vectors. Then:*

$$\begin{aligned} &|\mathrm{bias}_{(G' \otimes_H G'')(U_{r+d})}(a_1, \ldots, a_t, b_1, \ldots, b_t)| \\ &\quad \leq \quad f_\lambda(|\mathrm{bias}_{G'(U_r)}(a_1, \ldots, a_t)|, |\mathrm{bias}_{G''(U_r)}(b_1, \ldots, b_t)|) \end{aligned}$$

where $f_\lambda(x,y) = xy + \lambda\sqrt{1-x^2}\sqrt{1-y^2}$.

The bounds of [RV05] imply that if $\max_{k \in \mathbb{Z}_M \setminus 0} |\mathrm{bias}_{G'(U_r)}(ka_1, \ldots, ka_t)| \leq x$ then $\max_{k \in \mathbb{Z}_M \setminus 0} |\mathrm{bias}_{(G' \otimes_H G')(U_{r+d})}(a_1, \ldots, a_t, a_1, \ldots, a_t)| \leq x^2 + \lambda \cdot (1 - x^2) = f_\lambda(x,x)$. If also $\max_{k \in \mathbb{Z}_M \setminus 0} |\mathrm{bias}_{G''(U_r)}(kb_1, \ldots, kb_t)| \leq y$, then [RV05] proof can be extended to show $\max_{k \in \mathbb{Z}_M \setminus 0} |\mathrm{bias}_{(G' \otimes_H G')(U_{r+d})}(ka_1, \ldots, ka_t, kb_1, \ldots, kb_t)| \leq xy + \lambda \cdot (1 - xy)$, which is a worse than our bound $f(x,y)$ in case $x \neq y$ and does not suffice for our purposes. In addition, our result only requires a bound on the bias for the specific coefficient vectors $(a_1, \ldots, a_t), (b_1, \ldots, b_t)$ of interest, and not multiples of those coefficient vectors; this is crucial for our analysis when M is composite (cf., discussion after Lemma 9.3). On the other hand, the results of [RV05] are more general in that they apply to generators G',G'' that correspond to random walks on any expander, not just Cayley graphs of $\mathbb{Z}_M$.

Construction of G_1

As in [INW94, RV05], we iterate the above product. Like [RV05] we can use the constant-degree expander graphs $H_1, H_2, \ldots$ of Lemma 9.11 (as opposed to the expanders of degree $\mathrm{poly}(nw/\epsilon)$ used by [INW94] to prove Lemma 9.8). We define $G'_\ell : \{0,1\}^{\ell q} \to \{0,1\}^{2^{\ell-1}q}$ iteratively. $G'_1 : \{0,1\}^q \to \{0,1\}^q$ is the identity mapping, and $G'_\ell = G'_{\ell-1} \otimes_{H_{\ell-1}} G'_{\ell-1}$. We set $G_1 = G'_\ell$ for the minimal ℓ such that $2^{\ell-1}q \geq n$. We have:

Lemma 9.13. *Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector, which is not all zeros. Then:*

$$\mathrm{bias}_{G_1}(a_1, \ldots, a_n) \leq 1 - \frac{1}{M^2}.$$

The seed-length of G_1 is $O(\log n)$.

Construction of G_2

We will construct G_2 based on G_1. Let $(a_1, \ldots, a_n)$ be a coefficient vector. Assume first a special case: Let $n = k2^s$, and partition the set of coefficients into 2^s consecutive parts, each of size k. Assume that each part contain at least one non-zero coefficient. By Lemma 9.13, applying G_1 to each part independently gives bias of at most $1 - 1/M^2$. We use this to analyze the bias of G_1 when applied in the special case:

Lemma 9.14. *Let $n = k2^s$. Let $a_1, \ldots, a_n$ be a coefficient vector such that for every $j \in [2^s]$,* $\text{weight}(a_{jk+1}, a_{jk+2}, \ldots, a_{(j+1)k}) > 0$. *Then:*

$$\text{bias}_{G_1}(a_1, \ldots, a_n) \leq \min\left(1 - \left(\frac{9}{8}\right)^s \frac{1}{M^2}, 0.9\right).$$

In particular if $s \geq 12 \log M$, we have $\text{bias}_{G_1}(a_1, \ldots, a_n) \leq 0.9$.

We now construct the generator G_2 in three steps:

- Obliviously partition the coefficients, using the hash function H_2. Re-order the coefficients according to the partition. This guarantees that with probability at least 0.99, the conditions of Lemma 9.14 hold.
- Use G_1 on the re-ordered coefficients.
- Return the pseudorandom bits back to the original order.

We have:

Lemma 9.15. *Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector, of weight at least $100M^{24}$. Then:*

$$\text{bias}_{G_2}(a_1, \ldots, a_n) \leq 0.91.$$

The seed length of G_2 is $O(\log n + \log^2 M)$.

Construction of G_3

We use G_2 to build our final ϵ-bit-biased generator G_3. The construction of G_3 has three parts:

- Use H_3 to partition the inputs to $O(\log(1/\epsilon))$ buckets, such that with probability $1 - \epsilon/100$, most buckets contain at least $100M^{24}$ non-zero coefficients.
- Use G_2 on each bucket.
- Combine the generators for the separate buckets using expander products, with expanders of growing degree as in [RV05].

Lemma 9.16. *Let $(a_1, \ldots, a_n) \in \mathbb{Z}_M^n$ be a coefficient vector, of weight at least $10^5 M^{24} \log(1/\epsilon)$. Then:*

$$\text{bias}_{G_3}(a_1, \ldots, a_n) \leq \epsilon/2.$$

The randomness required by G_3 is $O(\log n + \log(M/\epsilon) \log(M \log(1/\epsilon)))$.

Chapter 10

Pseudorandom bit-generators for low-degree polynomials

In this work we give the first construction of a pseudorandom generator, with seed length $O(\log n)$, for $\mathrm{CC}_0[p]$, the class of constant-depth circuits with unbounded fan-in MOD_p gates, for some prime p. More accurately, the seed length of our generator is $O(\log n)$ for any constant error $\epsilon > 0$. In fact, we obtain our generator by fooling distributions generated by low degree polynomials, over $\mathbb{F}_p$, *when evaluated on the Boolean cube*. This result significantly extends previous constructions that either required a long seed [LVW93] or that could only fool the distribution generated by linear functions over $\mathbb{F}_p$, when evaluated on the Boolean cube [LRTV09, MZ09].

Enroute of constructing our PRG, we prove two structural results for low degree polynomials over finite fields that can be of independent interest.

1. Let f be an n-variate degree d polynomial over $\mathbb{F}_p$. Then, for every $\epsilon > 0$ there exists a subset $S \subset [n]$, whose size depends only on d and ϵ, such that $\sum_{\alpha \in \mathbb{F}_p^n : \alpha \neq 0, \alpha_S = 0} |\hat{f}(\alpha)|^2 \leq \epsilon$. Namely, there is a constant size subset S such that the total weight of the nonzero Fourier coefficients that do not involve any variable from S is small.

2. Let f be an n-variate degree d polynomial over $\mathbb{F}_p$. If the distribution of f when applied to uniform zero-one bits is ϵ-far (in statistical distance) from its distribution when applied to biased bits, then for every $\delta > 0$, f can be approximated over zero-one bits, up to error δ, by a function of a small number (depending only on ϵ, δ and d) of lower degree polynomials.

Joint work with Partha Mukhopadhyay and Amir Shpilka.

10.1 Introduction

A *pseudorandom generator* (PRG for short), over a domain D,[1] for a family of *tests* $\mathcal{T}$ is an explicit function $G : D^r \to D^n$ such that no test $T \in \mathcal{T}$ can distinguish a random output of

[1] One should think of D as either the Boolean cube $\{0,1\}^n$ or as $\mathbb{F}_p^n$.

G from truly uniform input elements in D^n. Namely,

$$\max_{T \in \mathcal{T}} \left| \Pr_{x \in D^r}[T(G(x)) = 0] - \Pr_{x \in D^n}[T(x) = 0] \right| \leq \epsilon \ .$$

Ideally, one would like to have the *seed* r as short as possible and the error ϵ to be as small as possible. A pseudorandom generator is considered efficient if the seed length is $O(\log n)$ (as in this case, for some applications, one can enumerate over all seeds to find a 'good' one). Pseudorandom generators have been a major object of study in theoretical computer science for several decades, and have found applications in the area of computational complexity, cryptography, algorithms design and more (see [Gol08, AB09]).

A family of tests that was widely considered in the literature is low degree polynomials over finite fields. Before stating the formal definition of a PRG for low degree polynomials we fix some notation: let f be a function, and $\mathcal{D}$ a distribution over the inputs of f. We denote by $f(\mathcal{D})$ the output distribution of f given inputs sampled according to $\mathcal{D}$. For a set S we denote by $f(S)$ the output distribution given that the inputs are uniformly sampled in S (for example, $f(\{0,1\}^n)$ is the distribution of f over uniform input bits).

Definition 10.1 (Pseudorandom distributions for degree d polynomials)**.** *A distribution $\mathcal{D}$ taking values in $\mathbb{F}_p^n$ is* pseudorandom for degree d polynomials over $\mathbb{F}_p$ *with error ϵ if, for any degree d polynomial $f(x_1, \ldots, x_n)$ over $\mathbb{F}_p$, the distributions $f(\mathcal{D})$ and $f(\mathbb{F}_p^n)$ are ϵ-close in statistical distance. A function $G : \{0,1\}^r \to \mathbb{F}_p^n$ is a* pseudorandom generator for degree d polynomials over $\mathbb{F}_p$*, if the output distribution of G, given uniformly sampled seeds, is a pseudorandom distribution for degree d polynomials.*

PRGs for linear polynomials over $\mathbb{F}_2$ were first constructed in [NN93] who gave PRGs with $O(\log n)$ seed length. The distributions constructed in [NN93] are also known as ϵ-biased distributions. Alon et al. extended this construction to work over arbitrary finite fields [AGHP90]. In [LVW93] a pseudorandom generator for the class of bounded degree polynomials over finite fields was given.[2] The seed length of [LVW93] was not optimal and was later improved in a sequence of works [BV07, Lov08, Vio08]. Note that all these generators take as input vectors from $\mathbb{F}_p^r$ and output vectors in $\mathbb{F}_p^n$. In [LRTV09, MZ09] a different kind of PRGs for linear polynomials were obtained. Both works constructed a PRG $G : \{0,1\}^r \to \{0,1\}^n$ that fools distributions generated by linear polynomials over $\mathbb{F}_p$, when evaluated on $\{0,1\}^n$. Namely, if $f = \sum_{i=1}^n \alpha_i x_i$ is a linear polynomial over $\mathbb{F}_p$ then the two distributions $f(G(\{0,1\}^r))$ and $f(\{0,1\}^n)$ are close to each other. Thus, although f is a polynomial over $\mathbb{F}_p$ they restrict their attention to the behavior of f on Boolean inputs. We call such a generator a *bit-pseudorandom generator*. We shall later give a more formal definition of bit-PRGs.

Another family of tests that received a lot of attention is bounded depth circuits (i.e. AC_0 circuits). This is the class of constant-depth circuits with unbounded fan-in AND, OR and NOT gates. AC_0 is probably the most intensively studied amongst classes of small-depth circuits. Håstad [Hås86] showed that the PARITY function cannot be approximated by any

[2] This is not explicitly stated in [LVW93], but it follows from their result for depth 2 circuits with a symmetric function at the top.

polynomial size AC_0 circuit. I.e., that no polynomial size AC_0 circuit agrees with parity on more than $\frac{1}{2} + \exp(-n)$ fraction of inputs. In other words, the *correlation* of PARITY with AC_0 is exponentially small. This result was later used by Nisan [Nis91] for constructing efficient pseudorandom generators for AC_0 (these pseudorandom generators use $r = \text{polylog(n)}$ bits). Recently, following a breakthrough by Bazzi [Baz07], Braverman [Bra09] showed that any polylog-wise independent distribution is pseudorandom for AC_0 circuits, thus settling a conjecture of Linial and Nisan [LN90]. $AC_0[p]$ is another well studied class of circuits, consisting of all constant-depth circuits with unbounded fan-in AND, OR, NOT and MOD_p gates (a MOD_p gate outputs 1 if the sum of its inputs is divisible by p, and 0 otherwise). In contrast to the impressive success in constructing pseudorandom generators for AC_0, no PRGs are known for $AC_0[p]$. One reason is that no strong correlation lower bounds are known for this class. Razborov and Smolensky [Raz87, Smo87] proved exponential lower bounds for $AC_0[p]$ circuits and their results also imply correlation lower bounds, albeit those are much weaker than the ones known for AC_0. Namely, [Raz87, Smo87] showed that the MOD_q function has polynomially small correlation with $AC_0[p]$ when p and q are co-prime. The class of $AC_0[m]$ where m is not a prime power is only very weakly understood; in particular, currently we cannot separate it from NP!

10.1.1 Our results

Motivated by the problem of constructing pseudorandom generators for $AC_0[p]$, we study a natural subclass - $CC_0[p]$ circuits. The class $CC_0[p]$ is the class of constant depth circuits using only MOD_p gates. While exponential lower bounds for this class follow from the work of Smolensky [Smo87], no pseudorandom generator better than the one constructed in [LVW93] (whose seed length is $r = \exp(\sqrt{\log n})$) is known for it. Our main result is an explicit pseudorandom generator fooling any $CC_0[p]$ circuit while using only $r = O(\log n)$ random bits, for any fixed error $\epsilon > 0$. Actually, our construction gives pseudorandom generators for low-degree polynomials over finite fields, from which the result for $CC_0[p]$ follows: Let $\mathbb{F}_p$ be a prime finite field. The MOD_p function can be computed by a degree $p-1$ polynomial over $\mathbb{F}_p$

$$\text{MOD}_p(x_1, \ldots, x_n) = (x_1 + \ldots + x_n)^{p-1} \pmod{p} .$$

Hence, any depth k circuit in $CC_0[p]$ can be computed by a polynomial over $\mathbb{F}_p$ of degree $d = (p-1)k$. Thus, in order to fool $CC_0[p]$ we have to fool the distribution induced by low degree polynomials over $\mathbb{F}_p$, when evaluated on inputs from the Boolean cube. In other words, we have to generalize the aforementioned results of [LRTV09, MZ09] from linear polynomials to any constant degree polynomials. This motivates the following definition of bit-pseudorandom generators for polynomials.

Definition 10.2 (Bit-pseudorandom distributions for degree d polynomials)**.** *A distribution $\mathcal{D}$ taking values in $\{0,1\}^n$ is* bit-pseudorandom for degree d polynomials over $\mathbb{F}_p$ *with error ϵ if, for any degree d polynomial $f(x_1, \ldots, x_n)$ over $\mathbb{F}_p$, the distributions $f(\mathcal{D})$ and $f(\{0,1\}^n)$ are ϵ-close in statistical distance. A function $G : \{0,1\}^r \to \{0,1\}^n$ is a* bit-pseudorandom generator for degree d polynomials over $\mathbb{F}_p$ *if the output distribution of G over a uniform seed is a bit-pseudorandom distribution for degree d polynomials.*

Notice the difference between this definition and Definition 10.1 where one has to fool the distribution of the polynomial when evaluated over the entire space and not just over the Boolean cube. As mentioned above, PRGs for polynomials over small finite fields were studied in several works [LVW93, BV07, Lov08, Vio08]. The best result to date is by Viola.

Theorem 10.1 (Theorem 1 in [Vio08]). *There exists an explicit and efficient function $G : \{0,1\}^r \to \mathbb{F}_p^n$ for $r = O(d \cdot \log(pn) + 2^d \cdot \log(1/\epsilon))$ such that $G(\{0,1\}^r)$ is pseudorandom for degree d polynomials over $\mathbb{F}_p$ with error ϵ.*

The problem of construction bit-pseudorandom generators for linear polynomials (i.e. the case of $d = 1$) was first studied by [LRTV09, MZ09] in the context of small-space computations. Before describing their generator we need a few notations. For $a = (a_1, \ldots, a_n) \in \mathbb{F}_p^n$ define $a^{p-1} = (a_1^{p-1}, \ldots, a_n^{p-1}) \in \{0,1\}^n$ to be the $p-1$ power of a. Similarly for a distribution $\mathcal{D} \subset \mathbb{F}_p^n$, define $\mathcal{D}^{p-1} \subset \{0,1\}^n$ by raising each element of $\mathcal{D}$ to the $p-1$ power. [LRTV09, MZ09] discovered the following construction for a bit-pseudorandom generator for linear polynomials over $\mathbb{F}_p$: the bitwise-XOR of the $p-1$ power of a pseudorandom distribution for degree $(p-1)$ polynomial over $\mathbb{F}_p$, and a k-wise independent distribution.

Theorem 10.2 (Bit-pseudorandom distribution for linear polynomials [LRTV09, MZ09]). *Let $\mathbb{F}_p$ be a prime finite field and $\epsilon > 0$ be an error parameter. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $p-1$ polynomials over $\mathbb{F}_p$ with error ϵ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution for $k = O(p^3 \log 1/\epsilon)$. Then $\mathcal{D}^{p-1} \oplus K$ is bit-pseudorandom distribution for linear polynomials over $\mathbb{F}_p$ with error $O(\epsilon)$.*

Our main result extends Theorem 10.2 to any constant degree polynomial. We prove that the following is a bit-pseudorandom distribution for degree d polynomials over $\mathbb{F}_p$: the bitwise-XOR of the $p-1$ power of a pseudorandom distribution for degree $((p-1)d)$ polynomials over $\mathbb{F}_p$, and a k-wise independent distribution.

Theorem 10.3 (Main Theorem: Bit-pseudorandom distribution). *Let $\mathbb{F}_p$ be an odd prime finite field, $d \geq 1$ an integer and $\epsilon > 0$ an error parameter. Then there exist $\delta = \delta(p,d,\epsilon)$ and $k = k(p,d,\epsilon)$ such that the following holds. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $((p-1)d)$ polynomials with error δ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution. Then, the bitwise-XOR of the two distributions $\mathcal{D}^{p-1} \oplus K$ is a bit-pseudorandom distribution for degree d polynomials over $\mathbb{F}_p$ with error ϵ. The parameters k, δ satisfy*

$$k(p,d,\epsilon), \delta(p,d,\epsilon)^{-1} \leq \exp^{(2d+1)}(\epsilon^{-c_{p,d}})$$

where $\exp^{(t)}$ is the t-times iterated exponential function, and $c_{p,d} > 0$ is some constant which depends on p and d.

An immediate corollary is that there exists an efficient and explicit pseudorandom generator $G : \{0,1\}^r \to \{0,1\}^n$ fooling any depth-k circuit in $\mathrm{CC}_0[p]$ with error ϵ, where $r = c_{p,k,\epsilon} \cdot \log n$.

Corollary 10.1 (Pseudorandom generators for $\mathrm{CC}_0[p]$). *Let p be an odd prime number and $\epsilon > 0$ an error parameter. For any $k > 0$ there exists an explicit pseudorandom generator $G : \{0,1\}^r \to \{0,1\}^n$, where $r = c_{p,k,\epsilon} \cdot \log n$, such that for any depth k circuit $C \in \mathrm{CC}_0[p]$, the statistical distance between the two distributions $C(\{0,1\}^n)$ and $C(G(\{0,1\}^r))$ is at most ϵ.*

Our proof of Theorem 10.3 is based on two new structural results for low degree polynomials, over finite fields, which may be of independent interest:

The first result is on the Fourier spectrum of such polynomials. Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a function. The α-Fourier coefficient of f, for $\alpha \in \mathbb{F}_p^n$, is defined as

$$\widehat{f}(\alpha) = \mathbb{E}_{x \in \mathbb{F}_p^n} \left[\omega^{f(x) - \langle x, \alpha \rangle} \right],$$

where $\omega = e^{2\pi i/p}$ is a primitive p-root of unity, and $\langle x, \alpha \rangle = \sum_{i=1}^n x_i \alpha_i$ is the inner product of x and α. The structural result we prove is that the Fourier coefficients of any low-degree polynomial cannot be spread over many disjoint sets. In other words, we show that one can always find a small set $S \subset [n]$ such that almost all Fourier coefficients intersect S (that is, have some nonzero entry inside S). We note that while Theorem 10.3 is interesting only for odd p,[3] this structural result is non-trivial also for polynomials over $\mathbb{F}_2$.

Theorem 10.4 (The Fourier spectrum of low-degree polynomials over finite fields)**.** *For every prime finite field $\mathbb{F}_p$, degree $d \geq 1$ and error $\epsilon > 0$ there exists a constant $C(d, \epsilon) \leq (1/\epsilon)^{O(4^d)}$ such that the following holds. Let $f(x_1, \ldots, x_n)$ be a degree d polynomial over $\mathbb{F}_p$. Then there exists a subset $S \subset [n]$ of size at most $|S| \leq C(d, \epsilon)$ such that*

$$\sum_{\alpha \in \mathbb{F}_p^n : \alpha \neq 0, \alpha_S = 0} |\hat{f}(\alpha)|^2 \leq \epsilon \,,$$

where α_S is the restriction of α to coordinates in S. In words, almost all nonzero Fourier coefficients of f intersect S.

Our second structural result concerns the structure of polynomials with the following property. Denote with $\mathcal{U}_p$ the distribution over $\{0,1\}^n$ where each bit is chosen independently to be 0 with probability $1/p$ and 1 with probability $1 - 1/p$. We call $\mathcal{U}_p$ the *p-biased distribution*. We show that if the distributions $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ are ϵ-far, then f can be approximated, over $\{0,1\}^n$, by a function of a small number of lower degree polynomials. To formally state our theorem we need some definitions.

Definition 10.3 (Bit-Rank)**.** *Let $g : \{0,1\}^n \to \mathbb{F}_p$ be a function. The d-bit-rank of g, denoted* $\mathrm{bitrank}_d(g)$*, is the minimal number of degree d polynomials over $\mathbb{F}_p$ required to compute g over $\{0,1\}^n$. That is,* $\mathrm{rank}_d(g) = k$ *where k is the minimal number such that there exist k degree d polynomials $f_1, \ldots, f_k : \mathbb{F}_p^n \to \mathbb{F}_p$ and a function $\Gamma : \mathbb{F}_p^k \to \mathbb{F}_p$ such that for all $x \in \{0,1\}^n$*

$$g(x) = \Gamma(f_1(x), \ldots, f_k(x)).$$

Example 10.1. *Consider the function $g(x) = \sum_{i \neq j} x_i x_j$ over $\mathbb{F}_p$ for $p > 2$. We have that the 1-bit-rank of g is 1, as for all $x \in \{0,1\}^n$*

$$g(x) = (x_1 + \ldots + x_n)^2 - (x_1^2 + \ldots + x_n^2) = (x_1 + \ldots + x_n)^2 - (x_1 + \ldots + x_n) \,.$$

Thus, for $x \in \{0,1\}^n$, $g(x)$ is determined by the linear function $\ell(x) = x_1 + \ldots + x_n$. Notice that as a quadratic polynomial over $\mathbb{F}_p$, the rank of g (i.e. the minimal number of linear functions required to compute g on inputs from $\mathbb{F}_p^n$) is either $n - 1$ or n, depending on p.

[3] For $p = 2$ it reduces to the case of pseudorandom distributions.

Our second structural result is the following.

Theorem 10.5 (Structure of bit-biased polynomials)**.** *Let $f(x_1, \ldots, x_n)$ be a degree d polynomial over $\mathbb{F}_p$ such that the statistical distance between the distributions $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ is at least ϵ. Then, for every $\delta > 0$, there exists a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that $\Pr_{x \in \{0,1\}^n}[g(x) \neq f(x)] \leq \delta$ and $\mathrm{bitrank}_d(g) \leq p^{O(c)}$ where[4] $c = C((p-1)(d+1), \delta\epsilon^2/p^3)$.*

In fact, for our proof we require such a polynomial g that approximates f with respect to (an affine shift of) $\mathcal{U}_p$, but we find this statement more appealing.

10.1.2 Proof overview

Pseudorandom generators that fool low degree polynomials over $\mathbb{F}_p^n$ were obtained in [BV07, Lov08, Vio08]. In our case we only consider the distribution of the polynomial over $\{0,1\}^n$ (and not over $\mathbb{F}_p^n$ as the aforementioned results), which creates new obstacles, and requires a different approach.

We sketch below the proof of Theorem 10.3. Our proof is carried by induction on the degree d, and uses Theorem 10.4 and (a variant of) Theorem 10.5 as important technical ingredients. Let $f(x) = f(x_1, \ldots, x_n)$ be a polynomial of degree d over $\mathbb{F}_p$. The base case of $d = 1$ was established in [LRTV09, MZ09], hence we assume from now on $d \geq 2$.

Regular polynomials Consider the p-biased distribution $\mathcal{U}_p$. This distribution can be simulated by low-degree polynomials over $\mathbb{F}_p$: let $x \in \mathbb{F}_p^n$ be chosen uniformly at random; then, $x^{p-1} = (x_1^{p-1}, \ldots, x_n^{p-1})$ is distributed according to $\mathcal{U}_p$. Furthermore, it is easy to construct a pseudorandom distribution fooling $f(\mathcal{U}_p)$ as follows. Let $\widetilde{f}(x) = f(x^{p-1})$. Then $\widetilde{f}$ is a polynomial of degree $(p-1)d$, and the distributions $\widetilde{f}(\mathbb{F}_p^n)$ and $f(\mathcal{U}_p)$ are identical. In particular, any distribution fooling degree $(p-1)d$ polynomials over $\mathbb{F}_p$ (such as those guaranteed by Theorem 10.1) also fools $f(\mathcal{U}_p)$.

Thus, if the polynomial f is regular in the sense that it cannot distinguish between the uniform distribution over $\{0,1\}^n$ and the p-biased distribution $\mathcal{U}_p$, then one can simply use a pseudorandom generator for $\widetilde{f}$ to get a pseudorandom generator for f. Hence, it is not hard to deduce the following lemma.

Lemma (Lemma 10.1, informal statement)**.** *Let $f(x)$ be a degree d polynomial over $\mathbb{F}_p$ such that the distributions $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ are ϵ-close. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $((p-1)d)$ polynomials over $\mathbb{F}_p$ with error ϵ. Then $f(\mathcal{D}^{p-1})$ and $f(\{0,1\}^n)$ are $O(\epsilon)$-close.*

Non-regular polynomials We now have to deal with non-regular polynomials, i.e polynomials that distinguish between uniform bits and the p-biased distribution. This is the main challenge we tackle in the paper. In fact, we will show that this property is so strong that bit-pseudorandom generators for degree $d-1$ polynomials with small enough error suffice to fool any such degree d polynomial. The proof consists of two steps. First we

[4]The function $C(\cdot,\cdot)$ is defined in the statement of Theorem 10.4.

prove Theorem 10.6 (which is close in spirit to Theorem 10.5) that shows that f can be well-approximated, with respect to (an affine shift of) $\mathcal{U}_p$, by a few polynomials of degree $d-1$. We then prove that any distribution that fools degree $d-1$ polynomials (over $\{0,1\}^n$) also fools f (Lemma 10.2).

We now explain the idea behind the proof of Theorem 10.6. Bogdanov and Viola proved that if $f(x)$ is a degree d polynomial over $\mathbb{F}_p$ such that $f(\mathbb{F}_p^n)$ is far from the uniform distribution over $\mathbb{F}_p$, then f can be well-approximated by a function of a few polynomials of lower degree [BV07]. Following this motivating example, we would like to prove that if $f(\mathcal{U}_p)$ is far from uniform (a similar property can be easily obtained from the fact that f is not regular, see Claim 10.1) then f can be well-approximated over $\mathcal{U}_p$ by lower degree polynomials. However, the case of $f(\mathbb{F}_p^n)$ being far from uniform is easy to handle via directional derivatives, as the input space is invariant under shifts (i.e. the mapping $x \to x+a$ for $a \in \mathbb{F}_p^n$ maps the uniform distribution over $\mathbb{F}_p^n$ to itself). In our case, the input distribution $\mathcal{U}_p$ is not invariant under shifts, which creates a major obstacle for using existing techniques.

To overcome this obstacle we first 'complete' f to a polynomial over $\mathbb{F}_p^n$ that carries similar properties: Define $f^{\oplus a} = f(x^{p-1} \oplus a)$, for some $a \in \{0,1\}^n$. Then $f^{\oplus a}$ is a polynomial of degree $d' = (p-1)d$ and the distributions $f^{\oplus a}(\mathbb{F}_p^n)$ and $f(\mathcal{U}_p \oplus a)$ are identical. We show that as f is non-regular, there exists $a \in \{0,1\}^n$ such that $f^{\oplus a}$ is biased (Corollary 10.3). Similarly to [BV07] we get that $f^{\oplus a}$ can be approximated by a few of its *directional derivatives*, where the directional derivative of $f^{\oplus a}$ in direction $y \in \mathbb{F}_p^n$ is defined as $f_y^{\oplus a}(x) = f^{\oplus a}(x+y) - f^{\oplus a}(x)$. However, in our case we need a stronger property to hold. Define the *support* of y to be the set of nonzero entries in y, $\mathrm{Supp}(y) = \{i \in [n] : y_i \neq 0\}$. We would like to show that $f^{\oplus a}$ can be approximated by a few directional derivatives having small supports. To obtain this we need Theorem 10.4 that shows that most of the Fourier weight of $f^{\oplus a}$ is supported on coefficients that intersect a relatively small set S. Using this theorem we get

Claim (Claim 10.8, informal statement)**.** *Let $\widetilde{f}$ be a polynomial over $\mathbb{F}_p$ of degree d'. For every $\delta > 0$ there exist a small number of directions $y_1, \ldots, y_k \in \mathbb{F}_p^n$ such that $|\mathrm{Supp}(y_1) \cup \ldots \cup \mathrm{Supp}(y_k)|$ is small, and such that $\widetilde{f}$ can be well-approximated by some function Γ of $\widetilde{f}_{y_1}, \ldots, \widetilde{f}_{y_k}$. Namely,*

$$\Pr_{x \in \mathbb{F}_p^n}[\widetilde{f}(x) \neq \Gamma(\widetilde{f}_{y_1}(x), \ldots, \widetilde{f}_{y_k}(x))] \leq \delta.$$

This is still not enough as the derivatives of $f^{\oplus a}$ have degree $(p-1)d-1$. However, we further show that sparse directional derivatives of $f^{\oplus a}$ can be calculated by directional derivatives of f and a few variables.

Claim (Claim 10.9, informal statement)**.** *Any directional derivative $f_y^{\oplus a}(x)$ can be computed by some function of $f_y(x)$ and $\{x_i : i \in \mathrm{Supp}(y)\}$.*

We prove this claim by showing that any derivatives of $f^{\oplus a}$, with respect to a direction supported on S, satisfies $(f^{\oplus a})_y(x) = f_w(x^{p-1} \oplus a)$ for some w that depends only on y and a, and is supported on S. Combining Claims 10.8 and 10.9 yields the required approximation of f.

To complete the picture we shortly remark on the proof of Theorem 10.4. The proof is by induction on the degree using Fourier analysis. The basic idea is that for every linear

subspace $A \subseteq \mathbb{F}_p^n$ we have that

$$\sum_{\alpha \in \mathbb{F}_p^n : \alpha \neq 0, \alpha_S = 0} |\hat{f}(\alpha)|^2 \leq \mathbb{E}_{a \in A} \left[\sum_{\alpha \in \mathbb{F}_p^n : \alpha \neq 0, \alpha_S = 0} |\widehat{f_a}(\alpha)|^2 \right] + \mathbb{E}_{a \in A} \left[|\widehat{f_a}(0)|^2 \right] .$$

Using this useful inequality we break the analysis to two cases depending on whether f has a high Fourier coefficient or not. If all of f's Fourier coefficients are small, then we construct S in the following way: we pick a constant dimensional subspace A at random. For each derivative f_a, where $a \in A$, we find a set S_a as guaranteed by the induction hypothesis (for some ϵ' depending on ϵ and d). Finally, we set S to be the union of all the S_a-s. When f has a high Fourier coefficient, we approximate f using a small number of lower degree polynomials and set S to be the union of their corresponding sets.

10.1.3 Paper organization

In Section 10.2 we fix some notations. We prove our main theorem, Theorem 10.3, in Section 10.3. The proof is based on Theorem 10.6 whose proof is given in Section 10.5, where we also prove Theorem 10.5. The proof of Theorem 10.6 relies on Theorem 10.4 that we prove in Section 10.6. We conclude and give some open problems in Section 10.7. For completeness, we sketch the proof for the linear case of Theorem 10.3 (i.e. $d = 1$) in Section 10.8.

10.2 Preliminaries

We will be working over a fixed prime finite field $\mathbb{F}_p$. Let $f(x) = f(x_1, \ldots, x_n)$ be a degree d polynomial over $\mathbb{F}_p$. Let $\mathcal{D}$ be a distribution. The support of $\mathcal{D}$ is the set of elements which have positive probability under $\mathcal{D}$. If the support of $\mathcal{D}$ is contained in a set S, we denote this by $\mathcal{D} \subseteq S$. For a distribution $\mathcal{D} \subset \mathbb{F}_p^n$, we denote by $f(\mathcal{D})$ the output distribution of f given inputs samples according to $\mathcal{D}$. For a subset $S \subset \mathbb{F}_p^n$ we denote by $f(S)$ the distribution of f over inputs chosen uniformly from S. In particular, $f(\mathbb{F}_p^n)$ denotes the distribution of f over uniform field elements, and $f(\{0,1\}^n)$ denotes the distribution of f over uniform bits.

The statistical distance between two distributions $\mathcal{D}', \mathcal{D}''$ is given by $\mathrm{sd}(\mathcal{D}', \mathcal{D}'') = \frac{1}{2} \sum_x |\Pr[\mathcal{D}' = x] - \Pr[\mathcal{D}'' = x]|$. If the statistical distance is at most ϵ, the distributions are said to be *ϵ-close*. If the statistical distance is at least ϵ, the distributions are said to be *ϵ-far*. It is easy to verify that statistical distance satisfies the triangle inequality.

Denote $[n] = \{1, 2, \ldots, n\}$. A distribution $K \subset \{0,1\}^n$ is said to be *k-wise independent* if for any k distinct indices $i_1, \ldots, i_k \in [n]$, the distribution K restricted to these indices is uniform over $\{0,1\}^k$.

For a function $f : \mathbb{F}_p^n \to \mathbb{F}_p$ we denote by $\widehat{f} : \mathbb{F}_p^n \to \mathbb{C}$ its Fourier transform, defined as $\widehat{f}(\alpha) = \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega^{f(x) - \langle x, \alpha \rangle}]$, where $\omega = e^{2\pi i/p}$ and $\langle x, \alpha \rangle = \sum_{i=1}^n \alpha_i x_i$ is the inner product of x and α. The Fourier representation of $f(x)$ is given by $f(x) = \sum_{\alpha \in \mathbb{F}_p^n} \widehat{f}(\alpha) \omega^{\langle x, \alpha \rangle}$. Parseval's identity gives that $\sum_{\alpha \in \mathbb{F}_p^n} |\widehat{f}(\alpha)|^2 = 1$.

10.3 Bit pseudorandom generator for low degree polynomials

In this section we prove Theorem 10.3. As sketched in Section 10.1.2 we first prove the theorem for the (easy) case of *regular* polynomials (a notion that we shall soon define) and then for non-regular polynomials.

10.3.1 Regular polynomials

Definition 10.4. *The* p-biased distribution $\mathcal{U}_p \subset \{0,1\}^n$ *is the distribution in which we choose each bit independently to be* 0 *with probability* $\frac{1}{p}$ *and to be* 1 *with probability* $1 - \frac{1}{p}$.

We call a polynomial $f : \mathbb{F}_p^n \to \mathbb{F}_p$ ϵ-regular if $\mathrm{sd}(f(\mathcal{U}_p), f(\{0,1\}^n)) \leq \epsilon$. The following lemma shows that if f is a regular polynomial then it is fooled by the $p-1$ power of a pseudorandom distribution for degree $(p-1)d$ polynomials.

Lemma 10.1. *Let* $f(x_1, \ldots, x_n)$ *be an* ϵ*-regular polynomial of degree* d *over* $\mathbb{F}_p$*. Let* $\mathcal{D} \subset \mathbb{F}_p^n$ *be a pseudorandom distribution for degree* $(p-1)d$ *polynomials over* $\mathbb{F}_p$ *with error* ϵ*. Then* $\mathrm{sd}(f(\mathcal{D}^{p-1}), f(\{0,1\}^n)) \leq 2\epsilon$.

Proof. Let $\tilde{f} : \mathbb{F}_p^n \to \mathbb{F}_p$ be defined as $\tilde{f}(x_1, \ldots, x_n) = f(x_1^{p-1}, \ldots, x_n^{p-1})$. As f is a degree d polynomial, $\tilde{f}$ is a polynomial of degree $(p-1)d$. Since $\mathcal{D}$ is pseudorandom against polynomials of degree $(p-1)d$, we have that $\tilde{f}(\mathcal{D})$ and $\tilde{f}(\mathbb{F}_p^n)$ are ϵ-close. By the definition of $\tilde{f}$ it follows that $f(\mathcal{D}^{p-1})$ and $f(\mathcal{U}_p)$ are ϵ-close. Hence, $\mathrm{sd}(f(\{0,1\}^n), f(\mathcal{D}^{p-1})) \leq \mathrm{sd}(f(\{0,1\}^n), f(\mathcal{U}_p)) + \mathrm{sd}(f(\mathcal{U}_p), f(\mathcal{D}^{p-1})) \leq 2\epsilon$. □

10.3.2 Non-regular polynomials

We now turn to study non regular polynomials. Namely, polynomials that can distinguish between the uniform distribution over $\{0,1\}^n$ and the p-biased distribution. The main tool in the proof is (a variant of) Theorem 10.5 that shows that non regular polynomials possess a very special structure. Namely, that a non-regular polynomial can be well approximated by a function of a small number of lower degree polynomials.

We will start by proving that non-regular polynomials admit a non-uniform distribution when applied to inputs sampled from some shift of a p-biased distribution. For a distribution $\mathcal{D} \subset \{0,1\}^n$ and an element $a \in \{0,1\}^n$ denote by $\mathcal{D} \oplus a$ the distribution generated by bitwise-XORing the element a to all elements of $\mathcal{D}$.

Claim 10.1. *Let* $f(x_1, \ldots, x_n)$ *be a degree* d *polynomial over* $\mathbb{F}_p$ *such that the distributions* $f(\mathcal{U}_p)$ *and* $f(\{0,1\}^n)$ *are* ϵ*-far. Then there exists* $a \in \{0,1\}^n$ *such that the distribution* $f(\mathcal{U}_p \oplus a)$ *is* $\epsilon/2$*-far from the uniform distribution over* $\mathbb{F}_p$.

Proof. If $f(\mathcal{U}_p)$ and $f(\{0,1\}^n)$ are ϵ-far, at least one of them is $\epsilon/2$-far from the uniform distribution over $\mathbb{F}_p$. If it is $f(\mathcal{U}_p)$, then we are done with $a = 0$. Otherwise assume that $f(\{0,1\}^n)$ is $\epsilon/2$-far from the uniform distribution over $\mathbb{F}_p$. We can generate the uniform

distribution over $\{0,1\}^n$ by first choosing $a \in \{0,1\}^n$ uniformly at random, and then bitwise-XORing it to the distribution $\mathcal{U}_p$. In other words, the uniform distribution over $\{0,1\}^n$ is a convex combination of the distributions $\{\mathcal{U}_p \oplus a : a \in \{0,1\}^n\}$. Thus, the distribution $f(\{0,1\}^n)$ is a convex combination of the distributions $\{f(\mathcal{U}_p \oplus a) : a \in \{0,1\}^n\}$. In particular, there must exist some $a \in \{0,1\}^n$ such that the distribution $f(\mathcal{U}_p \oplus a)$ is $\epsilon/2$-far from uniform. □

We recall the definition of bitrank given in Subsection 10.1.1.

Definition (Bit-Rank)**.** *Let $g : \{0,1\}^n \to \mathbb{F}_p$ be a function. The d-bit-rank of g, denoted* $\mathrm{bitrank}_d(g)$*, is the minimal number of degree d polynomials over $\mathbb{F}_p$ required to compute g over $\{0,1\}^n$. That is,* $\mathrm{rank}_d(g) = k$ *where k is the minimal number such that there exist k degree d polynomials $f_1, \ldots, f_k : \mathbb{F}_p^n \to \mathbb{F}_p$ and a function $\Gamma : \mathbb{F}_p^k \to \mathbb{F}_p$ such that for all $x \in \{0,1\}^n$*

$$g(x) = \Gamma(f_1(x), \ldots, f_k(x)).$$

The following theorem, shows that non-regular polynomials have a low bit-rank. We shall later deduce Theorem 10.5 from it. We defer the proof of the theorem to Section 10.5.

Theorem 10.6. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a polynomial of degree $d+1$ for some $d \geq 1$. Assume that, for some $a \in \{0,1\}^n$, the distribution of $f(\mathcal{U}_p \oplus a)$ is ϵ-far from uniform. Then for every $\delta > 0$ there exists a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that $\Pr_{x \in \mathcal{U}_p \oplus a}[g(x) \neq f(x)] \leq \delta$ and* $\mathrm{bitrank}_d(g) \leq c + p^c$ *where*[5] $c = C((p-1)(d+1), \delta\epsilon^2/p^3)$.

The next lemma shows that if a degree $d+1$ polynomial $f(x)$ can be approximated, under some shift of the p-biased distribution, by a function with a low d-bit-rank, then bit-pseudorandom distributions for degree d polynomials also fool f.

Lemma 10.2. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree $d+1$ polynomial. Assume that there is a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that* $\mathrm{bitrank}_d(g) = k$ *and for some $a \in \{0,1\}^n$ it holds that*

$$\Pr_{x \in \mathcal{U}_p \oplus a}[f(x) = g(x)] \geq 1 - \delta.$$

Let $\mathcal{D} \subset \{0,1\}^n$ be a bit-pseudorandom distribution for degree d polynomials with error ϵ. Then $f(\mathcal{D})$ and $f(\{0,1\}^n)$ are $(c_1^k \epsilon + c_2 \delta)$-close, for $c_1 = p^{2^{(p-1)(d+1)}}$ and $c_2 = 4p \cdot 2^{(p-1)(d+1)}$.

To ease the reading we first show how to obtain Theorem 10.3 using Theorem 10.6 and Lemma 10.2. The proof of Lemma 10.2 is given in Section 10.4 and the proof of Theorem 10.6 is given in Section 10.5.

10.3.3 Proof of Theorem 10.3

For convenience we repeat the statement of the theorem.

[5] The function $C(\cdot,\cdot)$ is defined in the statement of Theorem 10.4.

Theorem. *Let $\mathbb{F}_p$ be an odd prime finite field, $d \geq 1$ be a degree and $\epsilon > 0$ be an error parameter. Then there exist $\delta = \delta(p,d,\epsilon)$ and $k = k(p,d,\epsilon)$ such that the following holds. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $((p-1)d)$ polynomials with error δ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution. Then the bitwise-XOR of the two distributions $\mathcal{D}^{p-1} \oplus K$ is bit-pseudorandom for degree d polynomials over $\mathbb{F}_p$ with error ϵ. The parameters k, δ satisfy*

$$k(p,d,\epsilon), \delta(p,d,\epsilon)^{-1} \leq \exp^{(2d+1)}(\epsilon^{-c_{p,d}})$$

where $\exp^{(t)}$ is the t-times iterated exponential function, and $c_{p,d} > 0$ is some constant which depends on $\mathbb{F}_p$ and d.

Proof. The proof is by induction on the degree d. The case $d = 1$ was established in [LRTV09, MZ09] (Theorem 10.2). We restate their result here. For completeness we give a sketch of the proof in Section 10.8.

Theorem (Bit-pseudorandom distribution for linear polynomials)**.** *Let $\mathbb{F}_p$ be a prime finite field and $\epsilon > 0$ be an error parameter. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $p-1$ polynomials over $\mathbb{F}_p$ with error ϵ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution for $k = O(p^3 \log 1/\epsilon)$. Then $\mathcal{D}^{p-1} \oplus K$ is bit-pseudorandom distribution against linear polynomials over $\mathbb{F}_p$ with error $O(\epsilon)$.*

We now proceed with the induction. Let $d > 1$, and let $f(x)$ be a polynomial of degree $d+1$. We divide the analysis into two cases. Assume first that $f(\mathcal{U}_p)$ is $\epsilon/2$-close to $f(\{0,1\}^n)$. Lemma 10.1 implies that if $\mathcal{D}_1 \subset \mathbb{F}_p^n$ is a pseudorandom distribution for degree $(p-1)(d+1)$ polynomials, with error $\epsilon/2$, then $f(\mathcal{D}_1^{p-1})$ and $f(\{0,1\}^n)$ are ϵ-close.

We now handle the case that $f(\mathcal{U}_p)$ is $\epsilon/2$-far from $f(\{0,1\}^n)$. By Claim 10.1 there exists some $a \in \{0,1\}^n$ such that $f(\mathcal{U}_p \oplus a)$ is $(\epsilon/4)$-far from uniform. Let $\delta > 0$ be determined later. Applying Theorem 10.6 there exists a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that $\Pr_{x \in \mathcal{U}_p \oplus a}[f(x) \neq g(x)] \leq \delta$ and $\mathrm{bitrank}_d(g) \leq p^c + c$ for $c = C((p-1)(d+1), \delta(\epsilon/4)^2/p^3) = O(p\delta^{-1}\epsilon^{-1})^{O(4^{(p-1)(d+1)})}$. Lemma 10.2 implies that if $\mathcal{D}' \subset \{0,1\}^n$ is a bit-pseudorandom distribution for degree d polynomials with error ξ (that will be determined soon) then $f(\mathcal{D}')$ and $f(\{0,1\}^n)$ are τ-close for

$$\tau = c_1^{p^c+c}\xi + c_2\delta$$

where $c_1 = p^{2^{(p-1)(d+1)}}$ and $c_2 = 4p \cdot 2^{(p-1)(d+1)}$. In order to get $\tau \leq \epsilon$ we set $\delta = \epsilon/2c_2$ and $\xi = \epsilon/2c_1^{p^c+c}$. Substituting the parameters yields the bound

$$1/\xi \leq \exp(\exp((1/\epsilon)^{O(4^{(p-1)(d+1)})}) .$$

We now put things together. Let $\mathcal{D}_2 \subset \mathbb{F}_p^n$ to be a pseudorandom distribution for degree $(p-1)d$ polynomials with error $\delta = \delta(p,d,\xi)$. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution for $k = k(p,d,\xi)$. By the induction hypothesis, $\mathcal{D}' = \mathcal{D}_2^{p-1} \oplus K$ is a bit-pseudorandom distribution for degree d polynomials with error ξ. Thus, if f is not $\epsilon/2$-regular then $\mathcal{D}'$ fools f with error ϵ. To combine the two case in our analysis we note that if $\mathcal{D} \subset \mathbb{F}_p^n$ is a pseudorandom distribution against degree $(p-1)(d+1)$ polynomials with error ξ then $\mathcal{D}$ satisfies the requirements of both $\mathcal{D}_1$ and $\mathcal{D}_2$ (recall that $\xi \ll \epsilon$). Hence $\mathcal{D}^{p-1} \oplus K$

is a bit-pseudorandom distribution against any polynomial of degree $d+1$ with error ϵ. To conclude the proof we note that as

$$\delta(p, d+1, \epsilon) = \delta\left(p, d, 1/\exp(\exp((1/\epsilon)^{O(4^{(p-1)(d+1)})}))\right)$$

and

$$k(p, d+1, \epsilon) = k\left(p, d, 1/\exp(\exp((1/\epsilon)^{O(4^{(p-1)(d+1)})}))\right)$$

then there is a constants $c_{p,d} > 0$ such that

$$k(p, d, \epsilon),\ \delta(p, d, \epsilon)^{-1} \leq \exp^{(2d+1)}(\epsilon^{-c_{p,d}})$$

as claimed.

□

10.4 Approximately low bit-rank polynomials

In this section we give the proof of Lemma 10.2. We first give an overview of the proof.

Step 1. The first step in the proof is showing that if f is a degree $d+1$ polynomial which can be approximated by a function g of low d-bit-rank, then there is a *distribution* on functions G, such that every function in the support of G has a low d-bit-rank and such that for every $x \in \mathbb{F}_p^n$ it holds that $\Pr_{h \in G}[f(x) = h(x)] \geq 1 - \delta$ (Lemma 10.3). That is, we move from one function that compute f on most of the space to a distribution that is 'good' for every point x. The main idea behind the proof of this step is to use the self-correction properties of low degree polynomials (Claim 10.2). This step is the main technical part of the proof

Step 2. In the second step we show that if a function has a low d-bit-rank then any bit-pseudorandom distribution for degree d polynomials fools it. The argument here is quite straightforward (Claim 10.4).

Step 3. Finally, we show that if a function can be computed using a distribution on functions that have low d-bit-rank (as we achieved in **Step 1** above) then it is fooled by bit-pseudorandom distributions for degree d polynomials (Claim 10.5).

10.4.1 Step 1: from average case to worst case approximation

As in the overview above we start by showing that there exists a distribution on low d-bit-rank functions that correctly computes f everywhere (w.h.p.). To construct such a distribution we shall refer to the self correction properties of polynomials over $\mathbb{F}_p$. Using these properties we will show that we can construct G by (roughly) considering many shifts of g (the polynomial that computes f on a $1-\delta$ fraction of $\mathbb{F}_p^n$). We start with the following well known fact.

Claim 10.2. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree $d+1$ polynomial. For every $x, y_1, \ldots, y_{d+2} \in \mathbb{F}_p^n$ the following holds*

$$f(x) = \sum_{I \subseteq [d+2], |I| \geq 1} (-1)^{|I|+1} f(x + \sum_{i \in I} y_I) \ .$$

Proof. Taking $d+2$ partial derivatives of $f(x)$, in directions[6] $y_1, \ldots, y_{d+2}$, iteratively, we obtain the constant zero function. That is $f_{y_1,\ldots,y_{d+2}}(x) \equiv 0$. The claim follows as, by definition, $f_{y_1,\ldots,y_{d+2}}(x) = (-1)^{d+2} \sum_{I \subseteq [d+2]} (-1)^{|I|} f(x + \sum_{i \in I} y_I)$. □

The following is an easy corollary.

Corollary 10.2. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree $d+1$ polynomial. Let $t = (p-1)(d+1)+1$. For every $x, y_1, \ldots, y_t \in \mathbb{F}_p^n$ and $a \in \{0,1\}^n$ the following holds*

$$f(x^{p-1} \oplus a) = \sum_{I \subseteq [t], |I| \geq 1} (-1)^{|I|+1} f((x + \sum_{i \in I} y_I)^{p-1} \oplus a) \ .$$

Proof. Define $g^{\oplus a}(x) = f(x^{p-1} \oplus a)$. Note that $g^{\oplus a}$ is a polynomial of degree $(p-1)(d+1)$ since

$$g^{\oplus a}(x_1, \ldots, x_n) = f(\alpha_1 x_1^{p-1} + \beta_1, \ldots, \alpha_n x_n^{p-1} + \beta_n)$$

where α_i, β_i are defined as follows. If $a_i = 0$ then $\alpha_i = 1, \beta_i = 0$ and if $a_i = 1$ then $\alpha_i = -1, \beta_i = 1$. The claim is proved by applying Claim 10.2 to the polynomial $g^{\oplus a}$. □

We now show that a shift of a low bit-rank polynomial also has low bit-rank.

Claim 10.3. *Let $g : \{0,1\}^n \to \mathbb{F}_p$ be a function. For $a, c \in \{0,1\}^n, b \in \mathbb{F}_p^n$ define $g^{\oplus a,+b,\oplus c} : \{0,1\}^n \to \mathbb{F}_p$ by $g^{\oplus a,+b,\oplus c}(x) = g(((x \oplus c)+b)^{p-1} \oplus a)$. Then* $\text{bitrank}_d(g^{\oplus a,+b,\oplus c}) \leq \text{bitrank}_d(g)$.

Proof. Assume $\text{bitrank}_d(g) = k$. Consequently, there are k degree d polynomials $f_1, \ldots, f_k$ and a mapping $\Gamma : \mathbb{F}_p^k \to \mathbb{F}_p$ such that $g(x) = \Gamma(f_1(x), \ldots, f_k(x))$. Thus

$$g^{\oplus a,+b,\oplus c}(x) = g(((x \oplus c)+b)^{p-1} \oplus a) = \Gamma(f_1(((x \oplus c)+b)^{p-1} \oplus a), \ldots, f_k(((x \oplus c)+b)^{p-1} \oplus a)) \ .$$

We will conclude the proof by showing that each $f_j(((x \oplus c) + b)^{p-1} \oplus a)$ is a polynomial of degree at most d (in x). Define $f'_j(x_1, \ldots, x_n) = f_j(\alpha_1 x_1 + \beta_1, \ldots, \alpha_n x_n + \beta_n)$ where α_i, β_i are defined such that $\alpha_i x_i + \beta_i = ((x_i \oplus c_i)+b_i)^{p-1} \oplus a_i$ for $x_i \in \{0,1\}$ (that is, $\beta_i = (c_i + b_i)^{p-1} \oplus a_i$ and $\alpha_i = -\beta_i + (((1 \oplus c_i) + b_i)^{p-1} \oplus a_i)$). As we applied an affine linear transformation to the inputs $x_1, \ldots, x_n$, we have $\deg(f'_j) \leq \deg(f_j) \leq d$. We conclude that for any $x \in \{0,1\}^n$

$$g^{\oplus a,+b,\oplus c}(x) = \Gamma(f'_1(x), \ldots, f'_k(x)),$$

hence $\text{bitrank}_d(g^{\oplus a,+b,\oplus c}) \leq k$. □

Let G be a distribution over functions $g : \{0,1\}^n \to \mathbb{F}_p$. The d-bit-rank of G is defined to be the maximal d-bit-rank of a function in the support of G. The following lemma concludes the idea sketched above and shows that if f is close to a function with a low bit-rank then there is a distribution on low bit-rank functions that pointwise computes f.

[6] Recall that the derivative of f in direction y is defined as $f_y(x) = f(x+y) - f(x)$. It is easy to verify that if f has degree $d+1$ then f_y has degree at most d.

Lemma 10.3. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree $d+1$ polynomial. Assume that there is a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that* $\text{bitrank}_d(g) = k$ *and such that for some $a \in \{0,1\}^n$ it holds that*

$$\Pr_{x \in \mathcal{U}_p \oplus a}[f(x) = g(x)] \geq 1 - \delta \ .$$

Then, there is a distribution G on functions such that $\text{bitrank}_d(G) \leq (2^{(p-1)(d+1)+1} - 1)k$ *and* $\Pr_{h \in G}[f(x) = h(x)] \geq 1 - (2^{(p-1)(d+1)+1} - 1)\delta$.

Proof. We start by noting that the distribution $\mathcal{U}_p \oplus a$ is equivalent to the distribution of $x^{p-1} \oplus a$ for uniform $x \in \mathbb{F}_p^n$. By our assumption we have that

$$\Pr_{x \in \mathbb{F}_p^n}[f(x^{p-1} \oplus a) \neq g(x^{p-1} \oplus a)] \leq \delta.$$

Applying Corollary10.2 to f, which is a degree $d+1$ polynomial, we obtain

$$f(x^{p-1} \oplus a) = \sum_{I \subseteq [t], |I| \geq 1} (-1)^{|I|+1} f((x + \sum_{i \in I} y_I)^{p-1} \oplus a) \tag{10.1}$$

for $t = (p-1)(d+1)+1$ and any $y_1, \ldots, y_t \in \mathbb{F}_p^n$. Fix some $x \in \{0,1\}^n$. Let $y_1, \ldots, y_t \in \mathbb{F}_p^n$ be chosen uniformly at random, and note that for any non-empty $I \subseteq [t]$, the distribution of $x + \sum_{i \in I} y_i$ is uniform over $\mathbb{F}_p^n$. Therefore, for every $I \neq \emptyset$ it holds that

$$\Pr_{y_1, \ldots, y_t \in \mathbb{F}_p^n} \left[f((x + \sum_{i \in I} y_i)^{p-1} \oplus a) \neq g((x + \sum_{i \in I} y_i)^{p-1} \oplus a) \right] \leq \delta \ .$$

As $x^{p-1} = x$ for $x \in \{0,1\}^n$ we have that for such x-s $f(x^{p-1} \oplus a) = f(x \oplus a)$. Therefore, by Equation (10.1) and the union bound we get

$$\Pr_{y_1, \ldots, y_t \in \mathbb{F}_p^n} \left[f(x \oplus a) \neq \sum_{I \subseteq [t], |I| \geq 1} (-1)^{|I|+1} g((x + \sum_{i \in I} y_i)^{p-1} \oplus a) \right] \leq (2^t - 1)\delta \ .$$

Hence, for every $x \in \{0,1\}^n$ we have

$$\Pr_{y_1, \ldots, y_t \in \mathbb{F}_p^n} \left[f(x) \neq \sum_{I \subseteq [t], |I| \geq 1} (-1)^{|I|+1} g(((x \oplus a) + \sum_{i \in I} y_i)^{p-1} \oplus a) \right] \leq (2^t - 1)\delta \ .$$

For any setting of $y_1, \ldots, y_t \in \mathbb{F}_p^n$, define

$$h^{(y_1, \ldots, y_t)}(x) = \sum_{I \subseteq [t], |I| \geq 1} (-1)^{|I|+1} g(((x \oplus a) + \sum_{i \in I} y_i)^{p-1} \oplus a) \ .$$

Let G denote the distribution over the functions $h^{(y_1, \ldots, y_t)}$ obtained by sampling $y_1, \ldots, y_t \in \mathbb{F}_p^n$ uniformly at random. We conclude that for every $x \in \{0,1\}^n$ it holds that

$$\Pr_{h \in G}[f(x) = h(x)] \geq 1 - (2^t - 1)\delta \ .$$

To complete the proof we bound the d-bit-rank of G. Each function $h \in G$ is a linear combination of $g(((x \oplus a) + \sum_{i \in I} y_i)^{p-1} \oplus a) = g^{\oplus a, + \sum_{i \in I} y_i, \oplus a}(x)$, and by Claim 10.3 we know that $\text{bitrank}_d(g^{\oplus a, + \sum_{i \in I} y_i, \oplus a}) \leq \text{bitrank}_d(g) = k$. Therefore, $\text{bitrank}_d(h) \leq (2^t - 1)k$. Consequently, $\text{bitrank}_d(G) \leq (2^t - 1)k$. □

10.4.2 Steps 2 and 3: fooling approximately low bit-rank polynomials

We start by arguing that bit-pseudorandom distributions for degree d polynomials also fool functions with low d-bit-rank.

Claim 10.4. *Let $g : \{0,1\}^n \to \mathbb{F}_p$ be a function with $\mathrm{bitrank}_d(g) = k$. Let $\mathcal{D} \subset \{0,1\}^n$ be a bit-pseudorandom distribution for degree d polynomials with error ϵ. Then $g(\mathcal{D})$ and $g(\{0,1\}^n)$ are $(p^{k/2}\epsilon)$-close.*

Proof. Let $g = \Gamma(f_1(x), \ldots, f_k(x))$ be a representation of g as a function of k polynomials of degree $\leq d$. Denote with $\mathcal{D}_1 \subset \mathbb{F}_p^k$ the joint distribution of $(f_1(x), \ldots, f_k(x))$ when $x \in \{0,1\}^n$ is chosen uniformly at random. Similarly, denote with $\mathcal{D}_2 \subset \mathbb{F}_p^k$ the joint distribution of $(f_1(x), \ldots, f_k(x))$ when $x \in \mathcal{D}$. We will prove that $\mathcal{D}_1$ and $\mathcal{D}_2$ are $(p^{k/2}\epsilon)$-close and hence $g(\mathcal{D})$ and $g(\{0,1\}^n)$ are $(p^{k/2}\epsilon)$-close.

For $\alpha \in \mathbb{F}_p^k$ define $\langle \mathcal{D}_i, \alpha \rangle \subset \mathbb{F}_p$ to be the distribution of the inner product $\langle y, \alpha \rangle$ where $y \in \mathbb{F}_p^k$ is sampled according to $\mathcal{D}_i$. In other words, for $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{F}_p^k$ we have that $\langle \mathcal{D}_1, \alpha \rangle$ is the distribution of $f_\alpha(x) = \sum \alpha_i f_i(x)$ over uniform $x \in \{0,1\}^n$, and that $\langle \mathcal{D}_2, \alpha \rangle$ is the distribution of $f_\alpha(x)$ for $x \in \mathcal{D}$. Since $\mathcal{D}$ is a bit-pseudorandom distribution for degree d polynomials with error ϵ and as each f_α is a degree d polynomial (it is a linear combination of polynomials of degree d), we get that the distributions $\langle \mathcal{D}_1, \alpha \rangle$ and $\langle \mathcal{D}_2, \alpha \rangle$ are ϵ-close. The following well-known fact shows that two distributions with similar Fourier coefficients must be close. For completeness we give the proof in Section 10.9.

Fact 10.1. *Let $\mathcal{D}_1, \mathcal{D}_2 \subset \mathbb{F}_p^k$ be two distributions. Assume that for every $\alpha \in \mathbb{F}_p^k$ the distributions $\langle \mathcal{D}_1, \alpha \rangle$ and $\langle \mathcal{D}_2, \alpha \rangle$ are ϵ-close. Then $\mathcal{D}_1$ and $\mathcal{D}_2$ are $(p^{k/2}\epsilon)$-close.*

It follows that $\mathcal{D}_1$ and $\mathcal{D}_2$ are $(p^{k/2}\epsilon)$-close which concludes the proof. □

We next prove that if a degree $(d+1)$ polynomial f can be pointwise approximated by a distribution with a low d-bit-rank, then f is in fact fooled by bit-pseudorandom distributions for degree d polynomials.

Claim 10.5. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree $d+1$ polynomial. Let G be a distribution over functions $h : \{0,1\}^n \to \mathbb{F}_p$ such that $\mathrm{bitrank}_d(G) = k$, and such that for every $x \in \{0,1\}^n$*

$$\Pr_{h \in G}[f(x) = h(x)] \geq 1 - \delta.$$

Let $\mathcal{D} \subset \{0,1\}^n$ be a bit-pseudorandom distribution for degree d polynomials with error ϵ. Then $f(\mathcal{D})$ and $f(\{0,1\}^n)$ are $(p^{k/2}\epsilon + p\delta)$-close.

Proof. We need to bound

$$\mathrm{sd}(f(\mathcal{D}), f(\{0,1\}^n)) = \tfrac{1}{2} \sum_{t \in \mathbb{F}_p} \left| \Pr_{x \in \mathcal{D}}[f(x) = t] - \Pr_{x \in \{0,1\}^n}[f(x) = t] \right| .$$

Let $E \subset \{0,1\}^n$ be some distribution. We now prove that for every $t \in \mathbb{F}_p$ it holds that

$$\left| \Pr_{x \in E}[f(x) = t] - \Pr_{x \in E, h \in G}[h(x) = t] \right| \leq \delta.$$

First, note that

$$\begin{aligned}\Pr_{x\in E}[f(x)=t] &= \Pr_{x\in E,h\in G}[f(x)=t \ \wedge \ f(x)=h(x)] + \Pr_{x\in E,h\in G}[f(x)=t \ \wedge \ f(x)\neq h(x)] \\ &= \Pr_{x\in E,h\in G}[h(x)=t \ \wedge \ f(x)=h(x)] + \Pr_{x\in E,h\in G}[f(x)=t \ \wedge \ f(x)\neq h(x)]\end{aligned}$$

and

$$\Pr_{x\in E,h\in G}[h(x)=t] = \Pr_{x\in E,h\in G}[h(x)=t \ \wedge \ f(x)=h(x)] + \Pr_{x\in E,h\in G}[h(x)=t \ \wedge \ f(x)\neq h(x)] \ .$$

Therefore, we get that

$$\begin{aligned}&|\Pr_{x\in E}[f(x)=t] - \Pr_{x\in E,h\in G}[h(x)=t]| = \\ &|\Pr_{x\in E,h\in G}[f(x)=t \ \wedge \ f(x)\neq h(x)] - \Pr_{x\in E,h\in G}[h(x)=t \ \wedge \ f(x)\neq h(x)]| \leq \\ &\Pr_{x\in E,h\in G}[f(x)\neq h(x)] = \mathbb{E}_{x\in E}\Pr_{h\in G}[f(x)\neq h(x)] \leq \delta.\end{aligned}$$

The claim now follows as

$$\begin{aligned}2\cdot \mathrm{sd}(f(\mathcal{D}), f(\{0,1\}^n)) &= \sum_{t\in\mathbb{F}_p} |\Pr_{x\in\mathcal{D}}[f(x)=t] - \Pr_{x\in\{0,1\}^n}[f(x)=t]| \\ &\leq \sum_{t\in\mathbb{F}_p} |\Pr_{x\in\mathcal{D},h\in G}[h(x)=t] - \Pr_{x\in\{0,1\}^n,h\in G}[h(x)=t]| \\ &+ \sum_{t\in\mathbb{F}_p} |\Pr_{x\in\mathcal{D}}[f(x)=t] - \Pr_{x\in\mathcal{D},h\in G}[h(x)=t]| \\ &+ \sum_{t\in\mathbb{F}_p} |\Pr_{x\in\{0,1\}^n}[f(x)=t] - \Pr_{x\in\{0,1\}^n,h\in G}[h(x)=t]| \\ &\leq E_{h\in G}[\sum_{t\in\mathbb{F}_p} |\Pr_{x\in\mathcal{D}}[h(x)=t] - \Pr_{x\in\{0,1\}^n}[h(x)=t]|] + 2p\delta \\ &\leq E_{h\in G}[2\cdot \mathrm{sd}(h(\mathcal{D}), h(\{0,1\}^n))] + 2p\delta \\ &\leq 2p^{k/2}\epsilon + 2p\delta.\end{aligned}$$

□

The proof of Lemma 10.2 now follows easily.

Proof of Lemma 10.2. By Lemma 10.3 there is a distribution G on functions such that $\mathrm{bitrank}_d(G) \leq (2^{(p-1)(d+1)+1}-1)k$ and $\Pr_{h\in G}[f(x)=h(x)] \geq 1-(2^{(p-1)(d+1)+1}-1)\delta$. Applying Claim 10.5 we get that the distance between $f(\mathcal{D})$ and $f(\{0,1\}^n)$ is bounded by

$$\mathrm{sd}(f(\mathcal{D}), f(\{0,1\}^n)) \leq p^{2^{(p-1)(d+1)}k}\epsilon + p2^{(p-1)(d+1)+2}\delta \ .$$

□

10.5 The structure of non-regular polynomials

In this section prove of Theorem 10.6. To ease the reading we repeat it here.

Theorem (Theorem 10.6)**.** *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a polynomial of degree $d+1$ for some $d \geq 1$. Assume that, for some $a \in \{0,1\}^n$, the distribution of $f(\mathcal{U}_p \oplus a)$ is ϵ-far from uniform. Then for every $\delta > 0$ there exists a function $g : \{0,1\}^n \to \mathbb{F}_p$ such that $\Pr_{x \in \mathcal{U}_p \oplus a}[g(x) \neq f(x)] \leq \delta$ and $\text{bitrank}_d(g) \leq c + p^c$ where[7] $c = C((p-1)(d+1), \delta\epsilon^2/p^3) = (p^3/\delta\epsilon^2)^{O(4^{(p-1)(d+1)})}$.*

Before giving the actual proof we first give an overview of the main steps.

Step 1: We start by showing that if f is non-regular then it must have a somewhat large 'Fourier coefficient' with respect to a shifted p-biased distribution (Corollary 10.3).

Step 2: Defining $f^{\oplus a}(x) = f(x^{p-1} \oplus a)$ it follows that $f^{\oplus a}$ is a degree $(p-1)(d+1)$ polynomial that has a (relatively) high bias. In addition, Theorem 10.4 implies that there is a relatively small set of variables S such that the weight of the Fourier mass of $f^{\oplus a}$, that is supported on $\bar{S}$, is small.

Step 3: Next we show that if a polynomial has the two properties found in **Step 2** then it can be well approximated by (a function of) a small number of its derivatives and the variables in S. This is the main technical part of the proof (Claim 10.8 and Corollary 10.4). Intuitively, the idea is the following: Note that when $\widehat{f}(0) \neq 0$ we have that $\omega^{-f(x)} = \widehat{f}(0)^{-1}\mathbb{E}_{y \in \mathbb{F}_p^n}\omega^{f_y(x)}$. In our case we show that we can actually get $\omega^{-f(x)} \approx \widehat{f}(0)^{-1}\mathbb{E}_{y \in \mathbb{F}_p^S}\omega^{f_y(x)}$ for most x's, where $\mathbb{F}_p^S$ is the set of n-tuples in $\mathbb{F}_p$ that are supported on S. The reason being that w^f takes discrete values that are 'far' from each other and the average contribution of the derivatives in directions that are not supported on S is small (this follows, after some manipulations, from the structure guaranteed by Theorem 10.4).

Step 4: Using the fact that $f^{\oplus a}(x) = f(x^{p-1} \oplus a)$ we show that each derivative of $f^{\oplus a}$ is actually a function of a small number of the partial derivatives of f and the variables in S (Claim 10.9).

Step 5: From the above steps we get that $f(x^{p-1} \oplus a)$ can be well approximated by the variables in S and a small number of derivatives $f_z(x^{p-1} \oplus a)$. In the last step of the proof we show that we can actually replace the variables in $\{x_i : i \in S\}$ with $\{x_i^{p-1} \oplus a : i \in S\}$. From this we shall conclude that f can be well approximated, with respect to the shifted p-biased distribution, by (a function of) a small number of its derivatives and the variables in S.

We now go to the formal proof according to the steps sketched above.

[7] The function $C(\cdot,\cdot)$ is defined in the statement of Theorem 10.4.

10.5.1 Steps 1 and 2: finding structure in the Fourier spectrum

We start with an easy claim regarding Fourier coefficients of distributions. Abusing notations, given a distribution $\mathcal{D} \subseteq \mathbb{F}_p$ we identify it with the function $\mathcal{D} : \mathbb{F}_p \to [0,1]$ in the following way $\mathcal{D}(y) = \Pr_{x\in\mathcal{D}}[x = y]$. Notice that under this definition $\hat{\mathcal{D}}(0) = 1/p$ and in general,[8] $\hat{\mathcal{D}}(t) = \mathbb{E}_{x\in\mathbb{F}_p}[\mathcal{D}(x)\cdot\omega^{-t\cdot x}] = \frac{1}{p}\mathbb{E}_{x\in\mathcal{D}}[\omega^{-t\cdot x}]$.

Claim 10.6. *Let $\mathcal{D} \subset \mathbb{F}_p$ be a distribution which is ϵ-far from uniform. Then there exists some $t \in \mathbb{F}_p \setminus 0$ such that*

$$p\cdot\hat{\mathcal{D}}(t) = \mathbb{E}_{x\in\mathcal{D}}[\omega^{-t\cdot x}] \geq \epsilon\cdot p^{-1/2}\ .$$

Proof. Let $\mathcal{U}$ denote the uniform distribution on $\mathbb{F}_p$. We have

$$\begin{aligned} 4\epsilon^2 \leq 4\cdot \mathrm{sd}(\mathcal{D},\mathcal{U})^2 &= \left(\sum_{t\in\mathbb{F}_p} |Pr[\mathcal{D}=t] - \Pr[\mathcal{U}=t]|\right)^2 \\ &\leq p^2\mathbb{E}_{t\in\mathbb{F}_p}[|Pr[\mathcal{D}=t]-\Pr[\mathcal{U}=t]|^2] = p^2\sum_{t\in\mathbb{F}_p}|\hat{\mathcal{D}}(t)-\hat{\mathcal{U}}(t)|^2, \end{aligned}$$

where the last equality follows from the Parseval identity. As $\hat{\mathcal{U}}(t) = 0$ for $t \neq 0$, and $\hat{\mathcal{D}}(0) = \hat{\mathcal{U}}(0) = 1/p$ we get

$$\sum_{t\in\mathbb{F}_p\setminus 0}|\hat{\mathcal{D}}(t)|^2 \geq 4\epsilon^2/p^2,$$

hence there is some $t \in \mathbb{F}_p \setminus 0$ such that $|\hat{\mathcal{D}}(t)| \geq \sqrt{\frac{4}{p-1}}\epsilon/p \geq \epsilon\cdot p^{-3/2}$. □

We obtain the following corollary.

Corollary 10.3. *If the distribution $f(\mathcal{U}_p \oplus a)$ is ϵ-far from uniform then there is some $0 \neq t \in \mathbb{F}_p$ such that*

$$\mathbb{E}_{x\in\mathcal{U}_p\oplus a}[\omega^{t\cdot f(x)}] \geq \epsilon\cdot p^{-1/2}\ .$$

Note that we can assume w.l.o.g that $t = 1$. Indeed, let $f' = t\cdot f(x)$. We shall prove that there is a function $g' : \{0,1\}^n \to \mathbb{F}_p$ such that $\Pr_{x\in\mathcal{U}_p\oplus a}[g'(x) \neq f'(x)] \leq \delta$ and $\mathrm{bitrank}_d(g') \leq c + p^c$. Setting $g(x) = t^{-1}\cdot g'(x)$ we get the required polynomial for f. Thus from now on we assume that $t = 1$, i.e. that

$$\mathbb{E}_{x\in\mathcal{U}_p\oplus a}[\omega^{f(x)}] \geq \epsilon\cdot p^{-1/2}\ .$$

Let $f^{\oplus a} : \mathbb{F}_p^n \to \mathbb{F}_p$ be defined as $f^{\oplus a}(x) = f(x^{p-1}\oplus a)$. Since the distribution of $x^{p-1}\oplus a$ for uniform $x \in \mathbb{F}_p^n$ is exactly $\mathcal{U}_p \oplus a$ we get

$$\mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{f^{\oplus a}(x)}] \geq \epsilon\cdot p^{-1/2}\ .$$

Let $\gamma = \delta\epsilon^2/p^2$, for some $\delta > 0$. As $f^{\oplus a}(x)$ is a polynomial of degree at most $(p-1)(d+1)$, Theorem 10.4 implies that there exists a subset $S \subset [n]$ of size $|S| \leq C((p-1)(d+1),\gamma)$ such that

$$\sum_{\alpha\in\mathbb{F}_p^S\setminus 0}|\widehat{f^{\oplus a}}(\alpha)|^2 \leq \gamma\ . \tag{10.2}$$

[8] When speaking of distributions we do not consider the function $\omega^{\mathcal{D}}$ as we do with polynomials.

10.5.2 Step 3: approximating $f^{\oplus a}$ by a few derivatives

Next, we show that the function $f^{\oplus a}(x)$ can be well approximated by a small set of its derivatives. We start with some definitions and a simple yet useful equality. For a subset $S \subset [n]$ let $\mathbb{F}_p^S$ denote the set of vectors $v \in \mathbb{F}_p^n$ which are supported on S, that is,

$$\mathbb{F}_p^S = \{v \in \mathbb{F}_p^n : v_i = 0 \ \forall i \notin S\} .$$

Similarly let $F_p^{\overline{S}}$ denote the set of vectors supported on $\overline{S} = [n] \setminus S$. For $x \in \mathbb{F}_p^n$ let $x_S \in \mathbb{F}_p^S$ denote the part of x which is supported on S, and $x_{\overline{S}}$ the part of x supported on $[n] \setminus S$.

Claim 10.7. *For any function $h : \mathbb{F}_p^n \to \mathbb{F}_p$ and $S \subseteq [n]$ we have*

$$\mathbb{E}_{x \in \mathbb{F}_p^n, y \in \mathbb{F}_p^S}[\omega^{h_y(x)}] = \sum_{\alpha \in \mathbb{F}_p^{\overline{S}}} |\hat{h}(\alpha)|^2 .$$

Proof. Using the Fourier decomposition $\omega^{h(x)} = \sum_{\alpha \in \mathbb{F}_p^n} \hat{h}(\alpha)\omega^{\langle \alpha, x \rangle}$ we get

$$\begin{aligned}
\mathbb{E}_{x \in \mathbb{F}_p^n, y \in \mathbb{F}_p^S}[\omega^{h_y(x)}] &= \mathbb{E}_{x \in \mathbb{F}_p^n, y \in \mathbb{F}_p^S}[\omega^{h(x+y)-h(x)}] \\
&= \mathbb{E}_{x \in \mathbb{F}_p^n, y \in \mathbb{F}_p^S} \sum_{\alpha, \beta \in \mathbb{F}_p^n} \hat{h}(\alpha)\overline{\hat{h}(\beta)}\omega^{\langle \alpha, x+y \rangle - \langle \beta, x \rangle} \\
&= \sum_{\alpha, \beta \in \mathbb{F}_p^n} \hat{h}(\alpha)\overline{\hat{h}(\beta)} \left(\mathbb{E}_{x \in \mathbb{F}_p^n} \omega^{\langle \alpha - \beta, x \rangle}\right) \left(\mathbb{E}_{y \in \mathbb{F}_p^S} \omega^{\langle \alpha, y \rangle}\right) \\
&= \sum_{\alpha \in \mathbb{F}_p^n : \alpha_S = 0} |\hat{h}(\alpha)|^2 = \sum_{\alpha \in \mathbb{F}_p^{\overline{S}}} |\hat{h}(\alpha)|^2 .
\end{aligned}$$

□

We next show that a function h that has a high bias and that satisfy that $\sum_{\alpha \in \mathbb{F}_p^{\overline{S}} \setminus 0} |\hat{h}(\alpha)|^2$ is small can be well approximated by its derivatives in directions from $\mathbb{F}_p^S$. The proof is based on the idea described in **Step 3** above.

Claim 10.8. *Let $h : \mathbb{F}_p^n \to \mathbb{F}_p$ be a function such that $|\mathbb{E}_{x \in \mathbb{F}_p^n}[\omega^{h(x)}]| \geq \epsilon$. Let $\delta > 0$ be an error parameter, and assume there is a subset $S \subset [n]$ such that $\sum_{\alpha \in \mathbb{F}_p^{\overline{S}} \setminus 0} |\hat{h}(\alpha)|^2 \leq \gamma$ for $\gamma = \delta\epsilon^2/p^2$. Then h can be approximated, on a $1 - \delta$ fraction of $\mathbb{F}_p^n$, by a function of its derivatives in directions supported on S. That is, there exists a function $\Gamma : \mathbb{F}_p^{p^{|S|}} \to \mathbb{F}_p$ such that*

$$\Pr_{x \in \mathbb{F}_p^n}[h(x) \neq \Gamma(\{h_y(x) : y \in \mathbb{F}_p^S\})] \leq \delta.$$

Proof. Let $\hat{h}(0) = \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega^{h(x)}]$. By the assumption in the claim, $|\hat{h}(0)| \geq \epsilon$. Define $\phi(x) = \hat{h}(0)^{-1}\mathbb{E}_{y \in \mathbb{F}_p^S}\omega^{h_y(x)}$. We will show that $\phi(x)$ can be used to compute $h(x)$ on most of $\mathbb{F}_p^n$. Define $\Delta(x) = |\omega^{-h(x)} - \phi(x)|$. Note that the minimal distance between different p-th roots of unity, i.e. distinct elements in $\{\omega^t : t \in \mathbb{F}_p\}$, is given by $|1 - \omega| = 2\sin(\frac{\pi}{p}) \geq 2/p$. We will

show that for most $x \in \mathbb{F}_p^n$ we have $\Delta(x) < 1/p$, and hence we can deduce $\omega^{-h(x)}$ uniquely (and therefore $h(x)$) given $\phi(x)$. To achieve this we shall bound $\mathbb{E}_{x\in\mathbb{F}_p^n}[\Delta(x)^2]$ and then use Markov's inequality.

$$\mathbb{E}_{x\in\mathbb{F}_p^n}[\Delta(x)^2] = \mathbb{E}_{x\in\mathbb{F}_p^n}[(\omega^{-h(x)} - \phi(x))\overline{(\omega^{-h(x)} - \phi(x))}] =$$
$$\mathbb{E}_{x\in\mathbb{F}_p^n,y,z\in\mathbb{F}_p^S}[(\omega^{-h(x)} - \hat{h}(0)^{-1}\omega^{h_y(x)})(\omega^{h(x)} - \overline{\hat{h}(0)^{-1}}\omega^{-h_z(x)})] =$$
$$\mathbb{E}_{x\in\mathbb{F}_p^n,y,z\in\mathbb{F}_p^S}[\omega^{-h(x)+h(x)} - \hat{h}(0)^{-1}\omega^{h_y(x)+h(x)} - \overline{\hat{h}(0)^{-1}}\omega^{-h_z(x)-h(x)} + |\hat{h}(0)|^{-2}\omega^{h_y(x)-h_z(x)}] =$$
$$\mathbb{E}_{x\in\mathbb{F}_p^n,y,z\in\mathbb{F}_p^S}[1 - \hat{h}(0)^{-1}\omega^{h(x+y)} - \overline{\hat{h}(0)^{-1}}\omega^{-h(x+z)} + |\hat{h}(0)|^{-2}\omega^{h(x+y)-h(x+z)}] =$$
$$\mathbb{E}_{x\in\mathbb{F}_p^n,y\in\mathbb{F}_p^S}[1 - 1 - 1 + |\hat{h}(0)|^{-2}\omega^{h(x+y)-h(x)}] =$$
$$\mathbb{E}_{x\in\mathbb{F}_p^n,y\in\mathbb{F}_p^S}[-1 + |\hat{h}(0)|^{-2}\sum_{\alpha\in\mathbb{F}_p^{\overline{S}}}|\hat{h}(\alpha)|^2]$$

where the last equality hollows from Claim 10.7. We thus have

$$\mathbb{E}_{x\in\mathbb{F}_p^n}[\Delta(x)^2] = |\hat{h}(0)|^{-2}\sum_{\alpha\in\mathbb{F}_p^{\overline{S}}\setminus 0}|\hat{h}(\alpha)|^2 \leq \frac{\gamma}{\epsilon^2}\ .$$

By Markov's inequality we obtain that

$$\Pr_{x\in\mathbb{F}_p^n}[\Delta(x) \geq 1/p] = \Pr_{x\in\mathbb{F}_p^n}[\Delta(x)^2 \geq 1/p^2] \leq \frac{\mathbb{E}_{x\in\mathbb{F}_p^n}[\Delta(x)^2]}{1/p^2} \leq \frac{p^2\gamma}{\epsilon^2} = \delta.$$

We now define $\Gamma : \mathbb{F}_p^{p^{|S|}} \to \mathbb{F}_p$ as the value of $h(x)$ for which $|\phi(x) - \omega^{-h(x)}|$ is minimized (breaking ties arbitrarily). Since $\phi(x)$ depends only on $\{h_y(x) : y \in \mathbb{F}_p^S\}$ so does Γ and, by the argument above, as long as $\Delta(x) < 1/p$ we know that $\Gamma(x) = h(x)$. Since $\Pr[\Delta(x) \geq 1/p] \leq \delta$, we conclude that

$$\Pr_{x\in\mathbb{F}_p^n}[h(x) \neq \Gamma(\{h_y(x) : y \in \mathbb{F}_p^S\})] \leq \delta\ .$$

□

From Equation (10.2) and Claim 10.8 we obtain the following corollary.

Corollary 10.4. *Let f, a, ϵ be as in the statement of Theorem 10.6. Then, for every $\delta > 0$ there is a function $\Gamma_1 : \mathbb{F}_p^{p^{|S|}} \to \mathbb{F}_p$ and a set $S \subset [n]$, of size $|S| \leq C((p-1)(d+1), \delta\epsilon^2/p^2)$, such that*

$$\Pr_{x\in\mathbb{F}_p^n}[f^{\oplus a}(x) \neq \Gamma_1(\{f_y^{\oplus a}(x) : y \in \mathbb{F}_p^S\})] \leq \delta\ .$$

10.5.3 Step 4: 'fixing' the derivatives

We now show that we can replace the derivatives of $f^{\oplus a}(x)$ in Corollary 10.4 by derivatives of f itself.

Claim 10.9. *For any fixed $y \in \mathbb{F}_p^S$ we have that $(f^{\oplus a})_y(x) = f^{\oplus a}(x+y) - f^{\oplus a}(x)$ is determined by the variables $\{x_i : i \in S\}$ and the derivatives of f supported on S when evaluated on $x^{p-1} \oplus a$. Namely, for every $y \in \mathbb{F}_p^S$ there is a function $\Psi^{(y)} : \mathbb{F}_p^{|S|+p^{|S|}} \to \mathbb{F}_p$ such that for every $x \in \mathbb{F}_p^n$*

$$(f^{\oplus a})_y(x) = \Psi^{(y)}(\{x_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\}) .$$

Proof. Fix $x \in \mathbb{F}_p^n$. Let $w \in \mathbb{F}_p^n$ be defined by the equation $w = ((x+y)^{p-1} \oplus a) - (x^{p-1} \oplus a)$ (interpreted as a pointwise equality). Note that as $y \in \mathbb{F}_p^S$ we also have that $w \in \mathbb{F}_p^S$, i.e. $w_i = 0$ for all $i \notin S$. Moreover, note that we can compute w_S (hence also w) as a fixed function (depending on y, a) of $\{x_i : i \in S\}$. Hence we get

$$(f^{\oplus a})_y(x) = f((x+y)^{p-1} \oplus a) - f(x^{p-1} \oplus a) = f((x^{p-1} \oplus a) + w) - f(x^{p-1} \oplus a) = f_w(x^{p-1} \oplus a) .$$

Consequently, $(f^{\oplus a})_y(x)$ is a function of $\{x_i : i \in S\}$ and $\{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\}$. □

Combining Corollary 10.4 and Claim 10.9 we obtain the following corollary.

Corollary 10.5. *Let f, a, ϵ be as in the statement of Theorem 10.6. Then, for every $\delta > 0$ there is a set $S \subset [n]$, of size $|S| \leq C((p-1)(d+1), \delta\epsilon^2/p^2)$, and a function $\Gamma_2 : \mathbb{F}_p^{p^{|S|}+|S|} \to \mathbb{F}_p$ such that*

$$\Pr_{x \in \mathbb{F}_p^n}[f(x^{p-1} \oplus a) \neq \Gamma_2(\{x_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\})] \leq \delta .$$

10.5.4 Step 5: putting it all together

We now prove Theorem 10.6. Given Corollary 10.5 we basically have to complete **Step 5** in order to conclude the proof. That is, we have to show that f can be well approximated by a function of a small number of its derivatives and the variables in S.

Proof of Theorem 10.6. By Corollary 10.5 there is a function $\Gamma_2 : \mathbb{F}_p^{p^{|S|}+|S|} \to \mathbb{F}_p$ such that

$$\Pr_{x \in \mathbb{F}_p^n}\left[f(x^{p-1} \oplus a) \neq \Gamma_2(\{x_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\})\right] \leq \delta .$$

Thus, we have a function approximating f on Boolean inputs (under the distribution $x^{p-1} \oplus a$) which depends on $\{x_i : i \in S\}$. We will next show how these variables can be replaced by variables of the form $x_i^{p-1} \oplus a_i$, for $i \in S$. Let $u \in (\mathbb{F}_p \setminus 0)^n$ be chosen uniformly at random. Observe that the joint distribution of (x^{p-1}, x) over uniform $x \in \mathbb{F}_p^n$ is identical to the joint distribution of $(x^{p-1}, x^{p-1} \cdot u)$, where the product $x^{p-1} \cdot u$ is taken element-wise. It follows that

$$\Pr_{x \in \mathbb{F}_p^n, u \in (\mathbb{F}_p \setminus 0)^n}\left[f(x^{p-1} \oplus a) \neq \Gamma_2(\{(x_i)^{p-1} \cdot u_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\})\right] \leq \delta .$$

By averaging, there exist a value $u^* \in (\mathbb{F}_p \setminus 0)^n$ such that

$$\Pr_{x \in \mathbb{F}_p^n}\left[f(x^{p-1} \oplus a) \neq \Gamma_2(\{(x_i)^{p-1} \cdot u_i^* : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\})\right] \leq \delta .$$

Notice that given a, u^* we can compute $x_i^{p-1} \cdot u_i^*$ as a function of $x_i^{p-1} \oplus a_i$. Hence, we can define a function $\Gamma : \mathbb{F}_p^{p^{|S|}+|S|} \to \mathbb{F}_p$ such that

$$\Gamma(\{(x_i)^{p-1} \oplus a_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\}) = \\ \Gamma_2(\{(x_i)^{p-1} \cdot u_i^* : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\}) \ .$$

Therefore,

$$\Pr_{x \in \mathbb{F}_p^n}[f(x^{p-1} \oplus a) \neq \Gamma(\{x_i^{p-1} \oplus a_i : i \in S\}, \{f_z(x^{p-1} \oplus a) : z \in \mathbb{F}_p^S\})] \leq \delta \ .$$

As the distribution of $x^{p-1} \oplus a$, for uniform $x \in \mathbb{F}_p^n$, is the same as $\mathcal{U}_p \oplus a$, we conclude that

$$\Pr_{x \in \mathcal{U}_p \oplus a}[f(x) \neq \Gamma(\{x_i : i \in S\}, \{f_z(x) : z \in \mathbb{F}_p^S\})] \leq \delta \ .$$

Set $g(x) = \Gamma(\{x_i : i \in S\}, \{f_z(x) : z \in \mathbb{F}_p^S\})$. Clearly, $\text{bitrank}_d(g) \leq |S| + p^{|S|}$. This completes the proof of the theorem. □

We now give the proof of Theorem 10.5.

Proof of Theorem 10.5. Combining Claim 10.1, Theorem 10.6 and Lemma 10.3 we get that there is a distribution G on functions such that for $c = C((p-1)(d+1), \delta\epsilon^2/4p^3)$ it holds that $\text{bitrank}_d(G) \leq O(2^{(p-1)(d+1)}p^c)$ and $\Pr_{h \in G}[f(x) = h(x)] \geq 1 - (2^{(p-1)(d+1)+1} - 1)\delta$. A simple averaging argument implies that there is some $h \in G$ such that $\Pr_{x \in \{0,1\}^n}[f(x) = h(x)] \geq 1 - (2^{(p-1)(d+1)+1} - 1)\delta$. □

10.6 The Fourier spectrum of low degree polynomials

In this section we give the proof of Theorem 10.4. We start by defining the notion of an S-correlated distribution over $\mathbb{F}_p^n$, for a subset $S \subset [n]$. We recall that for $x \in \mathbb{F}_p^n$ we denote by $x_S \in \mathbb{F}_p^S$ the restriction of x to coordinates in S, and we denote the complement of S by $\bar{S} = [n] \setminus S$.

Definition 10.5. *Let $S \subset [n]$. The S-correlated distribution is a joint distribution over pairs $(X, Y) \in \mathbb{F}_p^n \times \mathbb{F}_p^n$ defined as follows. Choose $X_{\bar{S}} = Y_{\bar{S}}$ uniformly in $\mathbb{F}_p^{\bar{S}}$, and choose independently and uniformly $X_S, Y_S \in \mathbb{F}_p^S$. We denote the S-correlated distribution (X, Y) by $\mathcal{D}_S$. For $f, g : \mathbb{F}_p^n \to \mathbb{F}_p$ and $S \subset [n]$, we define the S-correlation of f and g to be*

$$\Delta_S(f, g) = \sum_{\alpha \in \mathbb{F}_p^n : \alpha_S = 0, \alpha \neq 0} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)} \ .$$

Note that an equivalent definition of $\mathcal{D}_S$ is to first sample $X \in \mathbb{F}_p^n$ uniformly, then to set $Y = X$ and finally to resample Y_S. We now restate Theorem 10.4 in terms of Δ_S.

Theorem 10.7 (Theorem 10.4, restated)**.** *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a degree d polynomial. For every $\epsilon > 0$ there exists $S \subset [n]$, of size $|S| \leq C(d, \epsilon) = O(1/\epsilon)^{O(4^d)}$, such that $\Delta_S(f, f) \leq \epsilon$.*

Before giving the formal proof we explain the idea behind it. We will prove the theorem by induction on the degree. The case of linear polynomials will be easy to handle by a direct calculation. For a general degree d we will use the following useful claims.

Claim 10.10. *Let A be any linear subspace of $\mathbb{F}_p^n$. For every $f : \mathbb{F}_p^n \to \mathbb{F}_p$ and $S \subset [n]$ it holds that $\Delta_S(f,f)^2 \leq \mathbb{E}_{a \in A}[\Delta_S(f_a, f_a)] + \mathbb{E}_{a \in A}[|\widehat{f_a}(0)|^2]$.*

Claim 10.11. *Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$. Let A be a random linear subspace of $\mathbb{F}_p^n$ of dimension r (i.e. A is picked at random amongst all r-dimensional subspaces of $\mathbb{F}_p^n$). Then*

$$\mathbb{E}_A \left[\mathbb{E}_{a \in A}[|\widehat{f_a}(0)|^2] \right] \leq \frac{1}{p^r} + \max_{\alpha} |\widehat{f}(\alpha)|^2 \ ,$$

where $\mathbb{E}_A$ means that we are averaging over a random choice of A.

These claims indicate that we have to consider two cases.

Case 1. All the Fourier coefficients of f are small: In this case, the claims above imply that if we set r to a large enough value and pick a random r-dimensional subspace A then setting S be the union of the corresponding sets for f_a, for $a \in A$, we get the required result (using the induction hypothesis).

Case 2. Some Fourier coefficient of f is large: In this case we first approximate f by a function of a small number of (linear shifts of) its partial derivatives. A simple calculation then gives that for some k, δ^* and σ we have

$$\Delta_S(f,f) \leq \frac{1}{k\delta^*} \sum_{i=1}^{k} |\Delta_S(\widetilde{h_{y_i}}, f)| + 2\sigma \ ,$$

where $\{\widetilde{h_{y_i}}\}_{i=1}^k$ is a set of (shifted) derivatives used to approximate f. Observing that for any g and $S \subseteq S'$ it holds that

$$|\Delta_{S'}(f,g)| \leq (\Delta_S(f,f))^{1/2} (\Delta_S(g,g))^{1/2} \ ,$$

we complete the proof for this case as well by picking S' to be the union of the corresponding sets for the polynomials $\widetilde{h_{y_i}}$.

10.6.1 Proofs of two useful claims

Following the proof outline above we start by proving Claims 10.10 and 10.11. As a first step we prove the following lemma.

Lemma 10.4. *Let $f, g : \mathbb{F}_p^n \to \mathbb{F}_p$. Then for any $S \subset [n]$ it holds that*

$$\Delta_S(f,g) = \mathbb{E}_{(x,y) \in \mathcal{D}_S}[\omega^{f(x)-g(y)}] - \mathbb{E}_{x \in \mathbb{F}_p^n}[\omega^{f(x)}] \overline{\mathbb{E}_{y \in \mathbb{F}_p^n}[\omega^{g(y)}]}$$

and for every $S' \supseteq S$ it holds that

$$|\Delta_{S'}(f,g)| \leq (\Delta_S(f,f))^{1/2} (\Delta_S(g,g))^{1/2} \ .$$

Proof. Recall that $\hat{f}(0) = \mathbb{E}[\omega^{f(x)}]$ and similarly for g. Calculating we get,

$$\begin{aligned}
\sum_{\alpha:\alpha_S=0} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)} &= \sum_{\alpha:\alpha_S=0} (\mathbb{E}_x \omega^{f(x)}\omega^{-\langle x,\alpha\rangle})(\mathbb{E}_y \omega^{-g(y)}\omega^{\langle y,\alpha\rangle}) \\
&= \frac{1}{p^{2n}} \sum_{x,y} \omega^{f(x)-g(y)} \sum_{\alpha:\alpha_S=0} \omega^{\langle y-x,\alpha\rangle} \\
&= \frac{1}{p^{2n}} \sum_{x_{\bar{S}}=y_{\bar{S}}} p^{n-|S|}\omega^{f(x)-g(y)} \\
&= \mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x)-g(y)}] \,.
\end{aligned}$$

Hence, $\Delta_S(f,g) = \mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x)-g(y)}] - \widehat{f}(0)\overline{\widehat{g}(0)}$. To show the second claim we apply the Cauchy-Schwarz inequality,

$$\begin{aligned}
|\Delta_{S'}(f,g)| = \left|\sum_{\alpha\neq 0,\alpha_{S'}=0} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)}\right| &\leq \left(\sum_{\alpha\neq 0,\alpha_{S'}=0} |\widehat{f}(\alpha)|^2\right)^{1/2} \left(\sum_{\alpha\neq 0,\alpha_{S'}=0} |\widehat{g}(\alpha)|^2\right)^{1/2} \\
&\leq \left(\sum_{\alpha\neq 0,\alpha_S=0} |\widehat{f}(\alpha)|^2\right)^{1/2} \left(\sum_{\alpha\neq 0,\alpha_S=0} |\widehat{g}(\alpha)|^2\right)^{1/2} = (\Delta_S(f,f))^{1/2}(\Delta_S(g,g))^{1/2} \,.
\end{aligned}$$

□

We now give the proofs of Claims 10.10 and 10.11.

Proof of Claim 10.10. By Lemma 10.4 we have

$$\Delta_S(f,f) = \mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x)-f(y)}] - |\widehat{f}(0)|^2 \leq \mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x)-f(y)}] \,.$$

For any fixed $a \in A$, the distribution $\{(x+a, y+a) : (x,y) \in \mathcal{D}_S\}$ is identical to $\mathcal{D}_S$. So we can express $\Delta_S(f,f)$ as follows,

$$\Delta_S(f,f) \leq \mathbb{E}_{a\in A}\mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x+a)-f(y+a)}] \,.$$

Applying the Cauchy-Schwarz inequality (and using the fact that A is a linear subspace) we get

$$\begin{aligned}
\Delta_S(f,f)^2 &\leq \mathbb{E}_{(x,y)\in\mathcal{D}_S}\left[|\mathbb{E}_{a\in A}[\omega^{f(x+a)-f(y+a)}]|^2\right] \\
&= \mathbb{E}_{(x,y)\in\mathcal{D}_S}\left[(\mathbb{E}_{a\in A}[\omega^{f(x+a)-f(y+a)}])(\mathbb{E}_{a'\in A}[\omega^{-f(x+a')+f(y+a')}])\right] \\
&= \mathbb{E}_{a,a'\in A}\mathbb{E}_{(x,y)\in\mathcal{D}_S}\left[\omega^{f(x+a)-f(x+a')}\omega^{f(y+a')-f(y+a)}\right] \\
&= \mathbb{E}_{a,a'\in A}\mathbb{E}_{(x',y')\in\mathcal{D}_S}\left[\omega^{f(x'+a-a')-f(x')}\omega^{f(y')-f(y'+a-a')}\right] \\
&= \mathbb{E}_{a\in A}\mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f_a(x)-f_a(y)}] \\
&= \mathbb{E}_{a\in A}[\Delta_S(f_a,f_a) + |\widehat{f_a}(0)|^2] \,.
\end{aligned}$$

□

Proof of Claim 10.11. We begin by showing an identity on $\mathbb{E}_{a\in A}[|\widehat{f_a}(0)|^2]$ for any subspace A.

Claim 10.12. *For any function $f : \mathbb{F}_p^n \to \mathbb{F}_p$ and any subspace $A \subset \mathbb{F}_p^n$*

$$\mathbb{E}_{a\in A}[|\widehat{f_a}(0)|^2] = \sum_{\beta\in\mathbb{F}_p^n,\gamma\in A^\perp} |\widehat{f}(\beta)|^2|\widehat{f}(\beta+\gamma)|^2,$$

where $A^\perp$ is the dual space of A.

Proof. Using the Fourier decomposition formula, the R.H.S of the above expression is

$$\sum_{\beta\in\mathbb{F}_p^n,\gamma\in A^\perp} (\mathbb{E}_{x,x'\in\mathbb{F}_p^n}[\omega^{f(x)-f(x')}\omega^{\langle\beta,x'-x\rangle}])(\mathbb{E}_{y,y'\in\mathbb{F}_p^n}[\omega^{f(y)-f(y')}\omega^{\langle\beta+\gamma,y'-y\rangle}])$$

which is equivalent to

$$\sum_{\gamma\in A^\perp} \mathbb{E}_{x,x',y,y'\in\mathbb{F}_p^n}\left[\omega^{f(x)-f(x')+f(y)-f(y')}\omega^{\langle\gamma,y'-y\rangle}\sum_{\beta\in\mathbb{F}_p^n}\omega^{\langle\beta,x'-x+y'-y\rangle}\right] .$$

Considering the inner sum over β, the above expression can be simplified as

$$\frac{1}{p^{3n}}\sum_{x-x'=y'-y}\omega^{f(x)-f(x')+f(y)-f(y')}\sum_{\gamma\in A^\perp}\omega^{\langle\gamma,y'-y\rangle} .$$

Now the inner sum over γ is nonzero only when $y'-y\in A$. Denote $a = y'-y \in A$. Recalling that we sum over $x-x' = y'-y = a$, we can further simplify the above expression as

$$\frac{|A^\perp|}{p^{3n}}\sum_{a\in A}\sum_{x',y\in\mathbb{F}_p^n}\omega^{f(x'+a)-f(x')+f(y)-f(y+a)} = \mathbb{E}_{a\in A}[|\widehat{f_a}(0)|^2] .$$

□

We now have that

$$\mathbb{E}_{a\in A}[|\widehat{f_a}(0)|^2] = \sum_{\beta\in\mathbb{F}_p^n,\gamma\in A^\perp}|\widehat{f}(\beta)|^2|\widehat{f}(\beta+\gamma)|^2 = \sum_{\beta\in\mathbb{F}_p^n,\alpha\in\mathbb{F}_p^n}|\widehat{f}(\beta)|^2|\widehat{f}(\alpha)|^2\chi_{A^\perp}(\alpha-\beta) ,$$

where $\chi_{A^\perp}$ is the characteristic function of $A^\perp$. Let A be a random subspace of dimension r. The probability for $\alpha\neq\beta$ that $(\alpha-\beta)\in A^\perp$ is $1/p^r$. Since $\sum_\alpha|\hat{f}(\alpha)|^2 = 1$ by Parseval's identity, we obtain that

$$\begin{aligned}\mathbb{E}_A\left[\mathbb{E}_{a\in A}[|\widehat{f_a}(0)|^2]\right] &= \sum_{\beta\neq\alpha\in\mathbb{F}_p^n}|\widehat{f}(\beta)|^2|\widehat{f}(\alpha)|^2\mathbb{E}_A[\chi_{A^\perp}(\alpha-\beta)] + \sum_{\alpha\in\mathbb{F}_p^n}|\widehat{f}(\alpha)|^4 \\ &\leq \frac{1}{p^r} + \sum_{\alpha\in\mathbb{F}_p^n}|\widehat{f}(\alpha)|^4 \leq \frac{1}{p^r} + \max_\alpha|\widehat{f}(\alpha)|^2 .\end{aligned}$$

□

10.6.2 Concluding the proof

We now have all the required ingredients to prove Theorem 10.7.

Proof of Theorem 10.7. The proof is by induction on d. The base case is $d = 1$. Let $f(x) = \sum_{i=1}^{n} a_i x_i$ be any linear polynomial. Consider $S = \{i\}$ such that $a_i \neq 0$. Then for any $\alpha \in \mathbb{F}_p^n$ such that $\alpha_S = 0$ we get $\widehat{f}(\alpha) = \mathbb{E}_{x_i \in \mathbb{F}_p}[\omega^{a_i x_i}] \prod_{j \neq i} \mathbb{E}_{x_j \in \mathbb{F}_p}[\omega^{(a_j - \alpha_j) x_j}] = 0$. Hence, $\sum_{\alpha : \alpha_S = 0} |\widehat{f}(\alpha)|^2 = 0$ and the claim is proved.

By induction hypothesis, let the result be true for any degree $\leq d-1$ polynomial. As outlined above, the proof proceeds by considering two cases, whether f has some large Fourier coefficient or not.

Case 1: Assume that $|\widehat{f}(\alpha)| \leq \delta^*$, for all $\alpha \in \mathbb{F}_p^n$, for an appropriate choice of δ^* (that we will suitably fix later). Let $\epsilon_d = \epsilon$. By Claim 10.12 we get that for any $S \subset [n]$ and a subspace $A \subseteq \mathbb{F}_p^n$

$$\Delta_S(f,f)^2 \leq \mathbb{E}_{a \in A}[\Delta_S(f_a, f_a)] + \mathbb{E}_{a \in A}[|\widehat{f_a}(0)|^2] \; .$$

Notice that for each $a \in A$, $\deg f_a \leq d-1$. Hence, by induction hypothesis, for each $a \in A$, there exist S_a of size $C(d-1, \epsilon_{d-1})$ such that $\Delta_{S_a}(f_a, f_a) \leq \epsilon_{d-1}$ (for some ϵ_{d-1} that will be soon determined). Let A be a linear subspace of dimension r that minimizes $\mathbb{E}_{a \in A}[|\widehat{f_a}(0)|^2]$. Consider $S = \cup_{a \in A} S_a$. Claim 10.11 implies that

$$\Delta_S(f,f)^2 \leq \epsilon_{d-1} + \frac{1}{p^r} + \max_{\alpha} |\widehat{f}(\alpha)|^2 \; .$$

Now it is enough to choose r, ϵ_{d-1} and δ^* such that $\epsilon_{d-1} + \frac{1}{p^r} + (\delta^*)^2 \leq \epsilon_d^2$. Also, notice that $|S| = C(d, \epsilon_d) \leq p^r C(d-1, \epsilon_{d-1})$.

Case 2: Let β be a Fourier coefficient such that $|\widehat{f}(\beta)| \geq \delta^*$. Set $\delta = \overline{\widehat{f}(\beta)}$. Let $h(x) = f(x) - \langle x, \beta \rangle$. Then the bias of $-h(x)$ is $\mathbb{E}_{x \in \mathbb{F}_p^n}[\omega^{-h(x)}] = \delta$. Notice that for every $x \in \mathbb{F}_p^n$ we have $\omega^{h(x)} \mathbb{E}_y[\omega^{-h(x+y)}] = \mathbb{E}_y[\omega^{-h_y(x)}]$. As for every fixed x we have $\mathbb{E}_y[\omega^{-h(x+y)}] = \delta$ it is clear that we can get the following decomposition of $f(x)$

$$\omega^{f(x)} = \omega^{\langle x, \beta \rangle} \cdot \omega^{h(x)} = \omega^{\langle x, \beta \rangle} \cdot \frac{1}{\delta} \mathbb{E}_y[\omega^{-h_y(x)}] = \frac{1}{\delta} \mathbb{E}_y[\omega^{\langle x, \beta \rangle - h_y(x)}] \; .$$

Define $\widetilde{h_y}(x) = \langle x, \beta \rangle - h_y(x)$. Notice that since $h(x)$ has degree $d \geq 2$ we have $\deg(\widetilde{h_y}) \leq d-1$. Now we can expect that if we sample enough y's uniformly and independently at random, and take the average of the corresponding $\omega^{\widetilde{h_y}(x)}$, then we can get a good estimate of $\omega^{f(x)}$. In particular for a parameter $\sigma \in (0,1)$ to be determined later, we find k such that the following holds

$$\mathbb{E}_{x, y_1, \ldots, y_k \in \mathbb{F}_p^n}[|\omega^{f(x)} - \frac{1}{\delta k} \sum_{i=1}^{k} \omega^{\widetilde{h_{y_i}}(x)}|] \leq \sigma.$$

By simple application of Chebyshev's inequality, we estimate the parameter k.

Claim 10.13. *To get an approximation* $\mathbb{E}_{x,y_1,\ldots,y_k\in\mathbb{F}_p^n}[|\omega^{f(x)} - \frac{1}{\delta k}\sum_{i=1}^k \omega^{\widetilde{h_{y_i}}(x)}|] \leq \sigma$, *it is enough to take* $k = O(|\delta|^{-3}\sigma^{-3})$.

Proof. It is enough to choose k such that $\mathbb{E}[|\Re(\omega^{f(x)} - \frac{1}{\delta k}\sum_{i=1}^k \omega^{\hat{h}_{y_i}(x)})|] \leq \sigma/2$, and $\mathbb{E}[|\mathrm{Img}(\omega^{f(x)} - \frac{1}{\delta k}\sum_{i=1}^k \omega^{\hat{h}_{y_i}(x)})|] \leq \sigma/2$. Let $Y_i = \Re(\frac{1}{\delta}\omega^{\widetilde{h_{y_i}}(x)})$. Then $\mathbb{E}_{y_i}[Y_i] = \Re(\omega^{f(x)})$. It is clear that $\mathrm{Var}(Y_i) \leq \frac{1}{|\delta|^2}$. Hence, by Chebyshev's inequality we get that

$$\Pr(|\Re(\omega^{f(x)}) - \frac{1}{k}\sum_{i=1}^k Y_i| \geq \frac{\sigma}{4}) \leq \frac{16}{|\delta|^2 k\sigma^2}\ .$$

Therefore, as always $|\Re(\omega^{f(x)}) - \frac{1}{k}\sum_{i=1}^k Y_i| \leq 1 + \delta^{-1} \leq 2\delta^{-1}$ we get that $\mathbb{E}[|\Re(\omega^{f(x)}) - \frac{1}{k}\sum_{i=1}^k Y_i|] \leq \sigma/2$ for $k \geq \frac{128}{|\delta|^3\sigma^3}$. The imaginary part can be approximated similarly. □

Fix $\{y_i\}_{i\in[k]}$ in such a way that $\mathbb{E}_{x\in\mathbb{F}_p^n}[|\omega^{f(x)} - \frac{1}{\delta k}\sum_{i=1}^k \omega^{\widetilde{h_{y_i}}(x)})|] \leq \sigma$. Let $F(x) = \frac{1}{k\delta}\sum_{i=1}^k \omega^{\widetilde{h_{y_i}}(x)}$. As $\mathbb{E}_{x\in\mathbb{F}_p^n}[|\omega^{f(x)} - F(x)|] \leq \sigma$ we can upper bound $\Delta_S(f,f)$ as follows

$$\begin{aligned}
\Delta_S(f,f) &= \mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{f(x)-f(y)}] - \mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{f(x)}]\cdot\overline{\mathbb{E}_{y\in\mathbb{F}_p^n}[\omega^{f(y)}]} \\
&\leq |\mathbb{E}_{(x,y)\in\mathcal{D}_S}[(\omega^{f(x)} - F(x))\omega^{-f(y)}] - \mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{f(x)} - F(x)]\cdot\overline{\mathbb{E}_{y\in\mathbb{F}_p^n}[\omega^{f(y)}]}| \\
&\quad + |\mathbb{E}_{(x,y)\in\mathcal{D}_S}[F(x)\omega^{-f(y)}] - \mathbb{E}_{x\in\mathbb{F}_p^n}[F(x)]\cdot\overline{\mathbb{E}_{y\in\mathbb{F}_p^n}[\omega^{f(y)}]}| \\
&\leq 2\sigma + |\mathbb{E}_{(x,y)\in\mathcal{D}_S}[F(x)\omega^{-f(y)}] - (\mathbb{E}_x[F(x)])(\mathbb{E}_y[\omega^{-f(y)}])| \\
&\leq 2\sigma + \frac{1}{k\delta}\sum_{i=1}^k |\mathbb{E}_{(x,y)\in\mathcal{D}_S}[\omega^{\widetilde{h_{y_i}}(x)-f(y)}] - (\mathbb{E}_x[\omega^{\widetilde{h_{y_i}}(x)}])(\mathbb{E}_y[\omega^{-f(y)}])| \\
&\leq 2\sigma + \frac{1}{k\delta^*}\sum_{i=1}^k |\Delta_S(\widetilde{h_{y_i}}, f)|.
\end{aligned}$$

As $\deg(\widetilde{h_{y_i}}) \leq d-1$ we get, by the induction hypothesis, that for each $\widetilde{h_{y_i}}$ there exists a set S_i, of size $C(d-1,\epsilon_{d-1})$, such that $\Delta_{S_i}(\widetilde{h_{y_i}}, \widetilde{h_{y_i}}) \leq \epsilon_{d-1}$. Consider $S = \cup_{i=1}^k S_i$. Obviously, $|S| \leq kC(d-1,\epsilon_{d-1})$. Lemma 10.4 implies that

$$|\Delta_S(\widetilde{h_{y_i}}, f)| \leq (\Delta_{S_i}(\widetilde{h_{y_i}}, \widetilde{h_{y_i}}))^{1/2}(\Delta_{S_i}(f,f))^{1/2} \leq (\Delta_{S_i}(\widetilde{h_{y_i}}, \widetilde{h_{y_i}}))^{1/2} \leq \epsilon_{d-1}^{1/2}\ .$$

In order to achieve $\Delta_S(f,f) \leq \epsilon_d$ we need to fix the parameters $\delta^*, \epsilon_{d-1}, k, \sigma$ so that $\frac{1}{\delta^*}\epsilon_{d-1}^{1/2} + 2\sigma \leq \epsilon_d$.

We now show how to pick the parameters adequately. We need to satisfy both $\epsilon_{d-1} + \frac{1}{p^r} + (\delta^*)^2 \leq \epsilon_d^2$ and $\frac{1}{\delta^*}\epsilon_{d-1}^{1/2} + 2\sigma \leq \epsilon_d$. Fix $\sigma = \frac{\epsilon_d}{4}$ and $\delta^* = \frac{\epsilon_d}{2}$. Then it is enough to choose $\epsilon_{d-1} = O(\epsilon_d^4)$ and $r = \log_p(\epsilon_d^2/4)$. We now estimate $|S|$. Recall that $|S| \leq \max(p^r, k)C(d-1,\epsilon_{d-1})$ where $k = O(|\delta^*|^{-3}\sigma^{-3})$. This yields the following bound

$$|S| \leq O(\epsilon_d^{-6})C(d-1,\Omega(\epsilon_d^4))$$

Solving the recurrence for $C(d,\epsilon)$ we get that $C(d,\epsilon) \leq O(\epsilon)^{O(4^d)}$. This completes the proof of the theorem. □

10.7 Conclusions and open problems

We construct efficient and explicit bit-pseudorandom generators for constant degree polynomials over finite fields. These yield pseudorandom generators for $\mathrm{CC}_0[p]$ which achieve any small constant error while using only $O(\log n)$ random bits. The proof is based on a new characterization of the Fourier spectrum of low degree polynomials over finite fields.

We state several open problems.

- Construct pseudorandom generators for $\mathrm{AC}_0[p]$. The next step, following this work, is to construct pseudorandom generators for sparse polynomials over $\mathbb{F}_p$ (i.e. polynomials of degree $O(\log n)$ with only a polynomial number of monomials). Any such polynomial can be realized by a depth-2 $\mathrm{AC}_0[p]$ circuit.

- Generalize our results for $\mathrm{CC}_0[m]$ for composite m. As a first step, generalize our results for bit-pseudorandom generators for low degree polynomials over $\mathbb{Z}_m$.

- Improve the parameters of Theorem 10.4. For $d = 1$ it is an easy observation that a set S of size $|S| = 1$ suffices. For $d = 2$, it is not difficult to see that all nonzero Fourier coefficients of a quadratic polynomial form an affine space and have the same absolute value. Using this observation one can get a set of size $|S| = O(\log 1/\epsilon)$. We do not have any example of a constant degree polynomial requiring sets of size $\omega(\log 1/\epsilon)$.

- Improve the dependence of the seed length on ϵ in Theorem 10.3. Currently, the seed length is logarithmic in n but a tower of height $O(d)$ in $1/\epsilon$.

10.8 Proof for linear polynomials

In this section we give the proof Theorem 10.2. For convenience we repeat it here.

Theorem (Bit-pseudorandom distribution for linear polynomials)**.** *Let $\mathbb{F}_p$ be a prime finite field and $\epsilon > 0$ be an error parameter. Let $\mathcal{D} \subset \mathbb{F}_p^n$ be a pseudorandom distribution for degree $p-1$ polynomials over $\mathbb{F}_p$ with error ϵ. Let $K \subset \{0,1\}^n$ be a k-wise independent distribution for $k = O(p^3 \log 1/\epsilon)$. Then $\mathcal{D}^{p-1} \oplus K$ is bit-pseudorandom distribution against linear polynomials over $\mathbb{F}_p$ with error $O(\epsilon)$.*

Proof. Let $f(x) = \sum_{i=1}^n a_i x_i$ be some linear polynomial. Define the *weight* of f, $\mathrm{wt}(f)$, to be the number of nonzero coefficients in f. We consider two cases. Consider first the case that $\mathrm{wt}(f) \leq k$. In such a case any k-wise independent distribution fools f completely. The distribution $\mathcal{D}^{p-1} \oplus K$ is k-wise independent, hence it fools f with error 0.

We now move to the second case, where $\mathrm{wt}(f) > k$. We will prove that in this case the distribution of $f(x)$ is $O(\epsilon)$-close to the uniform distribution over $\mathbb{F}_p$, both when $x \in \{0,1\}^n$ is chosen uniformly at random and when we choose $x \in \mathcal{D}^{p-1} \oplus K$. Hence these two distributions are $O(\epsilon)$-close to each other. In fact, we shall prove a stronger claim: for any fixed $v \in \{0,1\}^n$, the distribution of $f(x)$ where $x \in \mathcal{D}^{p-1} \oplus v$ is $O(\epsilon)$-close to uniform.

We first note that if $X \in \mathbb{F}_p$ is a distribution, then Fact 10.1 shows that in order to prove that X is ϵ-close to uniform it suffices to prove that for any $c \in \mathbb{F}_p \setminus 0$ it holds that

$\mathbb{E}[\omega^{cX}] \leq \epsilon/\sqrt{p}$. Since multiplying by $c \neq 0$ does not change the weight of f, it is enough to prove that if $\mathrm{wt}(f) > k$ then $|\mathbb{E}_{x\in\{0,1\}^n}[\omega^{f(x)}]| \leq O(\epsilon)$ and $|\mathbb{E}_{x\in\mathcal{D}^{p-1}\oplus a}[\omega^{f(x)}]| \leq O(\epsilon)$.

We first prove the claim for uniform inputs. Note that if $z \in \{0,1\}$ is uniform and $a \neq 0$, then

$$\left|\mathbb{E}_{z\in\{0,1\}}[\omega^{az}]\right| \leq 1 - \Omega(1/p^2) \; .$$

Therefore, as $\mathrm{wt}(f) > k$ we get

$$\left|\mathbb{E}_{x\in\{0,1\}^n}[\omega^{f(x)}]\right| = \prod_{i=1}^{n} \left|\mathbb{E}_{x_i\in\{0,1\}}[\omega^{a_i x_i}]\right| \leq (1 - \Omega(1/p^2))^k = O(\epsilon) \; .$$

We now move to proving the claim for $x \in \mathcal{D}^{p-1} \oplus v$. That is, we wish to prove that

$$|\mathbb{E}_{x\in\mathcal{D}}[\omega^{\sum a_i(x_i^{p-1}\oplus v_i)}]| = O(\epsilon) \; .$$

Define $g : \mathbb{F}_p^n \to \mathbb{F}_p$ as $g(x) = \sum a_i(x_i^{p-1} \oplus v_i)$. Note that g is a polynomial of degree $p-1$, as $x_i^{p-1} \oplus v_i$ is equal to x_i^{p-1} when $v_i = 0$ and is equal to $1 - x_i^{p-1}$ when $v_i = 1$. Since $\mathcal{D}$ is a pseudorandom distribution for degree $p-1$ polynomials with error ϵ, we get that

$$\left|\mathbb{E}_{x\in\mathcal{D}}[\omega^{g(x)}] - \mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{g(x)}]\right| \leq \epsilon.$$

Hence it is enough to prove that $\left|\mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{g(x)}]\right| = O(\epsilon)$. For that end, let $y \in \{0,1\}^n$ be distributed as follows: $y_1, \ldots, y_n$ are chosen independently such that $\Pr[y_i = v_i] = 1/p$. Then, for $x \in \mathbb{F}_p^n$ chosen uniformly at random, the distributions of $x^{p-1} \oplus v$ and of y are identical. Moreover it is straightforward to verify that for any $a_i \neq 0$ we have

$$|\mathbb{E}_{y_i}[\omega^{a_i y_i}]| \leq 1 - \Omega(1/p^3) \; .$$

The claim now follows as

$$\begin{aligned}\left|\mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{g(x)}]\right| &= \left|\mathbb{E}_{x\in\mathbb{F}_p^n}[\omega^{\sum a_i(x_i^{p-1}\oplus v_i)}]\right| = \left|\mathbb{E}_{y}[\omega^{\sum a_i y_i}]\right| \\ &= \prod_{i=1}^{n} |\mathbb{E}_{y_i}[\omega^{a_i y_i}]| \leq (1 - \Omega(1/p^3))^k = O(\epsilon) \; .\end{aligned}$$

□

10.9 Proof of Fact 10.1

For completeness we give the proof of the following well known fact.

Fact (Fact 10.1)**.** *Let $\mathcal{D}_1, \mathcal{D}_2 \subset \mathbb{F}_p^k$ be two distributions. Assume that for every $\alpha \in \mathbb{F}_p^k$ the distributions $\langle \mathcal{D}_1, \alpha\rangle$ and $\langle \mathcal{D}_2, \alpha\rangle$ are ϵ-close. Then $\mathcal{D}_1$ and $\mathcal{D}_2$ are $(p^{k/2}\epsilon)$-close.*

Proof of Fact 10.1. We need to bound

$$\mathrm{sd}(\mathcal{D}_1, \mathcal{D}_2) = \tfrac{1}{2} \sum_{x \in \mathbb{F}_p^k} |\Pr[\mathcal{D}_1 = x] - \Pr[\mathcal{D}_2 = x]| \ .$$

By the Cauchy-Schwarz inequality we get

$$4 \cdot \mathrm{sd}(\mathcal{D}_1, \mathcal{D}_2)^2 \leq p^k \sum_{x \in \mathbb{F}_p^k} |\Pr[\mathcal{D}_1 = x] - \Pr[\mathcal{D}_2 = x]|^2 \ .$$

Let $\widehat{\mathcal{D}_i} : \mathbb{F}_p^k \to \mathbb{C}$ be the Fourier transform of $\mathcal{D}_i$, when we think of $\mathcal{D}_i$ as the function $\mathcal{D}_i(y) = \Pr_{x \in \mathcal{D}_i}[x = y]$. In other words, $\widehat{\mathcal{D}_i}(\alpha) = \mathbb{E}_{x \in \mathbb{F}_p^k}[\Pr[\mathcal{D}_i = x]\omega^{-\langle x, \alpha \rangle}]$. By the Parseval identity we get that

$$4 \cdot \mathrm{sd}(\mathcal{D}_1, \mathcal{D}_2)^2 \leq p^{2k} \cdot \sum_{\alpha \in \mathbb{F}_p^k} |\widehat{\mathcal{D}_1}(\alpha) - \widehat{\mathcal{D}_2}(\alpha)|^2 \ .$$

From the assumption that for every $\alpha \in \mathbb{F}_p^k$ the distributions $\langle \mathcal{D}_1, \alpha \rangle$ and $\langle \mathcal{D}_2, \alpha \rangle$ are ϵ-close we obtain

$$\begin{aligned} |\widehat{\mathcal{D}_1}(\alpha) - \widehat{\mathcal{D}_2}(\alpha)| &= \left| E_{t \in \mathbb{F}_p^k}[(Pr[\langle \mathcal{D}_1, \alpha \rangle = t] - Pr[\langle \mathcal{D}_2, \alpha \rangle = t]) \cdot \omega^{-t}] \right| \\ &\leq \mathbb{E}_{t \in \mathbb{F}_p^k} |[Pr[\langle \mathcal{D}_1, \alpha \rangle = t] - Pr[\langle \mathcal{D}_2, \alpha \rangle = t]]| \\ &\leq 2\epsilon / p^k \ . \end{aligned}$$

Thus we conclude that

$$4 \cdot \mathrm{sd}(\mathcal{D}_1, \mathcal{D}_2)^2 \leq p^{2k} \sum_{\alpha \in \mathbb{F}_p^k} (2\epsilon / p^k)^2 = 4p^k \epsilon^2 \ .$$

The bound $\mathrm{sd}(\mathcal{D}_1, \mathcal{D}_2) \leq p^{k/2}\epsilon$ now follows. □

Part IV

Coding theory

Chapter 11

List Size vs. Decoding Radius for Reed-Muller Codes

The weight distribution and list-decoding size of Reed-Muller codes are studied in this work. Given a weight parameter, we are interested in bounding the number of Reed-Muller codewords with a weight of up to the given parameter. Additionally, given a received word and a distance parameter, we are interested in bounding the size of the list of Reed-Muller codewords that are within that distance from the received word. In this work, we make a new connection between computer science techniques used for studying low-degree polynomials and these coding theory questions. Using this connection we progress significantly towards resolving both the weight distribution and the list-decoding problems.

Obtaining tight bounds for the weight distribution of Reed-Muller codes has been a long standing open problem in coding theory, dating back to 1976 and seemingly resistent to the common coding theory tools. The best results to date are by Azumi, Kasami and Tokura which provide bounds on the weight distribution that apply only up to 2.5 times the minimal distance of the code. We provide asymptotically tight bounds for the weight distribution of the Reed-Muller code that apply to *all* distances.

List-decoding has both theoretical and practical applications in various fields. To name a few, hardness amplification in complexity, constructing hard-core predicates from one way functions in cryptography and learning parities with noise in learning theory.

Many algorithms for list-decoding have the crux of their analysis lying in bounding the list-decoding size. The case for Reed–Muller codes is similar, and Gopalan, Klivans and Zuckerman gave a list-decoding algorithm, whose complexity is determined by the list-decoding size. Gopalan et. al provided bounds on the list-decoding size of Reed–Muller codes which apply only up to the minimal distance of the code. We provide asymptotically tight bounds for the list-decoding size of Reed–Muller codes which apply to *all* distances.

Joint work with Tali Kaufman and Ely Porat.

11.1 Introduction

The weight distribution of an error correcting code counts, for every given weight parameter, the number of codewords with weight bounded by the given parameter. The weight distribution of a code is the main characteristic of the code, and governs the behavior of the code, from both theoretical and practical aspects.

Understanding the weight distribution of Reed-Muller codes is a 30-year-old standing open question in coding theory. The last progress on this question was made by Kasami and Tokura [KT70] that characterized the codewords of Reed-Muller codes of weight up to twice the minimal distance of the code, and hence obtained bounds for the weight distribution that apply till twice the minimal distance of the code. In this work we study the weight distribution of Reed Muller codes and provide asymptotically tight bounds that apply to all distances.

The problem of list-decoding an error correcting code is the following: given a received word and a distance parameter find all codewords of the code that are within the given distance from the received word. List-decoding is a generalization of the more common notion of unique decoding in which the given distance parameter ensures that there can be at most one codeword of the code that is within the given distance from the received word. The notion of list-decoding has numerous practical and theoretical implications. The breakthrough results in this field are due to Goldreich and Levin [GL89] and Sudan [Sud97] who gave efficient list decoding algorithms for the Hadamard code and the Reed-Solomon code. See surveys by Guruswami [Gur04] and Sudan [Sud00] for further details. In complexity, list-decodable codes are used to perform hardness amplification of functions [STV99]. In cryptography, list-decodable codes are used to construct hard-core predicates from one way functions [GL89]. In learning theory, list decoding of Hadamard codes implies learning parities with noise [KM93].

In this work we study the question of list-decoding Reed-Muller codes. Specifically, we are interested in bounding the list sizes obtained for different distance parameters for the list-decoding problem. Our work provides asymptotically tight bounds that apply to all distances. The improved bounds, imply improved algorithms for list-decoding Reed-Muller codes.

Our results are obtained by making a new connection between computer science techniques used for studying low-degree polynomials and the discussed coding theory questions. Using this connection we manage to progress significantly towards resolving these two important open problems.

Our proofs are technically relatively simple. We view this as evidence to the importance of this new connection, since these were considered as open problems, resistent to the more common coding theory tools. We view this as the main innovation of our work.

11.1.1 Reed–Muller codes

Reed-Muller codes are a very fundamental and well studied family of codes. $\mathrm{RM}(n,d)$ is a linear code, whose codewords $f \in \mathrm{RM}(n,d) : \mathbb{F}_2^n \to \mathbb{F}_2$ are evaluations of polynomials in n variables of total degree at most d over $\mathbb{F}_2$. In this work we study the code $\mathrm{RM}(n,d)$ when

$d \ll n$, and are interested in particular in the case of constant d.

The following facts regarding $\mathrm{RM}(n,d)$ are straight-forward: It has block length of 2^n, dimension $\sum_{i \le d} \binom{n}{i}$ and minimum relative distance $\frac{2^{n-d}}{2^n} = 2^{-d}$.

11.1.2 Weight distribution of Reed-Muller codes

We now formally define the weight distribution of a code, and discuss previous known bounds for the weight distribution of Reed-Muller codes.

Definition 11.1 (Relative weight)**.** *The relative weight of a function/codeword $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is the fraction of non-zero elements,*

$$\mathrm{wt}(f) = \frac{1}{2^n} |\{x \in \mathbb{F}_2^n : f(x) = 1\}|$$

Definition 11.2 (Accumulative weight distribution)**.** *The accumulative weight distribution of $\mathrm{RM}(n,d)$ at a relative weight α is the number of codewords up to this weight, i.e.*

$$A(\alpha) = |\{p \in \mathrm{RM}(n,d) : \mathrm{wt}(p) \le \alpha\}|$$

where $0 \le \alpha \le 1$.

It is well-known that for any $p \in \mathrm{RM}(n,d)$ which is not identically zero, $\mathrm{wt}(p) \ge 2^{-d}$. Thus, $A(2^{-d} - \epsilon) = 1$ for any $\epsilon > 0$. Kasami and Tokura [KT70] characterized the codewords in $\mathrm{RM}(n,d)$ of weight up to twice the minimal distance of the code (i.e up to distance 2^{1-d}). Based on their characterization one could conclude the following.

Corollary 11.1 (Corollary 10 in [GKZ08])**.**

$$A(2^{1-d} - \epsilon) \le (1/\epsilon)^{2(n+1)}$$

Corollary 11.1 and simple lower bounds (which we show later, see Lemma 11.5) show that $A(\alpha) = 2^{\Theta(n)}$ for $\alpha \in [2^{-d}, 2^{1-d} - \epsilon]$ for any $\epsilon > 0$ (and constant d).

11.1.3 List-decoding size of Reed-Muller codes

We now formally define the list-decoding size of a code, and discuss previous known bounds for the list-decoding size of Reed-Muller codes. Moreover we discuss known list-decoding algorithms for Reed-Muller codes. We start with the following definition.

Definition 11.3 (Relative distance between two functions)**.** *The relative distance between two functions $f, g : \mathbb{F}_2^n \to \mathbb{F}_2$ is defined as*

$$\mathrm{dist}(f,g) = \mathbb{P}_{x \in \mathbb{F}_2^n}[f(x) \ne g(x)]$$

Our work focuses on understanding the asymptotic growth of the list size in list-decoding of Reed-Muller codes, as a function of the distance parameter. Specifically we are interested in obtaining bounds on the following.

Definition 11.4 (List-decoding size)**.** *For a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ let the ball at relative distance α around f be*

$$B(f, \alpha) = \{p \in \mathrm{RM}(n, d) : \mathrm{dist}(p, f) \leq \alpha\}$$

The list-decoding size of $\mathrm{RM}(n, d)$ *at distance α, denoted by $L(\alpha)$, is the maximal size of $B(f, \alpha)$ over all possible functions f, i.e.*

$$L(\alpha) = \max_{f:\mathbb{F}_2^n \to \mathbb{F}_2} |B(f, \alpha)|$$

In a recent work, Gopalan, Klivans and Zuckerman [GKZ08] proved that for distances up to the minimal distance of the code, the list-decoding size of Reed-Muller codes remains constant.

Theorem 11.1 (Theorem 11 in [GKZ08])**.**

$$L(2^{-d} - \epsilon) \leq O\left((1/\epsilon)^{8d}\right)$$

Their result of bounding the list-decoding size of Reed-Muller codes is inherently limited to work up to the minimum distance of the code, since it uses the structural theorem of Kasami and Takura on Reed-Muller codes [KT70], which implies a bound on the weight distribution of Reed-Muller codes that works up to twice the minimum distance of the code.

Additionally, the work of [GKZ08] has developed a list-decoding algorithm for $\mathrm{RM}(n, d)$ whose running time is polynomial in the worst list-decoding size and in the block length of the code.

Theorem 11.2 (Theorem 4 in [GKZ08])**.** *Given a distance parameter α and a received word $R : \mathbb{F}_2^n \to \mathbb{F}_2$, there is an algorithm that runs in time poly$(2^n, L(\alpha))$ and produces a list of all $p \in \mathrm{RM}(n, d)$ such that* $\mathrm{dist}(p, R) \leq \alpha$*.*

Since Gopalan et al. could obtain non-trivial bounds on the list-decoding size for distance parameter α that is bounded by the minimum distance of the Reed-Muller code, their algorithm running time could be analyzed only for α that is less than the minimum distance of the code. This supports our earlier statement, that the crux of the analysis of list-decoding algorithms is in bounding the list-decoding size.

11.1.4 Our Results

The weight distribution of $\mathrm{RM}(n, d)$ codes beyond twice the minimum distance was widely open prior to our work. See e.g. Research Problem (15.1) in [MS83] and the related discussion in that chapter. In this work we provide asymptotic bounds for the weight distribution of $\mathrm{RM}(n, d)$ that applied for all weights $2^{-d} \leq \alpha \leq 1/2$. We state now our results for constant d, where the notation $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ hides constants depending only on d. Our first main result gives exact boundaries on the range of α for which $A(\alpha) = 2^{\Theta(n^\ell)}$, for any $\ell = 1, 2, ..., d$, showing there are "cut-off distances", at which the accumulative weight distribution jumps from $2^{\Theta(n^\ell)}$ to $2^{\Theta(n^{\ell+1})}$.

Theorem 11.3 (First main theorem - accumulative weight distribution)**.** *Let* $1 \leq \ell \leq d-1$ *be an integer, and let* $\epsilon > 0$*. For any* $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$

$$2^{\Omega(n^\ell)} \leq A(\alpha) \leq (1/\epsilon)^{O(n^\ell)}$$

and $A(\alpha) = 2^{\Theta(n^d)}$ *for any* $\alpha \geq 1/2$*.*

We also address the more general problem of bounding the list-decoding size. Gopalan et al. [GKZ08] left as an open problem the question of bounding the list-decoding size of Reed-Muller codes beyond the minimal distance. We give tight bounds on the list-decoding size of Reed–Muller codes that apply to all distances. In fact, we show that the behavior of the list-decoding size is asymptotically identical to that of the accumulative weight distribution.

Theorem 11.4 (Second main theorem - list-decoding size)**.** *Let* $1 \leq \ell \leq d-1$ *be an integer, and let* $\epsilon > 0$*. For any* $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$

$$2^{\Omega(n^\ell)} \leq L(\alpha) \leq (1/\epsilon)^{O(n^\ell)}$$

and $L(\alpha) = 2^{\Theta(n^d)}$ *for any* $\alpha \geq 1/2$*.*

Using Theorem 11.4 and Theorem 11.2, we obtain the following algorithmic result for list-decoding Reed-Muller codes.

Theorem 11.5 (List-decoding algorithm)**.** *Let* $R : \mathbb{F}_2^n \to \mathbb{F}_2$ *be a received word. Let* $\alpha \in [2^{\ell-d-1}, 2^{\ell-d} - \epsilon]$ *be a required distance parameter, where* $1 \leq \ell \leq d-1$ *is integer and* $\epsilon > 0$*. There exists an algorithm that runs in time* $(1/\epsilon)^{O(n^\ell)}$ *and produces a list of all* $p \in \mathrm{RM}(n,d)$ *such that* $\mathrm{dist}(p, R) \leq \alpha$*.*

Observe that Theorems 11.3 and 11.4 are asymptotically tight even for sub-constant values of ϵ. The smallest possible value is $\epsilon = 2^{-n}$, and indeed for $\alpha = 2^{\ell-d} - \epsilon$ we get that both $A(\alpha)$ and $L(\alpha)$ are upper bounded by $(1/\epsilon)^{O(n^\ell)} = 2^{O(n^{\ell+1})}$, while for $\alpha = 2^{\ell-d}$ they are lower bounded by $2^{O(n^{\ell+1})}$.

11.1.5 Techniques

Our results are obtained by making a new connection between computer science techniques used for studying low-degree polynomials and weight distribution and list-decoding size of Reed-Muller codes. Evidence of the importance of this new connection is the technical simplicity of our proofs that solve these well-known open problems. Following is a detailed discussion of our techniques.

The bounds on the accumulative weight distribution of the Reed-Muller code are obtained using the following novel strategy. We study the structure of functions $f : \mathbb{F}_2^n \to \mathbb{F}_2$ based on their discrete derivatives. The discrete derivative of f in direction $y \in \mathbb{F}_2^n$ is given by

$$f_y(x) = f(x+y) + f(x).$$

The k-iterated derivative of f in directions $y_1, \ldots, y_k \in \mathbb{F}_2^n$ is given by

$$f_{y_1,\ldots,y_k}(x) = (f_{y_1,\ldots,y_{k-1}})_{y_k}(x) = \sum_{I \subseteq \{1,\ldots,k\}} f(x + \sum_{i \in I} y_i).$$

Note that if f is an n-variate polynomial over $\mathbb{F}_2$ of total degree d, then any derivative of it is a polynomial of total degree at most $d-1$, and any k-iterated derivative of it is a polynomial of total degree at most $d-k$. This is an important property that is crucial to our proof.

As a first step to bounding the weight distribution of Reed-Muller codes we establish the following general result. We show that a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ whose weight is bounded by $\mathrm{wt}(f) \leq 2^{-k}(1-\epsilon)$ can be approximated by a universal function $\mathcal{A}$ of a small number of the k-iterated derivatives of f (Lemma 11.1). That is, for any approximation parameter $\delta > 0$, there exist $c = c(k, \epsilon, \delta)$ sets of k-iterated derivatives $\{y_{i,1}, \ldots, y_{i,k}\}_{1 \leq i \leq c}$, such that

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) \neq \mathcal{A}(\{y_{i,j}\}, f_{y_{1,1},\ldots,y_{1,k}}(x), \ldots, f_{y_{c,1},\ldots,y_{c,k}}(x))] < \delta.$$

We accomplish this in three steps. It will be useful to represent functions $f : \mathbb{F}_2^n \to \mathbb{F}_2$ as $(-1)^f : \mathbb{F}_2^n \to \{-1, 1\}$. First, we show that the function $(-1)^{f(x)}$ can be *computed* as an expectation of its $(k-1)$-iterated derivatives $(-1)^{f_{y_1,\ldots,y_{k-1}}(x)}$ multiplied by some bounded coefficients (Lemma 11.2). Moreover, we show that each of the $(k-1)$-iterated derivatives is biased (a function $g : \mathbb{F}_2^n \to \{-1, 1\}$ is biased if $\mathbb{E}[g] \neq 0$). Using standard sampling methods we convert this to approximation using only a few biased $(k-1)$-iterated derivatives (Lemma 11.4). The final step is approximating each biased $(k-1)$-iterated derivative by a small number of its derivatives (which are k-iterated derivatives of f). To this end we prove a general lemma showing that any biased function can be approximated in a concise manner by an algorithm having oracle access to a small number of its derivatives (Lemma 11.3).

We now apply the approximation by iterative derivative result we just described to bound the weight distribution of Reed-Muller codes. Fix $\delta = \delta(d)$ to be specified later. The approximation lemma implies that every $\mathrm{RM}(n, d)$ codeword of weight up to $2^{-k}(1-\epsilon)$ can be well approximated by a function of $c = c(k, \epsilon, \delta)$ of its k-iterated derivatives. Now we use the minimal distance of Reed-Muller codes. Any two distinct codewords $f', f'' \in \mathrm{RM}(n, d)$ have distance at least 2^{-d}. Thus, if we have a good enough approximation of f' (that is, for $\delta < 2^{-(d+1)}$), then such an approximation determines f' uniquely. Hence, to upper bound the number of Reed-Muller codewords it is enough to upper bound the number of δ-approximations for these codewords.

Using the approximation result we obtained, we get that the number of $\mathrm{RM}(n, d)$ codewords up to weight $2^{-k}(1-\epsilon)$, is bounded by the number of possible distinct inputs for the approximation function $\mathcal{A}$: the set of directions $\{y_{i,j}\}$ and the directional derivatives functions $f_{y_{1,1},\ldots,y_{1,k}}(x), \ldots, f_{y_{c,1},\ldots,y_{c,k}}(x)$. Each direction $y_{i,j}$ is an element of $\mathbb{F}_2^n$, hence has 2^n possible values; each k-iterated derivative is an n-variate polynomial of total degree at most $d-k$, hence has at most $2^{O(n^{d-k})}$ possible values. Thus, we get that the number of Reed-Muller codewords of weight up to $2^{-k}(1-\epsilon)$ can be bounded by

$$A(2^{-k}(1-\epsilon)) \leq 2^{n \cdot kc + O(n^{d-k}) \cdot c}.$$

Combining these with the estimates we get for c we get the required upper bound. We complement these upper bound estimations with matching lower bounds. The bounds on the list-decoding size of Reed-Muller codes are obtained using similar techniques.

A similar work along the same lines is the work of Bogdanov and Viola [BV07], which shows that a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ whose weight is bounded by $\mathrm{wt}(f) \leq 1/2 - \epsilon$ can

be well approximated by $c = c(k, \epsilon)$ of its 1^{st}-derivatives. Note that approximation by 1^{st}-derivatives *does not* imply in general approximation by k-iterated derivatives which is crucial for obtaining our bounds here.

11.1.6 Generalized Reed-Muller Codes

The problems of bounding both the accumulative weight distribution and the list-decoding size can be extended to Generalized Reed-Muller codes, the family of low-degree polynomials over larger fields. However, our techniques fail to prove tight results in these cases, as they do for Reed–Muller codes. We provide in Section 11.4 some partial results for this case and make a conjecture about the correct bounds.

11.1.7 Organization

The paper is organized as follows. In Section 11.2 we prove the main technical lemma, showing that a low-weight function can be approximated by its iterated derivatives. We then apply this lemma to bounding the weight distribution and list-decoding size of Reed-Muller codes in Section 11.3. We study the extension of our techniques for Generalized Reed-Muller codes in Section 11.4, where we provide some (non tight) bounds for these codes.

11.2 Approximation of biased functions by derivatives

We prove in this section the main technical lemma we use for bounding the weight distribution and list-decoding size of Reed–Muller codes. We require some definitions before stating it.

Definition 11.5 (Discrete derivatives)**.** *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ by a function. The discrete derivative of f in direction $a \in \mathbb{F}_2^n$ is defined as*

$$f_a(x) = f(x + a) + f(x).$$

The k-iterated discrete derivative of f in directions $a_1, ..., a_k \in \mathbb{F}_2^n$ is defined as

$$f_{a_1,...,a_k}(x) = (...((f_{a_1})_{a_2})...)_{a_k}(x) = \sum_{S \subseteq [k]} f(x + \sum_{i \in S} a_i)$$

We note that usually derivatives are defined as $f_a(x) = f(x+a) - f(x)$, but since we are working over $\mathbb{F}_2$, we can ignore signs. Another notion central to our proof is that of a bias of a function.

Definition 11.6 (Bias)**.** *The bias of a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is*

$$\begin{aligned}\mathrm{bias}(f) =& \mathbb{E}_{x \in \mathbb{F}_2^n}[(-1)^{f(x)}] = \\ & \mathbb{P}[f = 0] - \mathbb{P}[f = 1] = \\ & 1 - 2\mathrm{wt}(f)\end{aligned}$$

Our main lemma states that if f is a function with small weight, then it can be approximated by an algorithm having oracle access to a small number of its iterated derivatives. In the following when we assume an algorithm $\mathcal{A}$ receives as input a function $g(\cdot)$, we mean $\mathcal{A}$ has the ability to evaluate g on any input. One example is if $\mathcal{A}$ receives a representation of g in some canonical form (when g is a polynomial, $\mathcal{A}$ receives as input its list of coefficients).

Lemma 11.1. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that $\mathrm{wt}(f) \leq 2^{-k}(1-\epsilon)$. For every error parameter $\delta > 0$ there exists a universal algorithm $\mathcal{A}$ (that is, independent of f) with the following properties. $\mathcal{A}$ has two types of inputs. The first is an input $x \in \mathbb{F}_2^n$ on which $\mathcal{A}$ is required to guess $f(x)$. The second input is a family of $t = O(\log(1/\epsilon\delta) \cdot \log(1/\delta))$ sets of k directions $\{y_{i,j} \in \mathbb{F}_2^n : 1 \leq i \leq t, 1 \leq j \leq k\}$ and their corresponding k-iterated derivatives of f, $\{f_{y_{i,1},\ldots,y_{i,k}}(\cdot) : 1 \leq i \leq t\}$. For every function f there exists a set of directions $\{y_{i,j}\}$ such that*

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) \neq \mathcal{A}(x; \{y_{i,j} : 1 \leq i \leq t, 1 \leq j \leq k\}, \{f_{y_{i,1},\ldots,y_{i,k}}(\cdot) : 1 \leq i \leq t\})] \leq \delta.$$

Our starting point in the proof of Lemma 11.1 is the following lemma, which states that if a function f has weight less than $2^{-k}(1-\epsilon)$, then it can be computed exactly by a its iterated $(k-1)$-derivatives, and moreover each of theses derivatives is biased.

Lemma 11.2. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that $\mathrm{wt}(f) \leq 2^{-k}(1-\epsilon)$ for integer $k \geq 2$. Then the function $(-1)^{f(x)} : \mathbb{F}_2^n \to \{-1, 1\}$ can be written as*

$$(-1)^{f(x)} = \mathbb{E}_{a_1,\ldots,a_{k-1} \in \mathbb{F}_2^n}[\alpha_{a_1,\ldots,a_{k-1}}(-1)^{f_{a_1,\ldots,a_{k-1}}(x)}]$$

where

1. *The coefficients $\alpha_{a_1,\ldots,a_{k-1}}$ are real numbers of absolute value at most 10.*
2. *All the functions $f_{a_1,\ldots,a_{k-1}}$ are biased, $\mathrm{bias}(f_{a_1,\ldots,a_{k-1}}) \geq \epsilon$.*

We prove Lemma 11.2 in Subsection 11.2.1. The second lemma shows that biased functions can be approximated using a small number of their derivatives.

Lemma 11.3. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that $\mathrm{bias}(f) \geq \epsilon$. For every error parameter $\delta > 0$ there exists a universal algorithm $\mathcal{A}'$ (that is, independent of f) with the following properties. $\mathcal{A}'$ has two types of inputs. The first is an input $x \in \mathbb{F}_2^n$ on which $\mathcal{A}'$ is required to guess $f(x)$. The second input is a set of $t = \log(1/\epsilon\delta)+1$ directions $y_1, \ldots, y_t \in \mathbb{F}_2^n$ and the directional derivatives of f in these directions $f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)$. For every function f there exists a set of directions $y_1, \ldots, y_t$ such that*

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) \neq \mathcal{A}'(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot))] \leq \delta.$$

We prove Lemma 11.3 in Subsection 11.2.2. The last ingredient required for the proof of Lemma 11.1 is a standard sampling lemma showing how to transform exact computation by averaging many functions, to approximation by averaging few functions.

Lemma 11.4. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function, $H = \{h_1, ..., h_t\}$ a set of functions from $\mathbb{F}_2^n$ to $\mathbb{F}_2$, such that there exist constants $c_{h_1}, ..., c_{h_t}$ of absolute value at most C, such that*

$$(-1)^{f(x)} = \mathbb{E}_{i\in[t]}[c_{h_i}(-1)^{h_i(x)}] \qquad (\forall x \in \mathbb{F}_2^n)$$

Then f can be approximated by a small number of the functions $h_1, ..., h_t$. For any error parameter $\delta > 0$, there exist functions $h_1, ..., h_\ell \in H$ for $\ell = O(C^2 \log 1/\delta)$, and a function $F : \mathbb{F}_2^\ell \to \mathbb{F}_2$, such that the relative distance between $f(x)$ and $F(h_1(x), ..., h_\ell(x))$ is at most δ, i.e.

$$\mathbb{P}_{x\in\mathbb{F}_2^n}[f(x) \neq F(h_1(x), ..., h_\ell(x))] \leq \delta$$

The function F is a weighted majority, i.e. it is of the form

$$F(h_1(x), ..., h_\ell(x)) = \text{sign}(\sum_{i=1}^{\ell} s_i(-1)^{h_i(x)}).$$

Moreover, we can have $s_1, ..., s_\ell$ to be integers of absolute value at most $C + 1$.

We prove Lemma 11.4 in Subsection 11.2.3. We now prove Lemma 11.1 using Lemmas 11.2, 11.3 and 11.4.

Proof of Lemma 11.1. Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that $\text{wt}(f) \leq 2^{-k}(1 - \epsilon)$, and let $\delta > 0$ be an error parameter. We start by defining an algorithm $\mathcal{A}_1(x)$ approximating $f(x)$ using a small number of its $(k-1)$-iterated derivatives. If $k = 1$ simply set $\mathcal{A}_1(x) = f(x)$. For $k \geq 2$ apply Lemma 11.2 to get that $(-1)^{f(x)}$ can be exactly computed as

$$\mathbb{E}_{a_1,...,a_{k-1}\in\mathbb{F}_2^n}[\alpha_{a_1,...,a_{k-1}}(-1)^{f_{a_1,...,a_{k-1}}(x)}]$$

where $|\alpha_{a_1,...,a_k}| \leq 10$ and $\text{bias}(f_{a_1,...,a_{k-1}}(x)) \geq \epsilon$. Applying Lemma 11.4 we get that f can be approximated by a small number of its $(k-1)$-iterated derivatives,

$$\Pr_{x\in\mathbb{F}_2^n}[\text{Maj}(f_{a_{1,1},...,a_{1,k-1}}(x), \ldots, f_{a_{\ell,1},...,a_{\ell,k-1}}(x)) \neq f(x)] \leq \delta/2$$

where $\ell = O(\log 1/\delta)$. Define

$$\mathcal{A}_1(x) = \text{Maj}(f_{a_{1,1},...,a_{1,k-1}}(x), \ldots, f_{a_{\ell,1},...,a_{\ell,k-1}}(x)).$$

We now approximate each $(k-1)$-iterated derivative by a small number of its derivatives. We will use Lemma 11.3 to this end. Notice this can be done since by Lemma 11.2 all $(k-1)$-iterated derivatives $f_{a_{i,1},...,a_{i,k-1}}$ have bias of at least ϵ (and in the $k = 1$ case, $\text{bias}(f) \geq \epsilon$). Thus, for each $1 \leq i \leq \ell$ there exists $t = O(\log(1/\epsilon\delta))$ directions $y_{i,1}, \ldots, y_{i,t}$ such that

$$\Pr_{x\in\mathbb{F}_2^n}[f_{a_{i,1},...,a_{i,k-1}}(x) \neq \mathcal{A}'(x; y_{i,1}, \ldots, y_{i,k}, f_{a_{i,1},...,a_{i,k-1},y_{i,1}}(\cdot), \ldots, f_{a_{i,1},...,a_{i,k-1},y_{i,t}}(\cdot))] \leq \delta/(2\ell).$$

Plugging all these into $\mathcal{A}_1$, we get an algorithm $\mathcal{A}$ such that

$$\Pr_{x\in\mathbb{F}_2^n}[f(x) \neq A(x; \{y_{i,j} : 1 \leq i \leq \ell, 1 \leq j \leq t\}, \\ \{f_{a_{i,1},\ldots,a_{i,k-1},y_{i,j}} : 1 \leq i \leq \ell, 1 \leq j \leq t\})] \leq \delta.$$

In total, $\mathcal{A}$ has as input $\ell \cdot t = O(\log(1/\epsilon\delta) \cdot \log(1/\delta))$ k-iterated derivatives of f, and (a subset) of the directions of these derivatives. □

11.2.1 Proof of Lemma 11.2

Before proving Lemma 11.2, we need some claims regarding derivatives. The first claim shows that if a function has non-zero bias, it can be computed by an average of its derivatives.

Claim 11.1. *Let $g : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that* $\mathrm{bias}(g) \neq 0$. *Then*

$$(-1)^{g(x)} = \frac{1}{\mathrm{bias}(g)}\mathbb{E}_{a\in\mathbb{F}_2^n}[(-1)^{g_a(x)}]$$

where the identity holds for any $x \in \mathbb{F}_2^n$.

Proof. Fix x. We have

$$(-1)^{g(x)}\mathbb{E}_{a\in\mathbb{F}_2^n}[(-1)^{g_a(x)}] = \\ \mathbb{E}_{a\in\mathbb{F}_2^n}[(-1)^{g(x)-g_a(x)}] = \\ \mathbb{E}_{a\in\mathbb{F}_2^n}[(-1)^{g(x+a)}] = \mathrm{bias}(g)$$

□

The following claim shows that if a function has low weight, then derivatives of it will also have low weight, and thus large bias.

Claim 11.2. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that* $\mathrm{wt}(f) \leq 2^{-k}(1-\epsilon)$. *Let $a_1, \ldots, a_s \in \mathbb{F}_2^n$ for $1 \leq s \leq k-1$ be any derivatives, and consider* $\mathrm{bias}(f_{a_1,\ldots,a_s})$. *Then* $\mathrm{bias}(f_{a_1,\ldots,a_s}) \geq 1 - 2^{s+1-k}(1-\epsilon)$. *In particular*

1. *If $s < k-1$ then* $\mathrm{bias}(f_{a_1,\ldots,a_s}) \geq 1 - 2^{s+1-k}$.
2. *If $s = k-1$ then* $\mathrm{bias}(f_{a_1,\ldots,a_s}) \geq \epsilon$.

Proof. Consider $f_{a_1,\ldots,a_s}$

$$f_{a_1,\ldots,a_s} = \sum_{I\subseteq[s]} f(x + \sum_{i\in I} a_i)$$

For random x, the probability that $f(x+\sum_{i\in I} a_i) = 1$ is $\mathrm{wt}(f)$, which is at most $2^{-k}(1-\epsilon)$. Thus by the union bound,

$$\mathbb{P}_{x\in\mathbb{F}_2^n}[\exists I \subseteq [s],\ f(x + \sum_{i\in I} a_i) = 1] \leq 2^{s-k}(1-\epsilon)$$

In particular it implies that

$$\mathrm{wt}(f_{a_1,\ldots,a_s}) = \mathbb{P}_{x\in\mathbb{F}_2^n}[f_{a_1,\ldots,a_s}(x) = 1] \leq 2^{s-k}(1-\epsilon)$$

and we get the bound since $\mathrm{bias}(f_{a_1,\ldots,a_s}) = 1 - 2\mathrm{wt}(f_{a_1,\ldots,a_s})$. □

We now can prove Lemma 11.2 using Claims 11.1 and 11.2.

Proof of Lemma 11.2. Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function such that $\mathrm{wt}(f) \leq 2^{-k}(1-\epsilon)$. Thus $\mathrm{bias}(f) = 1 - 2\mathrm{wt}(f) > 0$ and by Claim 11.1 we can write

$$(-1)^{f(x)} = \frac{1}{\mathrm{bias}(f)}\mathbb{E}_{a_1\in\mathbb{F}_2^n}[(-1)^{f_{a_1}(x)}]$$

If $k = 1$ we are done. Otherwise by Claim 11.2, f_{a_1} also has positive bias,

$$\mathrm{bias}(f_{a_1}) \geq 1 - 2^{s+1-k}(1-\epsilon) > 0$$

and so again by Claim 11.1 we can write

$$(-1)^{f_{a_1}(x)} = \frac{1}{\mathrm{bias}(f_{a_1})}\mathbb{E}_{a_2\in\mathbb{F}_2^n}[(-1)^{f_{a_1,a_2}(x)}]$$

Thus we have

$$(-1)^{f(x)} = \frac{1}{\mathrm{bias}(f)}\mathbb{E}_{a_1\in\mathbb{F}_2^n}[\frac{1}{\mathrm{bias}(f_{a_1})}\mathbb{E}_{a_2\in\mathbb{F}_2^n}[(-1)^{f_{a_1,a_2}(x)}]]$$

We can continue this process as long as we can guarantee that $f_{a_1,\ldots,a_s}$ has non-zero bias for all $a_1,\ldots,a_s \in \mathbb{F}_2^n$. By Claim 11.2 we know this happens for $s \leq k-1$, and thus we have

$$(-1)^{f(x)} = \mathbb{E}_{a_1,\ldots,a_{k-1}\in\mathbb{F}_2^n}[\alpha_{a_1,\ldots,a_{k-1}}(-1)^{f_{a_1,\ldots,a_{k-1}}(x)}]$$

where

$$\alpha_{a_1,\ldots,a_k} = \frac{1}{\mathrm{bias}(f)}\frac{1}{\mathrm{bias}(f_{a_1})}\frac{1}{\mathrm{bias}(f_{a_1,a_2})}\cdots\frac{1}{\mathrm{bias}(f_{a_1,\ldots,a_{k-2}})}.$$

By Claim 11.2 we know that $\mathrm{bias}(f_{a_1,\ldots,a_{k-1}}) \geq \epsilon$ for all $(k-1)$-iterated derivatives. We now bound $\alpha_{a_1,\ldots,a_k}$. By Claim 11.2 we get that

$$1 \leq \alpha_{a_1,\ldots,a_k} \leq \prod_{s=0}^{k-2}\frac{1}{1-2^{s-k+1}} \leq \prod_{r\geq 1}\frac{1}{1-2^{-r}} \leq 10.$$

□

11.2.2 Proof of Lemma 11.3.

For a set of directions $y_1, \ldots, y_t \in \mathbb{F}_2^n$ and a subset $I \subseteq [t]$, define $y_I = \sum_{i \in I} y_i$. We start by showing that if we know the directions $y_1, \ldots, y_t$ and the directional derivatives of f in these directions $f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)$, then we can also compute all the derivatives in directions y_I, that is the functions $f_{y_I}(\cdot)$.

Claim 11.3. *Let $y_1, \ldots, y_t \in \mathbb{F}_2^n$ a set of directions, and $f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)$ the directional derivatives of a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$. For every non-empty $I \subseteq [t]$ there exists an algorithm $\mathcal{A}_I$ such that*

$$\mathcal{A}_I(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)) = f_{y_I}(x)$$

for all $x \in \mathbb{F}_2^n$.

Proof. Let $I = \{i_1, \ldots, i_r\}$. The algorithm $\mathcal{A}_I$ calculates

$$\mathcal{A}_I(x) = \sum_{a=1}^{r} f_{y_{i_a}}(x + \sum_{b=1}^{a-1} y_{i_b}).$$

It is straightforward to verify that $\mathcal{A}_I(x) = f_{y_I}(x)$ for all $x \in \mathbb{F}_2^n$. □

We turn to prove Lemma 11.3.

Proof of Lemma 11.3. Define the algorithm $\mathcal{A}'$ as follows. For a set of directions $y_1, \ldots, y_t \in \mathbb{F}_2^n$ and the directional derivatives of $f : \mathbb{F}_2^n \to \mathbb{F}_2$ in these directions $f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)$, define $\mathcal{A}'(x)$ to be the majority vote of $f_{y_I}(x)$, which according to Claim 11.3 can be computed by algorithms receiving $x, y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)$, that is,

$$\begin{aligned}
&\mathcal{A}'(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)) = \\
&\mathrm{Maj}\left\{\mathcal{A}_I(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot))\right\}_{\emptyset \neq I \subseteq [t]} = \\
&\mathrm{Maj}\left\{f_{y_I}(x)\right\}_{\emptyset \neq I \subseteq [t]}.
\end{aligned}$$

We will prove that there is a choice of $y_1, \ldots, y_t$ for which $\mathcal{A}'(x) = f(x)$ for almost all x. In fact, we will prove this occurs for a random choice of $y_1, \ldots, y_t$. First, we claim that $\mathcal{A}'(x) = f(x)$ iff

$$S = \sum_{\emptyset \neq I \subseteq [t]} (-1)^{f(x+y_I)} > 0.$$

This is because $f_{y_I}(x) = f(x)$ iff $f(x + y_I) = 0$. Having the majority of $f_{y_I}(x)$ being equal to $f(x)$ is equivalent to $S > 0$ (note we cannot have $S = 0$ as S is the sum of an odd number of $\{-1, 1\}$ summands). Let $x, y_1, \ldots, y_t \in \mathbb{F}_2^n$ be chosen uniformly and independently. We prove $S > 0$ with high probability using Markov's inequality. First we compute $\mathbb{E}[S]$.

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{\emptyset \neq I \subseteq [t]} (-1)^{f(x+y_I)}\right] = (2^t - 1)\mathrm{bias}(f).$$

To bound $\mathrm{Var}[S]$ we observe that the different summands in S are pairwise independent. This is because for distinct $I, J \subseteq [t]$ we have

$$\begin{aligned}
&\mathbb{E}[(-1)^{f(x+y_I)} \cdot (-1)^{f(x+y_I)}] = \\
&\mathbb{E}[(-1)^{f(x+y_I)+f(x+y_J)}] = \\
&\mathbb{E}[(-1)^{f(x+y_I)}] \cdot \mathbb{E}[(-1)^{f(x+y_J)}] = \\
&bias(f)^2,
\end{aligned}$$

where we used the fact that the two points $x + y_I$ and $x + y_J$ are uniform and independent given that $x, y_1, \ldots, y_t$ are chosen uniformly and independently. We thus conclude that

$$\begin{aligned}
\mathrm{Var}[S] &= \sum_{\emptyset \neq I \subseteq [t]} \mathrm{Var}[(-1)^{f(x+y_I)}] \\
&= (2^t - 1)\mathrm{Var}[(-1)^{f(x)}] \leq 2^t - 1.
\end{aligned}$$

Hence we conclude that

$$\begin{aligned}
\Pr[S \leq 0] &\leq \Pr[|S - \mathbb{E}[S]| \geq (2^t - 1)\mathrm{bias}(f)] \\
&\leq \frac{bias(f)}{2^t - 1}.
\end{aligned}$$

Thus, for $t = \log(1/\epsilon\delta) + 1$ we get that

$$\Pr[S \leq 0] \leq \delta,$$

Hence we get that for uniformly chosen $x, y_1, \ldots, y_t$,

$$\begin{aligned}
\Pr_{x, y_1, \ldots, y_t \in \mathbb{F}_2^n}[&\mathcal{A}'(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)) \\
&\neq f(x)] \leq \delta.
\end{aligned}$$

By an averaging argument, for every f there must exist a choice for $y_1, \ldots, y_t$ where

$$\Pr_{x \in \mathbb{F}_2^n}[\mathcal{A}'(x; y_1, \ldots, y_t, f_{y_1}(\cdot), \ldots, f_{y_t}(\cdot)) \neq f(x)] \leq \delta.$$

□

11.2.3 Proof of Lemma 11.4

The proof of Lemma 11.4 is based on a standard sampling argument.

Proof of Lemma 11.4. Choose $h_1, ..., h_\ell$ uniformly and independently from H. Fix $x \in \mathbb{F}_2^n$, and let Z_i be the random variable

$$Z_i = c_{h_i}(-1)^{h_i(x)}$$

and let $S = \frac{Z_1+...+Z_\ell}{\ell}$. We will use the fact that if $|S - (-1)^{f(x)}| < 1$ then $\text{sign}(S) = (-1)^{f(x)}$.

We first bound the probability that

$$|S - (-1)^{f(x)}| > 1/4$$

By regular Chernoff arguments for bounded independent variables, since $\mathbb{E}[S] = (-1)^{f(x)}$ and each Z_i is of absolute value of at most C, we get that

$$\mathbb{P}_{h_1,...,h_\ell \in H}[|S - (-1)^{f(x)}| > 1/4] \leq e^{-\frac{\ell}{32C^2}}$$

(see for example Theorem A.1.16 in [AS00]).

In particular for $\ell = O(C^2 \log 1/\delta)$ we get that

$$\mathbb{P}_{h_1,...,h_\ell \in H}[|S - (-1)^{f(x)}| > 1/4] \leq \delta$$

Thus by averaging arguments, there exists $h_1, ..., h_\ell$ such that

$$\mathbb{P}_{x \in \mathbb{F}_2^n}[|\frac{\sum_{i=1}^{\ell} c_{h_i}(-1)^{h_i(x)}}{\ell} - (-1)^{f(x)}| \geq 1/4] \leq \delta$$

We now round each coefficient to a close rational, without damaging the approximation error. The coefficient of $(-1)^{h_i(x)}$ is $\alpha_i = \frac{c_{h_i}}{\ell}$. If we round c_{h_i} to the closest integer $[c_{h_i}]$, we get that the coefficient of each $(-1)^{h_i(x)}$ is changed by at most $\frac{1}{2\ell}$, and thus the total approximation is changed by at most $1/2$. Hence we have

$$\mathbb{P}_{x \in \mathbb{F}_2^n}[|\frac{\sum_{i=1}^{\ell} [c_{h_i}](-1)^{h_i(x)}}{\ell}) - (-1)^{f(x)}| \geq 3/4] \leq \delta.$$

Thus we got that

$$\mathbb{P}_{x \in \mathbb{F}_2^n}[\text{sign}(\frac{\sum_{i=1}^{\ell} [c_{h_i}](-1)^{h_i(x)}}{\ell}) \neq (-1)^{f(x)}] \leq \delta.$$

Since dividing by ℓ does not change the sign we get

$$\mathbb{P}_{x \in \mathbb{F}_2^n}[\text{sign}(\sum_{i=1}^{\ell} [c_{h_i}](-1)^{h_i(x)}) \neq (-1)^{f(x)}] \leq \delta$$

□

11.3 Bounds for Reed-Muller codes

In this section we study the weight distribution and list-decoding size of Reed–Muller codes. Recall that $\text{RM}(n, d)$ denotes the code of multivariate polynomials $p(x_1, ..., x_n)$ over $\mathbb{F}_2$ of total degree at most d. In the following n and d will always stand for the number of variables and the total degree. We assume that $d \ll n$, and study in particular the case of constant d.

11.3.1 Weight distribution of Reed-Muller codes

We prove in this subsection our first main theorem, Theorem 11.3, which gives the asymptotic behavior of the weight distribution of Reed-Muller codes. It is a direct corollary of Theorem 11.6, giving an upper bound on the accumulative weight at distance $2^{\ell-d}-\epsilon$, and Lemma 11.5, giving a simple lower bound at distance $2^{\ell-d-1}$.

Theorem 11.6 (Upper bound on the accumulative weight)**.** *For any integer* $1 \leq k \leq d-1$*,*

$$A(2^{-k}(1-\epsilon)) \leq (1/\epsilon)^{O(\frac{d^2}{(d-k)!}n^{d-k})}.$$

In particular for constant d *we get that*

$$A(2^{-k}-\epsilon) \leq (1/\epsilon)^{O(n^{d-k})}.$$

Lemma 11.5 (Lower bound on the accumulative weight)**.** *For any integer* $1 \leq k \leq d$

$$A(2^{-k}) \geq 2^{\frac{n^{d-k+1}}{(d-k+1)!}(1+o(1))}.$$

In particular for constant d *we get that*

$$A(2^{-k}) \geq 2^{\Omega(n^{d-k+1})}.$$

We start by proving the lower bound.

Proof of Lemma 11.5. Single out k variables $x_1, ..., x_k$, and let q be any degree $d-k+1$ polynomials on the remaining $n-k$ variables. First, for any such q, the following degree d polynomial has relative weight exactly 2^{-k}

$$q'(x_1, ..., x_n) = x_1 x_2 ... x_{k-1}(x_k + q(x_{k+1}, ..., x_n))$$

The number of different polynomials q is

$$2^{\binom{n-k}{d-k+1}} = 2^{\frac{n^{d-k+1}}{(d-k+1)!}(1+o(1))}$$

□

We prove Theorem 11.6 using Lemma 11.1.

Proof of Theorem 11.6. Fix $1 \leq k \leq d-1$. We will bound the number of polynomials $p \in \mathrm{RM}(n,d)$ such that $\mathrm{wt}(p) \leq 2^{-k}(1-\epsilon)$. Let p be any such polynomial. Apply Lemma 11.1 to $p(x)$ with error parameter $\delta = 2^{-(d+2)}$. There exists a universal algorithm $\mathcal{A}$, and for each p a set of $t = O(d^2 + d\log(1/\epsilon))$ directions $\{y_{i,j} : 1 \leq i \leq t, 1 \leq j \leq k\}$ such that

$$\Pr_{x \in \mathbb{F}_2^n}[p(x) \neq \mathcal{A}(x; \{y_{i,j} : 1 \leq i \leq t, 1 \leq j \leq k\}, \{p_{y_{i,1},...,y_{i,k}}(\cdot) : 1 \leq i \leq t\})] \leq \delta.$$

Define $p'(x) = \mathcal{A}\left(x; \{y_{i,j}\}, \{p_{y_{i,1},\ldots,y_{i,k}}(\cdot)\}\right)$. We have that $\mathrm{dist}(p, p') = \Pr_x[p(x) \neq p'(x)] \leq \delta$. We claim that this guarantees that $p'(x)$ specifies $p(x)$ uniquely - it is the only element of $\mathrm{RM}(n, d)$ of distance at most δ from p'. This is because the minimal distance of $\mathrm{RM}(n, d)$ is 2^{-d}, and we chose δ to be less than half the minimal distance. Now, in order to compute $p'(x)$, we need to specify to the algorithm $\mathcal{A}$ the set of vectors $y_{i,j}$ and the polynomials $p_{y_{i,1},\ldots,y_{i,k}}(\cdot)$. To specify each vector $y_{i,j} \in \mathbb{F}_2^n$ we require n bits. Each polynomial $p_{y_{i,1},\ldots,y_{i,k}}(\cdot)$ is a k-iterated derivative of a degree-d polynomial $p(x)$, hence it is a degree $d-k$ polynomial. Thus, in order to specify it, we need to give the list of its $\sum_{i=0}^{d-k}\binom{n}{i}$ bits. Summing up, we need a total of

$$tk \cdot n + t \cdot \sum_{i=0}^{d-k} \binom{n}{i} = O\left(d^2 \log(1/\epsilon) \cdot \frac{n^{d-k}}{(d-k)!}\right)$$

bits in order to specify p' completely. Since each p' approximates at most a single p we get that the number of polynomials $p \in \mathrm{RM}(n, d)$ such that $\mathrm{wt}(p) \leq 2^{-k}(1-\epsilon)$ is bounded by the number of distinct p', which is bounded by

$$(1/\epsilon)^{O\left(\frac{d^2}{(d-k)!} n^{d-k}\right)}.$$

□

11.3.2 List-decoding size of Reed-Muller codes

We now turn to the problem of bounding the list-decoding size of Reed-Muller codes, and we prove our second main theorem, Theorem 11.4. We will show that the same techniques used to bound the weight distribution when proving Theorem 11.3 can be applied with minor variants to also bound the list-decoding size. We note this is an exception; commonly bounding the list-decoding size is a much harder task than bounding the weight distribution, and there exist codes where these two parameters behave very differently. However, we will see that in the case of Reed–Muller codes they share the same asymptotic behavior.

Theorem 11.4 giving the list-decoding size of Reed-Muller codes is a direct corollary of Theorem 11.7, giving an upper bound on the list-decoding size at distance $2^{\ell-d} - \epsilon$, and the same lower bound we used to bound the accumulative weight distribution, obtained in Lemma 11.5.

Theorem 11.7 (Upper bound on the list-decoding size)**.** *For any integer* $1 \leq k \leq d-1$*,*

$$L(2^{-k}(1-\epsilon)) \leq (1/\epsilon)^{O(\frac{d^2}{(d-k)!} n^{d-k})}.$$

In particular for constant d *we get that*

$$L(2^{-k} - \epsilon) \leq (1/\epsilon)^{O(n^{d-k})}.$$

Proof of Theorem 11.7. The proof follows the same lines as that of Theorem 11.6. Fix $f : \mathbb{F}_2^n \to \mathbb{F}_2$ to be any function. We will bound the number of polynomials $p \in \mathrm{RM}(n, d)$ such that $\mathrm{dist}(p, f) \leq 2^{-k}(1-\epsilon)$. Let $g = p + f$ such that $\mathrm{wt}(g) \leq 2^{-k}(1-\epsilon)$. Applying

Lemma 11.1 to $g(x)$ with the error parameter $\delta = 2^{-(d+2)}$, there exists a universal algorithm $\mathcal{A}$ and a set of direction $\{y_{i,j} : 1 \le i \le t, 1 \le j \le k\}$ such that

$$\Pr_{x \in \mathbb{F}_2^n}[g(x) \neq \mathcal{A}(x; \{y_{i,j} : 1 \le i \le t, 1 \le j \le k\}, \{g_{y_{i,1},\ldots,y_{i,k}}(\cdot) : 1 \le i \le t\})] \le \delta.$$

Since $g(x) = p(x) + f(x)$ we also have $g_{y_{i,1},\ldots,y_{i,k}}(\cdot) = p_{y_{i,1},\ldots,y_{i,k}}(\cdot) + f_{y_{i,1},\ldots,y_{i,k}}(\cdot)$. Hence, we can replace each instance of g or its derivatives in $\mathcal{A}$ with instances of p, f and their derivatives. Thus we get that

$$\Pr_{x \in \mathbb{F}_2^n}[p(x) \neq f(x) + \mathcal{A}(x; \{y_{i,j}\}, \{p_{y_{i,1},\ldots,y_{i,k}}(\cdot) + f_{y_{i,1},\ldots,y_{i,k}}(\cdot)\})] \le \delta.$$

Define $p'(x) = f(x) + \mathcal{A}(x; \{y_{i,j}\}, \{p_{y_{i,1},\ldots,y_{i,k}}(\cdot) + f_{y_{i,1},\ldots,y_{i,k}}(\cdot)\})$. Since we again have $\mathrm{dist}(p, p') \le \delta$, the function $p'(x)$ specifies $p(x)$ uniquely as the only element in $\mathrm{RM}(n, d)$ which has distance at most δ from p'. Now, in order to compute p', we may assume the algorithm $\mathcal{A}$ has oracle access to the function $f(\cdot)$, since we have fixed it in advance, and it is the same for all the polynomials we wish to bound. Thus, in order to calculate $p'(x)$, we need to provide to the algorithm $\mathcal{A}$ the set of directions $y_{i,j}$ and the polynomials $p_{y_{i,1},\ldots,y_{i,k}}(\cdot)$. Notice that $\mathcal{A}$ can compute $f_{y_{i,1},\ldots,y_{i,k}}(\cdot)$ using the oracle access to f and the set of directions $y_{i,j}$. As in the proof of Theorem 11.6, each direction $y_{i,j} \in \mathbb{F}_2^n$ requires n bits, and each polynomial $p_{y_{i,1},\ldots,y_{i,k}}(\cdot)$ being a degree $d-k$ polynomial requires $\sum_{i=0}^{d-k} \binom{n}{i}$ bits to specify. Following the same calculations as those in the proof of Theorem 11.6, we conclude that the number of distinct $p'(x)$ is bounded by

$$(1/\epsilon)^{O(\frac{d^2}{(d-k)!} n^{d-k})}.$$

Thus, for every fixed function f, this is also a bound on the number of $p \in \mathrm{RM}(n, d)$ such that $\mathrm{dist}(p, f) \le 2^{-k}(1 - \epsilon)$. □

11.4 Generalized Reed-Muller codes

The problems of bounding both the accumulative weight distribution and the list-decoding size can be extended to Generalized Reed-Muller, the code of low-degree polynomials over larger fields. However, our techniques fail to prove tight result in these cases. We briefly describe the reasons below, and give some partial results.

We start by making some basic definitions. Let q be a prime, and let $\mathrm{GRM}_q(n, d)$ denote the code of multivariate polynomials $p(x_1, ..., x_n)$ over the field $\mathbb{F}_q$, of total degree at most d.

Definition 11.7. *The relative weight of a function $f : \mathbb{F}_q^n \to \mathbb{F}_q$ is the fraction of non-zero elements,*

$$\mathrm{wt}(f) = \frac{1}{q^n} |\{x \in \mathbb{F}_q^n : f(x) \neq 0\}|$$

Definition 11.8. *The relative distance between two functions $f, g : \mathbb{F}_q^n \to \mathbb{F}_q$ is defined as*

$$\mathrm{dist}(f, g) = \mathbb{P}_{x \in \mathbb{F}_q^n}[f(x) \neq g(x)]$$

The accumulative weight distribution and the list-decoding size are defined analogously for $\mathrm{GRM}_q(n, d)$, using the appropriate definitions for relative weight and relative distance. We denote them by A_q and L_q. For each $1 \leq k \leq d$, we define a distance r_k as follows.

1. For $k = 1$, let $d = (q-1)a + b$, where $1 \leq b \leq q-1$. Define $r_1 = q^{-a}(1 - b/q)$. r_1 is the minimal distance of $\mathrm{GRM}_q(n, d)$.
2. For $2 \leq k \leq d-1$, let $d - k = (q-1)a + b$, where $1 \leq b \leq q-1$. Define $r_k = q^{-a}(1 - b/q)(1 - 1/q)$.
3. For $k = d$, define $r_d = 1 - 1/q$.

We conjecture that both for the accumulative weight distribution and the list-decoding size, the distances r_k are the thresholds for the exponential dependency in n.

Conjecture 11.1. *Let $\epsilon > 0$ be constant, and consider $\mathrm{GRM}_q(n, d)$ for constant d. Then*

- *For $\alpha \leq r_1 - \epsilon$ both $A_q(\alpha)$ and $L_q(\alpha)$ are constants.*
- *For $r_k \leq \alpha \leq r_{k+1} - \epsilon$ both $A_q(\alpha)$ and $L_q(\alpha)$ are $2^{\Theta(n^k)}$.*
- *For $\alpha \geq r_d$ both $A_q(\alpha)$ and $L_q(\alpha)$ are $2^{\Theta(n^d)}$.*

Proving lower bounds for $A_q(r_k)$ is similar to the case of $\mathrm{RM}(n, d)$.

Lemma 11.6 (Lower bound for A_q)**.** *For any integer $1 \leq k \leq d$,*

$$A_q(r_k) \geq 2^{\Omega(n^k)}$$

The problem is proving matching upper bounds. Using directly the derivatives method we used to give upper bounds for $\mathrm{RM}(n, d)$ gives the same bounds for $\mathrm{GRM}_q(n, d)$, alas they are not tight for $q > 2$.

$$A_q(2^{-k} - \epsilon) \leq 2^{O(n^{d-k})}$$

If we would like to get upper bounds closer to the lower bounds, a natural approach would be to generalize Lemma 11.2 to taking several derivatives in the same direction (which is possible over larger fields). This would give tight results for some values of k. The crucial point is generalizing Claim 11.1 to the case of taking multiple derivatives in the same direction. So far, we didn't find a way of doing so.

Instead, we give partial results for Conjecture 11.1 at both ends of the spectrum. We give results when $\alpha \leq r_1 - \epsilon$, and when $r_{d-1} \leq \alpha \leq r_d - \epsilon$ (when $\alpha \geq r_d$ Lemma 11.6 gives $L_q(\alpha)$ and $A_q(\alpha)$ are both exponential in n^d, and this is obviously tight).

First, the minimal distance of $\mathrm{GRM}_q(n, d)$ is known to be r_1. Thus, for any $\epsilon > 0$, $A_q(r_1 - \epsilon) = 1$. Gopalan, Klivans and Zuckerman [GKZ08] prove that $L_q(r_1 - \epsilon)$ is constant when $q - 1$ divides d.

Theorem 11.8 (Corollary 18 in [GKZ08])**.** *Assume $q-1$ divides d. Then*

$$L_q(r_1 - \epsilon) \leq c(q, d, \epsilon)$$

Moving to the case of $r_{d-1} \leq \alpha \leq r_d - \epsilon$, we prove

Lemma 11.7. *Let $\epsilon > 0$ be constant. then*

$$A_q(r_d - \epsilon) \leq 2^{O(n^{d-1})}$$

We now move on to prove Lemmas 11.6 and 11.7. We start with Lemma 11.6.

Proof of Lemma 11.6. We start by proving for $2 \leq k \leq d-1$. Let $d-k = (q-1)a + b$, where $1 \leq b \leq q-1$. Single out $a+2$ variables $x_1, ..., x_{a+2}$, and let g be any degree k polynomial on the remaining variables. The following polynomial has degree d and weight exactly $q^{-a}(1-b/q)(1-1/q)$.

$$g'(x_1, ..., x_n) = \left(\prod_{i=1}^{a}\prod_{j=1}^{q-1}(x_i - j)\right) \cdot \left(\prod_{j=1}^{b}(x_{a+1} - j)\right) \cdot (x_{a+2} + g(x_{a+3}, ..., x_n))$$

The number of distinct polynomial g is $2^{\Omega(n^k)}$.

The proofs for $k = 1$ and $k = d$ are similar: for $k = 1$, let $d = (q-1)a + b$. Let $l_1(x), ..., l_{a+1}(x)$ be any independent linear functions, and consider

$$g'(x_1, ..., x_n) = \left(\prod_{i=1}^{a}\prod_{j=1}^{q-1}(l_i(x) - j)\right)\left(\prod_{j=1}^{b}(l_{a+1}(x) - j)\right)$$

For $k = d$, let g be any degree d polynomial on variables $x_2, ..., x_n$, and consider $g'(x_1, ..., x_n) = x_1 + g(x_2, ..., x_n)$. □

We now continue to prove Lemma 11.7. We first make some necessary definitions.

Definition 11.9. *The bias of a polynomial $p(x_1, ..., x_n)$ over $\mathbb{F}_q$ is defined to be*

$$\text{bias}(p) = \mathbb{E}_{x \in \mathbb{F}_q^n}[\omega^{p(x)}]$$

where $\omega = e^{2\pi i/q}$ is a primitive q-th root of unity.

Kaufman and Lovett [KL08] prove that biased low-degree polynomials can be decomposed into a function of a constant number of lower degree polynomials.

Theorem 11.9 (Theorem 2 in [KL08])**.** *Let $p(x_1, ..., x_n)$ be a degree d polynomial, such that $|\text{bias}(p)| \geq \epsilon$. Then p can be decomposed as a function of a constant number of lower degree polynomials*

$$p(x) = F(g_1(x), ..., g_c(x))$$

where $deg(g_i) \leq d-1$, and $c = c(q, d, \epsilon)$.

We will use Theorem 11.9 to bound $A(r_d - \epsilon)$ for any constant $\epsilon > 0$.

Proof of Lemma 11.7. We will show that any polynomial $p \in \mathrm{GRM}_q(n, d)$ such that $\mathrm{wt}(p) \leq 1 - 1/p - \epsilon$ can be decomposed as

$$p(x) = F(g_1(x), ..., g_c(x))$$

where $deg(g_i) \leq d-1$, and c depends only on q, d and ϵ. Thus the number of such polynomials is bounded by the number of possibilities to choose c degree $d-1$ polynomials, and a function $F : \mathbb{F}_q^c \to \mathbb{F}_q$. The number of such possibilities is at most $2^{O(n^{d-1})}$. Let p be such that $\mathrm{wt}(p) \leq 1 - 1/p - \epsilon$. We will show there exists $\alpha \in \mathbb{F}_q$, $\alpha \neq 0$ such that $\mathrm{bias}(\alpha p) \geq \epsilon$. We will then finish by using Theorem 11.9 on the polynomial αp.

Consider the bias of αp for random $\alpha \in \mathbb{F}_q$.

$$\mathbb{E}_{\alpha \in \mathbb{F}_q}[\mathrm{bias}(\alpha p)] = \mathbb{E}_{\alpha \in F_q, x \in \mathbb{F}_q^n}[\omega^{\alpha p(x)}] = 1 - \mathrm{wt}(p)$$

since for x's for which $p(x) = 0$, $\mathbb{E}_{\alpha \in F_q}[\omega^{\alpha p(x)}] = 1$, and for x such that $p(x) \neq 0$, $\mathbb{E}_{\alpha \in F_q}[\omega^{\alpha p(x)}] = 0$. We thus get that

$$\mathbb{E}_{\alpha \in \mathbb{F}_q \setminus \{0\}}[\mathrm{bias}(\alpha p)] = 1 - \frac{q}{q-1}\mathrm{wt}(p) \geq \frac{q}{q-1}\epsilon$$

So, there must exist $\alpha \neq 0$ such that $\mathrm{bias}(\alpha p) \geq \epsilon$. □

Chapter 12

Holes in generalized Reed-Muller codes

The possible relative weights of codewords of Generalized Reed–Muller codes are studied. Let $\mathrm{RM}_q(r,m)$ denote the code of polynomials over the finite field $\mathbb{F}_q$ in m variables of total degree at most r. The relative weight of a codeword $f \in \mathrm{RM}_q(r,m)$ is the fraction of non-zero entries in f. The possible relative weights are studied, when the field $\mathbb{F}_q$ and the degree r are fixed, and the number of variables m tends to infinity. It is proved that the set of possible weights is sparse - for any α which is not rational of the form $\alpha = \ell/q^k$ there exists some $\epsilon > 0$ such that no weights fall in the interval $(\alpha - \epsilon, \alpha + \epsilon)$. This demonstrates a new property of the weight distribution of Generalized Reed-Muller codes.

12.1 Introduction

In this work we study the possible weights of codewords of Generalized Reed–Muller codes. For a prime power q, let $\mathbb{F}_q$ denote the field of q elements. The r^{th}-order Generalized Reed–Muller code over $\mathbb{F}_q$, denoted by $\mathrm{RM}_q(r,m)$, is a linear code over $\mathbb{F}_q$, whose codewords $f \in \mathrm{RM}_q(r,m) : \mathbb{F}_q^m \to \mathbb{F}_q$ are evaluations of polynomials over $\mathbb{F}_q$ in m variables of total degree at most r. Reed–Muller codes correspond to the special case of $q = 2$. Both Reed–Muller codes and the more general family of Generalized Reed–Muller codes have attracted research for many years; to quote [MS83],

Reed–Muller (or RM) codes are one of the oldest and best understood families of codes.

In this work we study Generalized Reed–Muller codes $\mathrm{RM}_q(r,m)$, when the field $\mathbb{F}_q$ and order r are fixed, and the number of variables m tends to infinity. The basic property that we study is the relative weights of codewords $f \in \mathrm{RM}_q(r,m)$.

Definition 12.1 (Relative weight)**.** *The* relative weight *of a codeword $f \in \mathrm{RM}_q(r,m)$ is the fraction of non-zero elements,*

$$\mathrm{wt}(f) = \frac{1}{q^m}|\{\mathbf{x} \in \mathbb{F}_q^m : f(\mathbf{x}) \neq 0\}|.$$

We denote by $A_q(r,m)$ the set of all weights of codewords $f \in \mathrm{RM}_q(r,m)$,

$$A_q(r,m) = \{\mathrm{wt}(f) : f \in \mathrm{RM}_q(r,m)\}.$$

There are two simple constraints on the values in $A_q(r,m)$. The first constraint relates to the the fact that the code is finite - since a relative weight is the fraction of inputs $\mathbf{x}$ for which $f(\mathbf{x}) \neq 0$, all values in $A_q(r,m)$ are rational of the form $\frac{\ell}{q^m}$. The second one relates to the minimal distance of the code.

Definition 12.2 (Minimal distance)**.** *The* minimal relative distance *of a code $\mathcal{C}$ is the minimal weight of a non-zero codeword $f \in \mathcal{C}$.*

The minimal distance of Generalized Reed–Muller codes is well-known (see for example [MS83]).

Fact 12.1. *Let $r = (q-1)a + b$ where $0 \leq b \leq q-1$. The minimal relative distance of* $\mathrm{RM}_q(r,m)$ *is*

$$\delta_q(r) = \frac{1}{q^a}\left(1 - \frac{b}{q}\right).$$

We are interested in the set of possible weights of $A_q(r,m)$ for fixed q and r when $m \to \infty$. Clearly $A_q(r,m) \subset A_q(r,m')$ when $m < m'$. Thus it makes sense to look at the limit

$$A_q(r) = \bigcup_{m=1}^{\infty} A_q(r,m).$$

Our main object of study is the set $A_q(r)$. A priory, one would think that the set $A_q(r)$ is dense inside the permissible range, given by the minimal distance of the code. However, our main result shows that the truth is quite far from this. First we define q-rational numbers.

Definition 12.3 (q-rational numbers)**.** *A rational number $\alpha \in [0,1]$ is* q-rational *if it is of the form $\alpha = \frac{\ell}{q^k}$ for some integers ℓ, k.*

Note that if $q = p^t$ for a prime p, then q-rational numbers and p-rational numbers define the same set.

Theorem 12.1 (Main theorem)**.** *Let $\alpha \in [0,1]$ be a number which is not q-rational. Then there exists some $\epsilon > 0$ such that $A_q(r)$ contains no value in the range $(\alpha - \epsilon, \alpha + \epsilon)$. Equivalently, there is no sequence of polynomials $f_1, f_2, \ldots$ over $\mathbb{F}_q$ of degree at most r, each possibly on a different number of variables, such that $\lim_{k\to\infty} \mathrm{wt}(f_k) = \alpha$.*

For example, there is no sequence of polynomials $f_1, f_2, \ldots$ over $\mathbb{F}_3$ of total degree at most 17, such that $\lim_{k\to\infty} \mathrm{wt}(f_k) = \frac{1}{2}$. The following is an immediate corollary of Theorem 12.1.

Corollary 12.1. *Let $\overline{C}$ denote the closure of $C \subset [0,1]$. We have*

$$\bigcup_{r=1}^{\infty} \overline{A_q(r)} = \{\textit{the set of } q\textit{-rationals}\}$$

Proof. Theorem 12.1 shows that $\overline{A_q(r)}$ contains only q-rational numbers. On the other hand, any q-rational number $\alpha = \frac{\ell}{q^k}$ can be realized as the relative weight of some function $f : \mathbb{F}_q^k \to \mathbb{F}_q$. Clearly $f \in \mathrm{RM}_q((q-1)k, k)$ and $\alpha \in A_q((q-1)k)$. □

Consider the following possible strengthening of Theorem 12.1: all weights in $A_q(r)$ are of the form $\frac{\ell}{q^t}$, where t is bounded. This is not true, as the following example shows.

Example 12.1. *Consider the set $A_2(2)$, i.e. the set of relative weights of quadratic polynomials over $\mathbb{F}_2$. For every $k \in \mathbb{N}$, let f be the polynomial*

$$f(x_1, \ldots, x_{2k}) = x_1x_2 + x_3x_4 + \cdots + x_{2k-1}x_{2k}$$

A straightforward calculation shows that

$$\mathrm{wt}(f) = \frac{2^k + 1}{2^{k+1}}$$

Thus, the set of attainable weights by quadratics over $\mathbb{F}_2$ contains 2-rational numbers with unbounded denumerators. In fact, the set $A_2(2)$ is given by

$$A_2(2) = \left\{0, \frac{1}{2}, 1\right\} \cup \left\{\frac{2^k + 1}{2^{k+1}} : k \in \mathbb{N}\right\} \cup \left\{\frac{2^k - 1}{2^{k+1}} : k \in \mathbb{N}\right\}$$

and the only limit point of $A_2(2)$ is $\frac{1}{2}$.

12.1.1 Related results

The weights of codewords of Reed–Muller and Generalized Reed–Muller codes have been extensively studied. A related line of research is that of divisibility of weights of codewords. The (non-relative) weight of a codeword is the number of non-zero elements in it. Ax [Ax64] proved that weights of codewords $f \in \mathrm{RM}_q(r, m)$ are divisible by $q^{\lceil r/m \rceil - 1}$. Notice that such results are incomparable to our results - if we write the weight of a codeword as a number in base q, divisibility results relate to the lower digits of the weight, while our results relate to the upper digits of the weight.

Another line of research, which is probably the most extensively studied one, is investigating the weight distribution of Generalized Reed–Muller codes. That is, estimating the number of codewords $f \in \mathrm{RM}_q(r, m)$ of relative weight at most ρ, for values of ρ ranging between the minimal distance of the code and 1.

No exact (or even asymptotically tight) answer is known in the general case. However, partial results are known in several cases. The case of $r = 1$, i.e. that of linear polynomials, is easy. Dixon (see [MS83]) provided a complete canonical characterization of quadratic polynomials. This amounts to a full understanding of the weight distribution of $\mathrm{RM}_q(2, m)$. The set of codewords $f \in \mathrm{RM}_q(r, m)$ which attain exactly the minimal weight was characterized by Delsarte et al. [DGW70]. In the case of Reed–Muller codes, i.e when $q = 2$, Kasami and Tokura [KT70] gave a complete characterization of codewords of weight at most *twice* the minimal distance of the code. Azumi et al. [AKT76] characterized codewords of weight at most 2.5 times the minimal distance of the code. Recently, Kaufman, Porat and the author [KLP10] gave an estimate on the number of codewords in Reed–Muller codes for all distances.

12.1.2 Organization

The paper is organized as follows. Theorem 12.1 is proved in Section 12.2. The proof is based on a technical lemma which is proved in Section 12.3.

12.2 Proof of Theorem 12.1

We study codewords $f \in \mathrm{RM}_q(r,m)$. Equivalently, we study polynomials: f is a polynomial over $\mathbb{F}_q$ in m variables of total degree at most r. First, we fix some notation. We denote elements of $\mathbb{F}_q^m$ by $\mathbf{x} = (x_1, \ldots, x_m)$, and polynomials/functions by $f(\mathbf{x}) = f(x_1, \ldots, x_m)$. When we speak about the degree of a polynomial, we always mean its total degree. We denote probabilities according to a distribution D by $\mathbb{P}_{z\sim D}[\texttt{event}]$. For a set S we denote by U_S the uniform distribution over S, and we shorthand $\mathbb{P}_{z\in S}$ for $\mathbb{P}_{z\sim U_S}$. We let $\mathbb{N} = \{1, 2, \ldots\}$ denote the set of natural numbers. We initiate the proof of Theorem 12.1 by showing that it suffices to prove it over prime fields.

Claim 12.1. *Let $\mathbb{F}_q$ be a finite field for $q = p^t$. Let $f(x_1, \ldots, x_m)$ be a degree-r polynomial over $\mathbb{F}_q$. There exists a polynomial $f'(x_1, \ldots, x_{mt})$ over $\mathbb{F}_p$ of degree rt such that $\mathrm{wt}(f') = \mathrm{wt}(f)$. In particular, $A_q(r) \subset A_p(rt)$.*

Proof. We use standard facts about extension fields (for details, see any standard algebra book, for example [BM65]): elements of $\mathbb{F}_q = \mathbb{F}_{p^t}$ correspond to vectors in $\mathbb{F}_p^t$, and multiplication over $\mathbb{F}_q$ corresponds to a bilinear map over $\mathbb{F}_p^t$. Thus if we denote the representation of $x_i \in \mathbb{F}_q$ by $\vec{x}_i \in \mathbb{F}_p^t$, then the value of the polynomial $f(x_1, \ldots, x_m)$ is represented by $(f_1(\vec{\mathbf{x}}), \ldots, f_t(\vec{\mathbf{x}}))$ where $\vec{\mathbf{x}} = \vec{x}_1 \ldots \vec{x}_m \in \mathbb{F}_p^{mt}$ and $f_1, \ldots, f_t$ are polynomials over $\mathbb{F}_p$ of degree r. Let $g : \mathbb{F}_p^t \to \mathbb{F}_p$ be defined as $g(0, \ldots, 0) = 0$ and $g(z) = 1$ for any $z \in \mathbb{F}_p^t \setminus 0^t$. Note that g is a polynomial over $\mathbb{F}_p$ of degree t. Define $f'(\vec{\mathbf{x}}) = g(f_1(\vec{\mathbf{x}}), \ldots, f_t(\vec{\mathbf{x}}))$. Observe that we have $f'(\vec{\mathbf{x}}) = 0$ iff $f(\mathbf{x}) = 0$. In particular $\mathrm{wt}(f') = \mathrm{wt}(f)$. To conclude the proof, observe that as $\deg(f_1), \ldots, \deg(f_t) = r$ and $\deg(g) = t$ we have that $\deg(f') \le rt$. □

Thus we restrict our attention from now on to polynomials over a prime finite field $\mathbb{F}_p$ of fixed degree r. In order to prove Theorem 12.1 we will show that for any degree-r polynomial $f(x_1, ..., x_m)$, there exists a function $g(x_1, ..., x_c)$ on a constant number of inputs (i.e. independent of m), such that $\mathrm{wt}(f) \approx \mathrm{wt}(g)$.

Lemma 12.1. *Let $\epsilon : \mathbb{N} \to (0, 1)$ be an arbitrary mapping from the natural numbers to $(0, 1)$. For any constant degree r there exists a constant $C = C(\mathbb{F}_p, r, \epsilon(\cdot))$ such that the following holds: for any degree-r polynomial $f(\mathbf{x}) = f(x_1, ..., x_m)$, there exists $c \le C$ and a function $g(x_1, ..., x_c)$, such that*

$$|\mathrm{wt}(f) - \mathrm{wt}(g)| < \epsilon(c)$$

Note that it is straightforward to find a function $g(x_1, \ldots, x_c)$ such that $\mathrm{wt}(f) \approx \mathrm{wt}(g)$ if the required approximation is fixed a priory. The novelty of Lemma 12.1 is that this can be achieved even if the error is allowed to depend arbitrarily on the number of inputs c.

Remark. *In fact, a somewhat stronger version of the lemma also holds. Not only is* $|\mathrm{wt}(f) - \mathrm{wt}(g)| < \epsilon(c)$*, but the statistical distance between the distributions of* f *and* g*, evaluated over uniform inputs, is bounded by* $\epsilon(c)$*. However, we will not need this stronger version in the proof of Theorem 12.1.*

We now prove Theorem 12.1 using Lemma 12.1.

Proof of Thereom 12.1. Let $\alpha \in (0,1)$ be a number which is not p-rational, and assume by contradiction there exists a sequence of polynomials $f_1, f_2, \ldots$ of degree at most r, where $f_k = f_k(x_1, ..., x_{m_k})$, such that $\lim_{k\to\infty} \mathrm{wt}(f_k) = \alpha$. The main idea of the proof is to show that we can approximate the weights of $f_1, f_2, \ldots$ by weights of functions depending on a small number of variables. However, as α is not p-rational it cannot be approximated too well by weights of such functions. We now proceed with the details.

Let δ be a mapping from the natural numbers to $(0,1)$ defined as follows. For every $c \in \mathbb{N}$, define $\delta(c)$ to be the distance of α from all rational numbers of the form $\frac{\ell}{p^c}$. Explicitly, $\delta(c)$ is given by

$$\delta(c) = \min\left\{\alpha - \frac{\lfloor \alpha p^c \rfloor}{p^c}, \frac{\lceil \alpha p^c \rceil}{p^c} - \alpha\right\}$$

Note that $\delta(\cdot)$ is non-increasing, and by our assumption that α is not p-rational, $\delta(c) > 0$ for all $c \in \mathbb{N}$.

Set $\epsilon(c) = \frac{\delta(c)}{4}$. Once we fix the mapping $\epsilon(\cdot)$, we can use Lemma 12.1: there exists some constant $C = C(\mathbb{F}_p, r, \epsilon(\cdot))$, such that for any polynomial f_k there exists $c_k \le C$, and a function $g_k(x_1, ..., x_{c_k})$, such that

$$|\mathrm{wt}(f_k) - \mathrm{wt}(g_k)| < \epsilon(c_k) = \frac{\delta(c_k)}{4} \tag{12.1}$$

Since $\lim_{k\to\infty} \mathrm{wt}(f_k) = \alpha$, and $\epsilon(\cdot)$ is positive, there exists some k such that

$$|\mathrm{wt}(f_k) - \alpha| < \epsilon(C) = \frac{\delta(C)}{4} \tag{12.2}$$

Combining (12.1) and (12.2), and since $\delta(\cdot)$ is non-increasing, we get that

$$|\mathrm{wt}(g_k) - \alpha| < \frac{\delta(c_k)}{4} + \frac{\delta(C)}{4} \le \frac{\delta(c_k)}{2} \tag{12.3}$$

We now show this cannot hold. The function g_k is a function on c_k inputs; thus, its weight is of the form $\frac{\ell}{p^{c_k}}$. By the definition of $\delta(\cdot)$:

$$|\mathrm{wt}(g_k) - \alpha| = \left|\frac{\ell}{p^{c_k}} - \alpha\right| \ge \delta(c_k) \tag{12.4}$$

Combining (12.3) and (12.4) yields a contradiction. Thus, α must be p-rational.

□

12.3 Proof of Lemma 12.1

The proof of Lemma 12.1 is based on regularity results for constant degree polynomials by Green and Tao [GT07] and by Kaufman and Lovett [KL08]. We first make some definitions. In this section, all polynomials will be polynomials over $\mathbb{F}_p$ in m variables.

Definition 12.4 (rank of polynomials)**.** *Let $f(\mathbf{x})$ be a degree-r polynomial. The $(r-1)$-rank of f, denoted by $rank_{r-1}(f)$, is the minimal number of degree-$(r-1)$ polynomials required to compute f. This means that $rank_{r-1}(f)$ is the minimal c such that there exist polynomials $g_1(\mathbf{x}), ..., g_c(\mathbf{x})$ of degree at most $r-1$ and a function $F : \mathbb{F}_p^c \to \mathbb{F}_p$ such that*

$$f(\mathbf{x}) = F\left(g_1(\mathbf{x}), ..., g_c(\mathbf{x})\right)$$

Definition 12.5 (regularity of polynomials)**.** *A degree-r polynomial $f(\mathbf{x})$ is T-regular if $rank_{r-1}(f) > T$. A set of polynomials $\{f_1(\mathbf{x}), ..., f_c(\mathbf{x})\}$ is T-regular if all their non-zero linear combinations are T-regular. That is, for any $a_1, ..., a_c \in \mathbb{F}_p$ not all zero, let $f'(\mathbf{x}) = a_1 f_1(\mathbf{x}) + \cdots + a_c f_c(\mathbf{x})$. We require that f' is not identically zero, and that if $degree(f') = k$, then $rank_{k-1}(f') > T$.*

We will need the following result from [GT07]: any degree-r polynomial f is a function of a constant number of regular polynomials $g_1, \ldots, g_c$, even if the regularity requirements on $g_1, \ldots, g_c$ depend on the number of polynomials c.

Lemma 12.2 (Lemma 2.3 in [GT07])**.** *Let $\mathbb{T} : \mathbb{N} \to \mathbb{N}$ by an arbitrary mapping. There exists a constant $C_1 = C_1(\mathbb{F}_p, r, \mathbb{T}(\cdot))$ such that the following holds. For any degree-r polynomial $f(\mathbf{x})$ there exists some $c \leq C_1$, a set of polynomials $g_1(\mathbf{x}), ..., g_c(\mathbf{x})$ of degree at most r and a function $F : \mathbb{F}_p^c \to \mathbb{F}_p$, such that:*

1. *$f(\mathbf{x}) = F(g_1(\mathbf{x}), ..., g_c(\mathbf{x}))$,*
2. *The set of polynomials $\{g_1(\mathbf{x}), ..., g_c(\mathbf{x})\}$ is $\mathbb{T}(c)$-regular.*

We also need a result relating regularity of polynomials to their joint distribution.

Definition 12.6 (distribution of polynomials)**.** *Let $f : \mathbb{F}_p^m \to \mathbb{F}_p$ be a polynomial. Its distribution $\mathcal{D}(f)$ is the distribution (taking values in $\mathbb{F}_p$) of applying f on a random input $\mathbf{x} \in \mathbb{F}_p^m$,*

$$\mathcal{D}(f) = f(\mathbf{x})_{\mathbf{x} \sim U_{\mathbb{F}_p^m}}.$$

For a set of polynomials $f_1, \ldots, f_c : \mathbb{F}_p^m \to \mathbb{F}_p$, their joint distribution $\mathcal{D}(f_1, \ldots, f_c)$ (taking values in $\mathbb{F}_p^c$) is the distribution of applying $f_1, \ldots, f_c$ on a common random input $\mathbf{x} \in \mathbb{F}_p^m$,

$$\mathcal{D}(f_1, \ldots, f_c) = (f_1(\mathbf{x}), \ldots, f_c(\mathbf{x}))_{\mathbf{x} \sim U_{\mathbb{F}_p^m}}.$$

Definition 12.7 (statistical distance)**.** *Let D', D'' be two distributions taking values in the same set S. Their statistical distance is*

$$dist(D', D'') = \frac{1}{2} \sum_{s \in S} \left| \mathbb{P}[D' = s] - \mathbb{P}[D'' = s] \right|.$$

The following result from [KL08] shows that polynomials whose distribution is not close to uniform must have low rank.

Lemma 12.3 (Theorem 4 in [KL08])**.** *Let $f(\mathbf{x})$ be a degree-r polynomial such that $dist(\mathcal{D}(f), U_{\mathbb{F}_p}) \geq \epsilon$. Then $rank_{r-1}(f) \leq C_2(\mathbb{F}_p, r, \epsilon)$.*

We combine Lemma 12.2 and Lemma 12.3 to prove the following lemma, showing that any degree-r polynomial is a function of a constant number of polynomials which are uncorrelated.

Lemma 12.4. *Let $\epsilon : \mathbb{N} \to (0,1)$ be an arbitrary mapping from the natural numbers to $(0,1)$. For any constant degree r there exists a constant $C = C(\mathbb{F}_p, r, \epsilon(\cdot))$ such that the following holds: For any degree-r polynomial $f(\mathbf{x})$ there exists some $c \leq C$, a set of polynomials $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$ of degree at most r and a function $F : \mathbb{F}_p^c \to \mathbb{F}_p$, such that:*

1. *$f(\mathbf{x}) = F(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x}))$,*
2. *$dist(\mathcal{D}(g_1, \ldots, g_c), U_{\mathbb{F}_p^c}) < \epsilon(c)$.*

Proof. We will choose $\mathbb{T} : \mathbb{N} \to \mathbb{N}$ large enough, to be specified later, and apply Lemma 12.2. Let $g_1, \ldots, g_c$ be the polynomials given by the lemma such that $f(\mathbf{x}) = F(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x}))$, and the set $\{g_1, \ldots, g_c\}$ is $\mathbb{T}(c)$-regular. We will show that if we choose $\mathbb{T}(\cdot)$ large enough, we can guarantee that $\mathcal{D}(g_1, \ldots, g_c)$ is close to uniform.

We first reduce the task to guaranteeing that all the non-zero linear combinations of $g_1, \ldots, g_c$ are close to uniform. We claim that in order to guarantee that $dist(\mathcal{D}(g_1, \ldots, g_c), U_{\mathbb{F}_p^c}) < \epsilon(c)$, it is enough to guarantee for every non-zero linear combination $g'(\mathbf{x}) = a_1 g_1(\mathbf{x}) + \cdots + a_c g_c(\mathbf{x})$ that $dist(\mathcal{D}(g'), U_{\mathbb{F}_p}) < p^{-c}\epsilon(c)$. The proof is by simple Fourier analysis, see for example Claim 33 in [BV07].

Given this reduction, we show it is enough to require that g' is regular. Assume $dist(\mathcal{D}(g'), U_{\mathbb{F}_p}) \geq p^{-c}\epsilon(c)$. Then either $g' \equiv 0$, or, by Lemma 12.3, if $degree(g') = k$ then

$$rank_{k-1}(g') \leq C_2(\mathbb{F}_p, k, p^{-c}\epsilon(c)) \tag{12.5}$$

In any case, if we set $\mathbb{T}(c) = \max_{1 \leq k \leq r} C_2(\mathbb{F}_p, k, p^{-c}\epsilon(c))$, we get that the set $\{g_1, \ldots, g_c\}$ is not $\mathbb{T}(c)$-regular, since g' is not $\mathbb{T}(c)$-regular. This is a contradiction to the guarantee of Lemma 12.2.

Hence we conclude that the joint distribution $\mathcal{D}(g_1, \ldots, g_c)$ has statistical distance of at most $\epsilon(c)$ to the uniform distribution $\mathbb{F}_p^c$, where $c \leq C$ and

$$C = C_1(\mathbb{F}_p, r, \mathbb{T}(\cdot))$$

□

We will also need the following simple claim: the statistical distance between distributions bounds the probability that any event can distinguish between them.

Claim 12.2. *Let D', D'' be two distributions taking values in the same set S. Then for any subset $E \subseteq S$ we have*

$$|\mathbb{P}_{z \sim D'}[z \in E] - \mathbb{P}_{z \sim D''}[z \in E]| \leq dist(D', D'')$$

We are now ready to prove Lemma 12.1.

Proof of Lemma 12.1. Let $f(\mathbf{x})$ be a degree-r polynomial. Apply Lemma 12.4. There exists some constant $C = C(\mathbb{F}_p, r, \epsilon(\cdot))$ such that there is $c \leq C$, a set of polynomials $g_1(\mathbf{x}), ..., g_c(\mathbf{x})$ and a function $F : \mathbb{F}_p^c \to \mathbb{F}_p$ such that

1. $f(\mathbf{x}) = F(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x}))$,
2. $dist(\mathcal{D}(g_1, \ldots, g_c), U_{\mathbb{F}_p^c}) < \epsilon(c)$.

We claim that the function $F(y_1, \ldots, y_c)$, where $y_1, \ldots, y_c \in \mathbb{F}_p$ are new independent variables, have approximately the same relative weight as that of $f(\mathbf{x}) = F(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x}))$. We bound

$$
\begin{aligned}
&|\mathrm{wt}(f) - \mathrm{wt}(F)| = \\
&|\mathbb{P}_{\mathbf{x}\in\mathbb{F}_p^m}\left[F(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})) \neq 0\right] - \\
&\ \mathbb{P}_{y_1,\ldots,y_c\in\mathbb{F}_p}\left[F(y_1, \ldots, y_c)\right] \neq 0| = \\
&|\mathbb{P}_{\mathbf{x}\in\mathbb{F}_p^m}\left[(g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})) \in F^{-1}(\mathbb{F}_p \setminus \{0\})\right] - \\
&\ \mathbb{P}_{y_1,\ldots,y_c\in\mathbb{F}_p}\left[(y_1, \ldots, y_c) \in F^{-1}(\mathbb{F}_p \setminus \{0\})\right]| \leq \\
&dist(\mathcal{D}(g_1, \ldots, g_c), \mathcal{D}(y_1, \ldots, y_c)) = \\
&dist(\mathcal{D}(g_1, \ldots, g_c), U_{\mathbb{F}_p^c}) < \epsilon(c).
\end{aligned}
$$

□

Chapter 13

Affine invariant codes, and extension to Weil bound

In this work we consider linear codes which are locally testable in a sublinear number of queries. We give the first general family of locally testable codes of exponential size. Previous results of this form were known only for codes of quasi-polynomial size (e.g. Reed-Muller codes). We accomplish this by showing that any affine invariant code $\mathcal{C}$ over $\mathbb{F}_{p^n}$ of size $p^{p^{\Omega(n)}}$ is locally testable using $poly(\log_p |\mathcal{C}|/n)$ queries. Previous general result for affine invariant codes were known only for sparse codes, i.e. codes of size $p^{O(n)}$. The main new ingredients used in our proof are a new extension of the Weil bound for character sums, and a Fourier-analytic approach for estimating the weight distribution of affine invariant codes.

Joint work with Tali Kaufman.

13.1 Introduction

We study in this work families of locally testable codes. Let $\mathbb{F}_N = \mathbb{F}_{p^n}$ be a finite field, where we think of p as either constant or small. A code is a family of functions $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$. All codes we consider in this work are linear[1]. The dimension of a code is $\dim(\mathcal{C}) = \log_p(|\mathcal{C}|)$.

A code is *locally testable* if there is a randomized algorithm, which when given as input a function $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$, probes f in a small number of locations and determines (with high probability) whether $f \in \mathcal{C}$ or f is far[2] from all codewords of $\mathcal{C}$. A code is q-locally testable if the number of probes is at most q, where q is sublinear in the code length, i.e. $q = o(N)$.

Most of the study of locally testable codes has been focused on codes testable with constant query complexity (i.e. $q = O(1)$) or with poly-logarithmic query complexity (i.e. $q = (\log N)^{O(1)}$). They appear as low-degree tests in the $IP = PSPACE$, $MIP = NEXP$

[1] A code $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ is linear if for any $f(x), g(x) \in \mathcal{C}$ also $h(x) = \alpha f(x) + \beta g(x) \in \mathcal{C}$ where $\alpha, \beta \in \mathbb{F}_p$.

[2] If f has distance ϵ from $\mathcal{C}$, i.e. if $\min_{g \in \mathcal{C}} \Pr_{x \in \mathbb{F}_{p^n}}[f(x) \neq g(x)] = \epsilon$, we require the local test to reject f with probability at least $\Omega(\epsilon)$.

and $PCP = NP$ theorems, and indeed the work of [GS06] (which was later partly derandomized by [BSSVW03]) elucidates their role as the "combinatorial heart" of PCPs.

In general, there is a tradeoff between the rate of the code $\dim(\mathcal{C})/N$ and the query complexity of testing this code. A major open problem in this field is whether one can enjoy the best of both worlds: a code of constant rate which is locally testable with a constant query complexity.

One line of research focuses on constructing explicit codes which try to approach this optimal tradeoff. The best results to date are by Ben-Sasson and Sudan [BSS05] and Dinur [Din07] (see also Meir [Mei08]) which achieve an explicit binary code of rate $\frac{1}{(\log N)^{O(1)}}$ which is testable using a constant number of probes.

A second line of research focuses on characterization of general families of codes that are locally testable [BLR93, RS93, NAR03, JPRZ04, KR04, KS08, KS07, KL05, GKS09, KS10]. Many results in this field apply only to *sparse* codes over binary fields $\mathbb{F}_{2^n}$, which are codes of dimension $O(\log N)$ [KL05, KS07, GKS09, KS10]. Another example is *Generalized Reed-Muller codes* which are the family of polynomials $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$ of total degree at most d. These codes are testable using $p^{\frac{d}{p-1}} = exp(d)$ queries, while having dimension $O(n^d)$ [NAR03, JPRZ04, KR04]. Such codes can be locally testable with sublinear number of queries for $d \leq O(\log n)$, which gives codes of quasi-logarithmic dimension $\dim(\mathcal{C}) \leq (\log N)^{\log\log N}$.

Our work falls into the latter line of research. We exhibit a general family of codes of almost optimal dimension $\dim(\mathcal{C}) = N^{\Omega(1)}$ which are locally testable with sublinear query complexity. We achieve this by studying *affine invariant codes*. A code $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ is affine invariant if it is invariant under affine transformation of the coordinates of input space. That is, if $f(x) \in \mathcal{C}$ then also $g(x) = f(ax+b) \in \mathcal{C}$ for any $a, b \in \mathbb{F}_{p^n}, a \neq 0$. Previous results [GKS09] showed that sparse affine invariant codes (i.e., codes of size $p^{O(n)}$) are locally testable. We significantly extend this to codes of up to exponential size, i.e. of size at most $p^{p^{\Omega(n)}}$.

Theorem 13.1 (Main result)**.** *Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be a linear code which is affine invariant of dimension $\dim(\mathcal{C}) \leq p^{\alpha n}$, where $\alpha > 0$ is an absolute constant. Then $\mathcal{C}$ is locally testable with query complexity $q = poly(\dim(\mathcal{C})/n) = o(p^n)$. In particular, any sparse affine invariant code (i.e. with $\dim(\mathcal{C}) = O(n)$) is locally testable with constant query complexity $q = O(1)$. The parameter α can be chosen to be any $\alpha < 1/32$ for large enough n.*

This generalizes previous works in several aspects: our result applies to codes of exponential size $exp(N^\alpha)$, while previous results apply only to codes of polynomial size $N^{O(1)}$ or quasi-polynomial size $exp(\log N^{\log\log N})$. Previous results on sparse codes applied only to binary fields $\mathbb{F}_{2^n}$, while our result applies to any field of small characteristic. Note that a recent result of Ben-Sasson and Sudan [BSS09, Sud10] shows that affine invariant codes that are testable with constant number of queries cannot have exponential rate. Thus, our testing result of exponentially large codes cannot be improved to testing with constant locality.

The main new ingredients in our work is a Fourier-analytic approach for estimating the weight distribution of affine invariant codes, and a new extension of the Weil bound for character sums of low-degree polynomials. We start by describing our new result for character sums for polynomials, and then discuss its relation to proving local testability of affine invariant codes. The proof of our new extension for the Weil bound relies on

techniques borrowed from additive combinatorics. This demonstrates yet another connection between additive combinatorics and theoretical computer science. Such connections were used before to establish results regarding pseudorandom generators [BV07, Lov08, Vio08] and list-decoding of codes [KLP10].

13.1.1 Character sums

Let $\mathbb{F}$ be a finite field. An additive character is a function $\chi : \mathbb{F} \to \mathbb{C}$ for which $\chi(x+y) = \chi(x)\chi(y)$ (and which is not the identically zero function). For example, if $\mathbb{F} = \mathbb{F}_q$ is a prime finite field then the additive characters are given by $\chi_a(x) = e^{\frac{2\pi i}{q}ax}$ for $a \in \mathbb{F}_q$. In the general case of $\mathbb{F} = \mathbb{F}_{p^n}$, the additive characters are given by $\chi_a(x) = e^{\frac{2\pi i}{p}\mathrm{Tr}(ax)}$, where $a \in \mathbb{F}_{p^n}$ and the Trace operator $\mathrm{Tr} : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is defined as $\mathrm{Tr}(x) = \sum_{i=0}^{n-1} x^{p^i}$.

The Weil bound for character sums [Wei48] is a general result regarding character sums of low-degree polynomials over a finite field $\mathbb{F}$. Let $f(x) \in \mathbb{F}[x]$ be a univariate polynomial of degree k. Let $\chi : \mathbb{F} \to \mathbb{C}$ be any additive character. Weil's bound states that either $\chi(f(x))$ is constant, or is distributed close to uniform when $x \in \mathbb{F}$ is uniformly chosen.

Theorem 13.2 (Weil bound [Wei48])**.** *Let $f(x)$ be a univariate polynomial over $\mathbb{F}$ of degree $\leq |\mathbb{F}|^{1/2-\delta}$. Let $\chi : \mathbb{F} \to \mathbb{C}$ be any additive character. Then either $\chi(f(x))$ is constant for all $x \in \mathbb{F}$, or*

$$|\mathbb{E}_{x\in\mathbb{F}}[\chi(f(x))]| \leq |\mathbb{F}|^{-\delta}.$$

The Weil bound is very effective to polynomials of degree $k \ll \sqrt{|\mathbb{F}|}$, however it fails for polynomials of degree $k \geq \sqrt{|\mathbb{F}|}$. We establish a general result in fields of small characteristics $\mathbb{F}_{p^n}$ which allows to extend polynomials by a small number of monomials of larger degree, as long as they have small *weight degree*.

Definition 13.1 (Weight degree)**.** *Let $t \in \{0, \ldots, p^n - 1\}$. The weight degree of t is the hamming weight of the digits of t in base p. That is, let $t = \sum_{i=0}^{n-1} t_i p^i$ be the representation of t in base p, where $0 \leq t_i \leq p-1$. The weight degree of t is*

$$\mathrm{wt}(t) = \sum_{i=0}^{n-1} t_i.$$

The weight degree of a monomial x^t is the weight degree of t, and the weight degree of a univariate polynomial $f(x)$ is the maximal weight degree of a monomial in it with a nonzero coefficient.

We prove the following extension of the Weil bound in case $f(x)$ is the sum of a low degree polynomial and a small number of monomials of bounded weight degree (but of arbitrary degree).

Theorem 13.3 (Extension of the Weil bound)**.** *Let $f(x) = g(x) + h(x)$ be a univariate polynomial over $\mathbb{F}_{p^n}$, where $g(x)$ is a polynomial of degree $\leq |\mathbb{F}|^{1/2-\delta}$ and $h(x)$ is the sum of at most $k \geq 1$ monomials, each of weight degree at most d. Let $\chi : \mathbb{F}_{p^n} \to \mathbb{C}$ be an additive character. Then either $\chi(f(x))$ is constant for all $x \in \mathbb{F}_{p^n}$, or*

$$|\mathbb{E}_{x\in\mathbb{F}}[\chi(f(x))]| \leq |\mathbb{F}_{p^n}|^{-\frac{\delta}{2d^2 2^{dk}}}.$$

Note that in order to get a meaningful bound, we need our parameters to obey $kd^2 2^d \leq O(n)$. Note that for $d \leq (1-\epsilon)\log_2(n)$ we may have $k = n^{O(1)}$. This can be compared to a relatively recent result of Bourgain [Bou05] of a similar flavor. We state it below informally, as the exact formulation is somewhat complex, and we will not require it in the paper.

Theorem 13.4 (Bourgain's extension of Weil bound [Bou05])**.** *Let $f(x) = g(x) + h(x)$ be a univariate polynomial over a prime finite field $\mathbb{F}_q$, where $g(x)$ is a polynomial of degree $\leq |\mathbb{F}_q|^{1/2-\delta}$ and $h(x)$ is the sum of at most $k = O(1)$ monomials, each of degree at most $|\mathbb{F}_q|^{1-\epsilon}$. Let $\chi : \mathbb{F}_q \to \mathbb{C}$ be an additive character. Then either $\chi(f(x))$ is constant for all $x \in \mathbb{F}_q$, or*

$$\left|\mathbb{E}_{x\in\mathbb{F}_q}[\chi(f(x))]\right| \leq |\mathbb{F}_q|^{-\Omega(1)}.$$

Comparing our result with the result of Bourgain, we note two important advantages of our work: first, we can handle non-prime finite fields; second, when $d \leq O(\log n)$ is small enough, we may have $k = poly(n)$ monomials of high degree, while in the result of Bourgain one can take at most $k = O(1)$ such monomials. In contrast, the result of Bourgain does not assume a bound on the weight degree of the monomials. The two advantages of our work are crucial for the application to locally testing of exponentially large affine invariant codes. Bourgain's result was used in a similar fashion by Grigorescu, Kaufman and Sudan [GKS09] to establish a similar result which holds only for sparse affine invariant codes, i.e. codes of polynomial size. Our new character sum result allows us to extend their techniques to handle exponentially large affine invariant codes.

13.1.2 Connection between character sums and affine invariant codes

Affine invariant codes can be characterized by trace codes. Let $S \subseteq \{0, \ldots, p^n - 1\}$. The S-trace code over $\mathbb{F}_{p^n}$ is defined as the family of functions $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$ given by

$$\mathcal{T}(S) = \left\{ \left(\mathrm{Tr}(\sum_{e\in S} a_e x^e) : F_{p^n} \to \mathbb{F}_p \right) : a_e \in \mathbb{F}_{p^n} \right\}.$$

where we recall that the Trace function $\mathrm{Tr} : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is given by $\mathrm{Tr}(x) = \sum_{i=0}^{n-1} x^{p^i}$. For example, Generalized Reed-Muller codes $\mathrm{RM}(n,d)$, which are the family of functions $f : \mathbb{F}_p^n \to \mathbb{F}_p$ where f is an n-variate polynomial of total degree at most d, can be equivalently characterized as

$$\mathrm{RM}(n,d) = \mathcal{T}(\{e \in \{0,\ldots,p^n-1\} : \mathrm{wt}(e) \leq d\}).$$

We define two important properties of trace codes.

Definition 13.2 (Shift closed)**.** *Let $S \subseteq \{0, \ldots, p^n - 1\}$. The set S is said to be* shift closed *if, for every $e \in S$, we also have that $ep^\ell \pmod{p^n} \in S$ for all $\ell = 1, \ldots, n$.*

The term *shift closed* comes from viewing elements $e \in S$ as vectors in $\mathbb{F}_p^n$, given by the representation of e in base p. In this case, $ep^\ell \pmod{p^n}$ corresponds to a cyclic shift of the vector by ℓ coordinates.

Definition 13.3 (Shadow closed)**.** *Let $S \subseteq \{0, \ldots, p^n - 1\}$. The set S is said to be* shadow closed *if the following holds. For any $e \in S$, let $e = \sum_{i=0}^{n-1} e_i p^i$ be the representation of e in base p. Define the* support *of e to be the set of nonzero digits of e,*

$$\text{support}(e) = \{0 \leq i \leq n-1 : e_i \neq 0\}.$$

Let e' be obtained from e by changing some of the non-zero digits of e, i.e.

$$e' = \sum_{i \in \text{support}(e)} e'_i p^i.$$

Then we should have that also $e' \in S$. That is, S is shadow closed if

$$\left\{ \sum_{i \in \text{support}(e)} e'_i p^i : e \in S, (e'_i)_{i \in \text{support}(e)} \in \mathbb{F}_p \right\} \subseteq S.$$

A set S is said to be *affine closed* if it is both shift closed and shadow closed. The following general result was established by Kafuman and Sudan [KS08]. They show that the class of affine invariant linear codes is equivalent to the class of trace codes of affine closed sets.

Theorem 13.5 (Monomial extraction [KS08])**.** *Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be an affine invariant linear code. Then there exists an affine closed set $S \subseteq \{0, \ldots, p^n - 1\}$ such that $\mathcal{C} = \mathcal{T}(S)$. Moreover, for any affine closed set S the code $\mathcal{T}(S)$ is linear and affine invariant.*

Thus, to study affine invariant codes, we need to study trace codes. We now introduce two notions. The dual of a code $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ is defined as

$$\mathcal{C}^\perp = \left\{ (g : \mathbb{F}_{p^n} \to \mathbb{F}_p) : \sum_{x \in \mathbb{F}_{p^n}} f(x)g(x) = 0 \quad \forall f \in \mathcal{C} \right\}.$$

The *affine closure* of a function $g : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is the set of functions obtained by applying affine transformations on the coordinates of the input space of f, that is

$$\overline{\text{affine}}(g) = \left\{ (g(ax + b) : \mathbb{F}_{p^n} \to \mathbb{F}_p) : a, b \in \mathbb{F}_{p^n} \right\}.$$

It is easy to verify that if $\mathcal{C}$ is an affine invariant code, and $g \in \mathcal{C}^\perp$, then in fact $\overline{\text{affine}}(g) \subseteq \mathcal{C}^\perp$. An important case is when in fact $\overline{\text{affine}}(g)$ spans the entire code $\mathcal{C}^\perp$.

Definition 13.4 (Single orbit property)**.** *Let $g \in \mathcal{C}^\perp$. We say that $\mathcal{C}$ has the* single orbit property *for g if the affine closure of g is a spanning set for $\mathcal{C}^\perp$, that is if*

$$\mathcal{C} = \text{Span}(\overline{\text{affine}}(g))^\perp.$$

We will shortly see that the single orbit property is tightly connected to locally testing properties of the code $\mathcal{C}$. First, define the *weight* of $g : \mathbb{F}_{p^n} \to \mathbb{F}_p$ to be the number of coordinates where g evaluates to a nonzero value,

$$\mathrm{wt}(g) = |\{x \in \mathbb{F}_{p^n} : g(x) \neq 0\}|.$$

The following result was established by Kaufman and Sudan [KS08]. If $\mathcal{C}$ is an affine invariant code which has the single orbit property for a codeword $g \in \mathcal{C}^\perp$ of small weight, then $\mathcal{C}$ can be locally tested[3].

Theorem 13.6 (Theorem 2.9 in [KS08]). *Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be a linear code which is affine invariant. Assume there exists $g \in \mathcal{C}^\perp$ such that $\mathcal{C}$ has the single orbit property for g. Then $\mathcal{C}$ can be locally tested with $O(\mathrm{wt}(g)^2)$ queries.*

Hence, to show that $\mathcal{C}$ can be locally tested, it is sufficient to demonstrate that $\mathcal{C}^\perp$ is spanned by the orbit of a short codeword under the affine group.

Let $\mathcal{C} = \mathcal{T}(S)$ for some affine closed set $S \subseteq \{0, \ldots, p^n - 1\}$. The dual code of $\mathcal{C}$ is a dual-trace code $d\mathcal{T}(S)$, which can be verified (Claim 13.1) to be

$$d\mathcal{T}(S) = \left\{ (f : \mathbb{F}_{p^n} \to \mathbb{F}_p) : \sum_{x \in \mathbb{F}_{p^n}} f(x)x^e = 0 \quad \forall e \in S \right\}.$$

We need to establish that there exists $f \in d\mathcal{T}(S)$ of small weight such that $\mathrm{Span}(\overline{\mathrm{affine}}(f)) = d\mathcal{T}(S)$. Assume that this is false, i.e. that $\mathrm{Span}(\overline{\mathrm{affine}}(f)) \subsetneq d\mathcal{T}(S)$. Using the fact that S is affine invariant, we show (Corollary 13.1) that in fact $f \in d\mathcal{T}(S \cup \{e\})$ where $e \in \{0, \ldots, p^n - 1\} \setminus S$ has small weight.

Hence, in order to conclude the proof, we will show that for a suitably chosen weight ℓ, there exist codewords on weight ℓ in $d\mathcal{T}(S)$ which are not in any of $d\mathcal{T}(S \cup \{e\})$ for any $e \notin S$ which has small weight.

The main tool we develop in order to do so, is a tight estimate on the number of codewords of weight ℓ in dual-trace codes. We show the following result.

Lemma (Lemma 13.1, informal statement). *Let $S \subseteq \{0, \ldots, p^n - 1\}$ be affine closed of size $|S| \leq p^{\Omega(n)}$. Then there exists $\ell_{\min} = poly(|S|)$ and $\ell_{\max} = p^{\Omega(n)}$, such that for any $\ell_{\min} \leq \ell \leq \ell_{\max}$ the following holds. The number of codewords in $d\mathcal{T}(S)$ of weight exactly ℓ is given by*

$$\frac{C(p,\ell)}{\ell!} p^{n(\ell - |S'|)}(1 + o(1))$$

where $S' = \{e \in S : (p, e) = 1\}$ is the set of elements in S which are co-prime to p, and where $C(p, \ell)$ is given by

$$C(p,\ell) = \left| \left\{ (v_1, \ldots, v_\ell) \in (\mathbb{F}_p \setminus \{0\})^\ell : v_1 + \ldots + v_\ell = 0 \right\} \right|.$$

[3] In fact, the local test for $\mathcal{C}$ is performed by computing $\sum f(ax + b)g(x)$ for a small random subset of $a, b \in \mathbb{F}_{p^n}$. Note that to perform each such test, we only need to query $f(x)$ only on $x \in \mathbb{F}_{p^n}$ for which $g(x) \neq 0$.

Similar results were previously obtained over binary fields $\mathbb{F}_{2^n}$ using properties of Krawtchouk polynomials [KL05, KS07]. Our technique is different, and relies on methods from additive combinatorics and Fourier analysis. In particular it allows us to extend the result to arbitrary fields and allows to obtain bounds for a wider range of values of ℓ. The proof of this lemma relies on the new extension of the Weil bound we establish.

Given the lemma, the proof of Theorem 13.1 can be easily concluded. Recall that we showed that in order to prove local testability of an affine invariant code $\mathcal{T}(S)$, we need to show that there is a short codeword whose affine closure linearly spans $d\mathcal{T}(S)$. We showed that any $f \in d\mathcal{T}(S)$ for which this does not occur, is in fact contained in some $d\mathcal{T}(S \cup \{e\})$ for some $e \notin S$ of small weight. Thus, to conclude the proof we need to show that there exist small weight codewords in

$$d\mathcal{T}(S) \setminus \bigcup_{e \notin S : e \text{ has small weight}} d\mathcal{T}(S \cup \{e\}).$$

To this end we apply the tight bounds we obtain for the number of codewords of weight ℓ in dual-trace codes. We first show that if $\mathcal{C}$ is affine invariant of size $|\mathcal{C}| \leq p^{p^{O(n)}}$ then in fact $\mathcal{C} = d\mathcal{T}(S)$ where S is affine invariant of size $|S| \leq p^{O(n)}$, so our estimates for the number of codewords apply for $d\mathcal{T}(S)$. Fix a suitable weight ℓ. The number of codewords of weight ℓ in $d\mathcal{T}(S)$ is given by

$$W_\ell = \frac{C(p,\ell)}{\ell!} p^{n(\ell - |S'|)}(1 + o(1)),$$

where we recall that $S' = \{e \in S : (e, p) = 1\}$. On the other hand, as S is affine closed and $e \notin S$, we can bound the number of codewords of weight ℓ in any of the codes $d\mathcal{T}(S \cup \{e\})$ by

$$\leq \frac{C(p,\ell)}{\ell!} p^{n(\ell - |S'| - 1)}(1 + o(1)) \approx p^{-n} W_\ell.$$

Thus to conclude we just need to verify that the number of distinct e of small weight is $\ll p^n$. This then can be verified by a routine calculation.

13.1.3 New extension to the Weil bound

We sketch in high level how we achieve the new extension to the Weil bound. Let $f(x) = g(x) + h(x)$ be a univariate polynomial over $\mathbb{F}_{p^n}$, where $\deg(g) \leq |\mathbb{F}_{p^n}|^{1/2-\delta}$ and $h(x)$ is the sum of k monomials, each of weight degree at most d. We need to prove that either $\mathrm{Tr}(f) : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is a constant function, or that it is highly unbiased (note that proving the result for the Trace operator implies it immediately for all additive characters).

The analysis divides into two cases: either g has high weight-degree $\mathrm{wt}(g) \geq d + 1$, or g has low weight-degree $\mathrm{wt}(g) \leq d$. The first case is the easier one, and both cases rely on an analysis of directional derivatives of polynomials. The directional derivative of a polynomial $f(x)$ in direction $y \in \mathbb{F}_{p^n}$ is given by $f_y(x) = f(x + y) - f(x)$, and iterated derivatives are defined as $f_{y_1,\ldots,y_k}(x) = (f_{y_1,\ldots,y_{k-1}})_{y_k}(x)$.

The case of high weight g The first case, where $\mathrm{wt}(g) \geq d+1$ is easy to analyze by taking enough derivatives that eliminate $h(x)$, and reducing to a theorem of Deligne [Del78], which is a multivariate analog of Weil's bound. Specifically, For any $y_1, \ldots, y_{d+1}$ one can verify that since $\mathrm{wt}(h) \leq d$ then

$$h_{y_1,\ldots,y_{d+1}} \equiv 0,$$

hence $f_{y_1,\ldots,y_{d+1}} \equiv g_{y_1,\ldots,y_{d+1}}$. An iterated application of the Cauchy-Schwarz inequality yields that

$$\left|\mathbb{E}_{x\in\mathbb{F}_{p^n}}\left[\omega^{\mathrm{Tr}(f(x))}\right]\right|^{2^{d+1}} \leq \left|\mathbb{E}_{x,y_1,\ldots,y_{d+1}\in\mathbb{F}_{p^n}}\left[\omega^{\mathrm{Tr}(f_{y_1,\ldots,y_{d+1}}(x))}\right]\right|$$

where $\omega = e^{\frac{2\pi i}{p}}$. Hence to prove that $\mathrm{Tr}(f(x))$ in unbiased for uniform x, it is sufficient to prove that $\mathrm{Tr}(f_{y_1,\ldots,y_{d+1}}(x))$ is unbiased for uniform $x, y_1, \ldots, y_{d+1}$. We then verify that as g is of weight degree at least $d+1$, it is not eliminated by taking generic $d+1$ derivatives, and we get that $f_{y_1,\ldots,y_{d+1}}(x)$ is a nonzero polynomial in the variables $x, y_1, \ldots, y_{d+1}$ of total degree at most $\deg(g) \leq |\mathbb{F}_{p^n}|^{1/2-\delta}$. Moreover, we can prove that $\mathrm{Tr}(f_{y_1,\ldots,y_{d+1}}(x))$ is not a constant function; hence by Deligne's theorem we deduce that

$$\left|\mathbb{E}_{x,y_1,\ldots,y_{d+1}\in\mathbb{F}_{p^n}}\left[\omega^{\mathrm{Tr}(f_{y_1,\ldots,y_{d+1}}(x))}\right]\right| \leq |\mathbb{F}|^{-\delta}$$

and the bound on the bias of $\mathrm{Tr}(f(x))$ follows.

The case of low weight g The harder case is handling g of small weight $\mathrm{wt}(g) \leq d$, since h cannot simply be eliminated by taking enough iterated derivatives, without eliminating f altogether. We solve this problem by taking a smaller number of derivatives, such that f is not eliminated, but instead is transformed into a special class of polynomials (p-multilinear polynomials). We then proceed to study this family of polynomials, and are able to bound the bias of such polynomials, given that they came from a polynomial $f = g + h$ where g has low degree and h is the sum of a small number of low weight degree monomials. Most of the technical challenges of the proof are in this part.

13.1.4 Paper organization

We prove our main result, Theorem 13.1, on the local testing properties of affine invariant codes in Section 13.2. The proof uses our new extension to the Weil bound, which we prove in Section 13.3. Both sections are written in a self-contained manner, so that readers that are interested in the details of only one of these results can read only the relevant section. We note that throughout the paper we do not attempt to optimize constants.

13.2 Testing of affine invariant codes

We study affine invariant codes in this section. We begin with some definitions and stating our main theorem formally. We then proceed to prove some properties of affine invariant codes, and then apply those to prove our main result, Theorem 13.1.

13.2.1 Basic codes definitions

Let $\mathbb{F} = \mathbb{F}_{p^n}$ be a finite field. A code is a set of functions $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$. A code is called *linear* if it forms a linear space, i.e. if $f(x), g(x) \in \mathcal{C}$ then also $h(x) = \alpha f(x) + \beta g(x) \in \mathcal{C}$ where $\alpha, \beta \in \mathbb{F}_p$. We will only consider linear codes in this paper. For a linear code $\mathcal{C}$, its dual is the set functions which are normal to all codewords of $\mathcal{C}$.

Definition 13.5 (Dual code)**.** *Let $\mathcal{C} = \{f : \mathbb{F}_p^n \to \mathbb{F}_p\}$ be some linear code over $\mathbb{F}_p$. The dual code $\mathcal{C}^\perp$ is defined as*

$$\mathcal{C}^\perp = \left\{ (g : \mathbb{F}_p^n \to \mathbb{F}_p) : \sum_{x \in \mathbb{F}_p^n} f(x)g(x) = 0 \quad \forall f \in \mathcal{C} \right\}.$$

Note that the dual of the dual is the original code, i.e. $(\mathcal{C}^\perp)^\perp = \mathcal{C}$. We next define the weight and support of a codeword.

Definition 13.6 (Weight and support of codeword)**.** *The* support *of a codeword $f : \mathbb{F}_p^n \to \mathbb{F}_p$ is the set of $x \in \mathbb{F}_p^n$ for which $f(x) \neq 0$,*

$$\text{support}(f) = \{x \in \mathbb{F}_{p^n} : f(x) \neq 0\}.$$

The weight *of a codeword is the size of its support,*

$$\text{wt}(f) = |\text{support}(f)| = |\{x \in \mathbb{F}_{p^n} : f(x) \neq 0\}|.$$

13.2.2 Trace codes

Definition 13.7 (trace codes)**.** *Let $S \subseteq \{0, \ldots, p^n - 1\}$. The S-trace code is a code whose codewords are evaluations of functions $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$ given by*

$$\mathcal{T}(S) = \left\{ \left(\sum_{e \in S} \text{Tr}(\alpha_e x^e) : \mathbb{F}_{p^n} \to \mathbb{F}_p \right) : \alpha_e \in \mathbb{F}_p \right\},$$

where the Trace function $\text{Tr} : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is given by $\text{Tr}(x) = \sum_{i=0}^{n-1} x^{p^i}$.

For example, dual-BCH codes of weight t correspond to the special case

$$\text{dBCH}(t) = \mathcal{T}(\{1, 2, \ldots, t\}).$$

Generalized Reed-Muller codes over $\mathbb{F}_p^n$ of total degree d are equivalent to

$$\text{RM}(n, d) = \mathcal{T}(\{e \in \{0, \ldots, p^n - 1\} : \text{wt}(e) \leq d\}).$$

The following fact gives some simple properties of the Trace operator. For a proof, see any standard Algebra textbook, e.g. [BM65].

Fact 13.1 (Facts on the trace operator)**.** *Let $\text{Tr}(x) = \sum_{i=0}^{n-1} x^{p^i}$ be the trace operator over $\mathbb{F}_{p^n}$. Then*

1. *For any* $x \in \mathbb{F}_{p^n}$, $\mathrm{Tr}(x) \in \mathbb{F}_p$. *That is,* $\mathrm{Tr} : \mathbb{F}_{p^n} \to \mathbb{F}_p$.

2. *The trace operator is linear. That is, for any* $x, y \in \mathbb{F}_{p^n}$ *and* $a, b \in \mathbb{F}_p$ *we have*

$$\mathrm{Tr}(ax + by) = a\mathrm{Tr}(x) + b\mathrm{Tr}(y).$$

3. *The trace operator is invariant under the Frobenius map. That is, for any* $x \in \mathbb{F}_{p^n}$ *and* $0 \leq i \leq n-1$ *we have*

$$\mathrm{Tr}(x^{p^i}) = \mathrm{Tr}(x).$$

4. *Let* $x \in \mathbb{F}_{p^n}$, *and assume that for any* $\alpha \in \mathbb{F}_{p^n}$ *we have* $\mathrm{Tr}(\alpha x) = 0$. *Then* $x = 0$.

We denote the dual codeword to $\mathcal{T}(S)$ by $d\mathcal{T}(S) = \mathcal{T}(S)^{\perp}$. The following claim characterizes dual-trace codes.

Claim 13.1 (Characterization of dual-trace codes)**.** *Let* $S \subseteq \{0, \ldots, p^n - 1\}$. *Then*

$$d\mathcal{T}(S) = \left\{ (g : \mathbb{F}_{p^n} \to \mathbb{F}_p) : \sum_{x \in \mathbb{F}_{p^n}} g(x)x^e = 0 \quad \forall e \in S \right\}.$$

Proof. Let $g : \mathbb{F}_{p^n} \to \mathbb{F}_p$ be a function such that $\sum g(x)x^e = 0$ for all $e \in S$. We first verify that $g \in d\mathcal{T}(S)$. To do so, we need to show that $\sum_x f(x)g(x) = 0$ for any $f \in \mathcal{T}(S)$. Let $f = \sum_{e \in S} \mathrm{Tr}(\alpha_e x^e) \in \mathcal{T}(S)$. Then we have

$$\begin{aligned} \sum_{x \in \mathbb{F}_{p^n}} f(x)g(x) &= \sum_{x \in \mathbb{F}_{p^n}} \sum_{e \in S} \mathrm{Tr}(\alpha_e x^e)g(x) \\ &= \sum_{e \in S} \mathrm{Tr}(\alpha_e \sum_{x \in \mathbb{F}_{p^n}} x^e g(x)) = 0, \end{aligned}$$

where we used the fact that Trace is a linear operator over $\mathbb{F}_{p^n}$, thus $\mathrm{Tr}(ax+by) = a\mathrm{Tr}(x) + b\mathrm{Tr}(y)$ for any $a, b \in \mathbb{F}_p$ and $x, y \in \mathbb{F}_{p^n}$. Thus, to prove the claim we need to establish that for any $g \in d\mathcal{T}(S)$ and any $e \in S$ we have $\sum g(x)x^e = 0$. Note that for any $\alpha_e \in \mathbb{F}_{p^n}$ we have $f(x) = \alpha_e x^e \in \mathcal{T}(S)$, thus we have

$$\sum_{x \in \mathbb{F}_{p^n}} \mathrm{Tr}(\alpha_e x^e g(x)) = 0.$$

Let $z = \sum_{x \in \mathbb{F}_{p^n}} g(x)x^e$. We obtained that for any $\alpha_e \in \mathbb{F}_{p^n}$ we have

$$\mathrm{Tr}(\alpha_e z) = 0.$$

This can only hold if $z = 0$, thus we conclude that we must have that $\sum_x g(x)x^e = 0$ for all $e \in S$. □

The next claim shows that if $S_1 \subseteq S_2$ then $\mathcal{T}(S_1) \subseteq \mathcal{T}(S_2)$ and $d\mathcal{T}(S_1) \supseteq d\mathcal{T}(S_2)$.

Claim 13.2 (Monotonicity of trace codes)**.** *Let* $S_1 \subseteq S_2 \subseteq \{0, \ldots, p^n - 1\}$. *Then we have the following inclusions*

1. $\mathcal{T}(S_1) \subseteq \mathcal{T}(S_2)$.

2. $d\mathcal{T}(S_1) \supseteq d\mathcal{T}(S_2)$.

Proof. The claim follows immediately from the definition of trace codes and of dual codes. □

We will consider in the following few claims only trace codes for $S \subseteq \{1, \ldots, p^n - 1\}$, i.e. we disallow $0 \in S$. We will later also deal with sets containing 0. We now define irreducible degrees and reduced forms. We will see that it is enough to study trace codes over reduced form sets.

Definition 13.8 (Irreducible degrees and reduced form)**.** *We define R as the set of co-prime elements to p,*

$$R = \{1 \leq e \leq p^n - 1 : (e, p) = 1\}.$$

For $1 \leq e \leq p^n - 1$ define its reduced form $e' \in R$ *as follows. Let $e = p^k m$ where $(p, m) = 1$. Then the reduced form of e is $e' = m$. For a subset $S \subseteq \{1, \ldots, p^n - 1\}$ define its reduced form $S' \subseteq R$ as $S' = \{e' : e \in S\}$.*

Claim 13.3 (Trace codes are defined over reduce form sets)**.** *Let $S \subseteq \{1, \ldots, p^n - 1\}$. Let $S' \subseteq R$ be the reduced form of S. Then $d\mathcal{T}(S) = d\mathcal{T}(S')$ and $\mathcal{T}(S) = \mathcal{T}(S')$.*

Proof. By Claim 13.1 we have that $g \in d\mathcal{T}(S)$ iff $\sum g(x)x^e = 0$ for all $e \in S$. For any $0 \leq k \leq n - 1$ we have

$$\left(\sum g(x)x^e\right)^{p^k} = \sum g(x)x^{ep^k} = \sum g(x)x^{ep^k \pmod{p^n}},$$

where we used the facts that $x \to x^{p^k}$ is a linear map over $\mathbb{F}_{p^n}$, and that for any $x \in \mathbb{F}_{p^n}$ we have $x^{p^n} = x$. Hence we get that $\sum g(x)x^e = 0$ iff $\sum g(x)x^{e'} = 0$ for any e' such that $e' = ep^k \pmod{p^n}$. This shows that $d\mathcal{T}(S) = d\mathcal{T}(S')$, since for every element $e \in S$ there is some $e' = ep^k \pmod{p^n} \in S'$ and vice versa. Since $d\mathcal{T}(S) = d\mathcal{T}(S')$ we also get by the uniqueness of dual codes that $\mathcal{T}(S) = d\mathcal{T}(S)^\perp = d\mathcal{T}(S')^\perp = \mathcal{T}(S')$. □

The next claim establishes the size of trace codes defined over reduced form sets $S \subseteq R$.

Claim 13.4 (Size of trace codes)**.** *Let $S \subseteq \{1, \ldots, p^n - 1\}$. Let $S' \subseteq R$ be the reduced form of S. Then $|\mathcal{T}(S)| = p^{n|S'|}$.*

Proof. By Claim 13.3 we know that $\mathcal{T}(S) = \mathcal{T}(S')$. The codewords of $\mathcal{T}(S')$ are functions of the form

$$f(x) = \sum_{e \in S'} \mathrm{Tr}(\alpha_e x^e),$$

where $\alpha_e \in \mathbb{F}_{p^n}$. The number of combinations of $\{\alpha_e : e \in S'\}$ is $|\mathbb{F}_{p^n}|^{|S'|} = p^{n|S'|}$. Hence to conclude we need to show any two such settings are distinct. Since the code is linear, it is

enough to show that if the coefficients α_e are not all zero, then the codeword is not the all zeros codeword, i.e. there is some $x \in \mathbb{F}_{p^n}$ such that

$$\sum_{e \in S'} \mathrm{Tr}(\alpha_e x^e) \neq 0.$$

Let $p(x) = \sum_{e \in S'} \mathrm{Tr}(\alpha_e x^e)$, and note that

$$\begin{aligned} p(x) &= \sum_{e \in S'} \sum_{i=0}^{n-1} \alpha_e^{p^i} x^{ep^i} \\ &= \sum_{e \in S'} \sum_{i=0}^{n-1} \alpha_e^{p^i} x^{ep^i \pmod{p^n}}, \end{aligned}$$

where we used the facts that $\mathrm{Tr}(x) = \sum_{i=0}^{n-1} x^{p^i}$ as well as the identity $x^t = x^{t \pmod{p^n}}$ which holds for any t. Since $S' \subseteq R$ is a set of

all the monomials x^{ep^i} for $e \in S'$ are disjoint. Hence $p(x)$ is not the all zeros polynomial. As $\deg(p) \leq p^n - 1$ there must exist some $x \in \mathbb{F}_{p^n}$ such that $p(x) \neq 0$, and the codeword defined by f is not the all zeros codeword. □

13.2.3 Characterization of affine invariant codes by trace codes

We start by recalling *affine invariant codes*, which are codes that are closed under an affine transformation of the input space coordinates.

Definition 13.9 (Affine closure, and affine invariant codes)**.** *Let $f : \mathbb{F}_{p^n} \to \mathbb{F}_p$ be a function. The* affine closure *of f is the set of functions*

$$\overline{\mathrm{affine}}(f) = \left\{ (f(ax+b) : \mathbb{F}_{p^n} \to \mathbb{F}_p) : a, b \in \mathbb{F}_{p^n} \right\}.$$

A code $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ is called affine invariant *if for any $f \in \mathcal{C}$, we have $\overline{\mathrm{affine}}(f) \subseteq \mathcal{C}$. A codeword $f \in \mathcal{C}$* affinely generates *$\mathcal{C}$ if*

$$\mathcal{C} = \mathrm{Span}(\overline{\mathrm{affine}}(f)).$$

We can characterize linear codes which are affine invariant as a special subfamily of trace codes. To this end we will require some definitions. We first define shift closure of a set, which is tightly related to the reduced form we previously defined.

Definition 13.10 (Shift closed)**.** *Let $e \in \{0, \ldots, p^n - 1\}$. The* shift closure *of e is defined as the set*

$$\overline{\mathrm{shift}}(e) = \{ep^\ell \ (mod\ p^n) : \ell = 1, \ldots, n\}.$$

The shift closure of a set $S \subseteq \{0, \ldots, p^n - 1\}$ is defined as the union of the shift closures of its elements,

$$\overline{\mathrm{shift}}(S) = \cup_{e \in S} \overline{\mathrm{shift}}(e).$$

A set $S \subseteq \{0, \ldots, p^n - 1\}$ is said to be shift closed *if $S = \overline{\mathrm{shift}}(S)$.*

The term *shift closed* comes from viewing elements $e \in S$ as vectors in $\mathbb{F}_p^n$, given by the representation of e in base p. In this case, $ep^\ell \pmod{p^n}$ corresponds to a cyclic shift of the vector by ℓ coordinates. The following claim shows that trace codes are invariant under shift closure.

Claim 13.5. *Let $S \subseteq \{0, \ldots, p^n - 1\}$. Then*

$$d\mathcal{T}(S) = d\mathcal{T}(\overline{\text{shift}}(S)), \quad \mathcal{T}(S) = \mathcal{T}(\overline{\text{shift}}(S)).$$

Proof. The proof is identical to the proof of Claim 13.3. □

We next define the notion of shadow closed sets.

Definition 13.11 (Shadow closed)**.** *Let $S \subseteq \{0, \ldots, p^n - 1\}$. The set S is said to be* shadow closed *if the following holds. For any $e \in S$, let $e = \sum_{i=0}^{n-1} e_i p^i$ be the representation of e in base p. Define the* support *of e to be the set of nonzero digits of e,*

$$\text{support}(e) = \{0 \leq i \leq n-1 : e_i \neq 0\}.$$

Let e' be obtained from e by changing some of the non-zero digits of e, i.e.

$$e' = \sum_{i \in \text{support}(e)} e'_i p^i.$$

Then we should have that also $e' \in S$. That is, S is shadow closed if

$$\left\{ \sum_{i \in \text{support}(e)} e'_i p^i : e \in S, (e'_i)_{i \in \text{support}(e)} \in \mathbb{F}_p \right\} \subseteq S.$$

Definition 13.12 (Affine closed)**.** *A set $S \subseteq \{0, \ldots, p^n - 1\}$ is* affine closed *if it is both shift closed and shadow closed.*

We recall the following theorem of Kaufman and Sudan [KS08] that we presented in the introduction. It shows that affine invariant linear codes are equivalent to trace codes over affine closed sets.

Theorem (Theorem 13.5: Equivalence of affine invariant codes and trace codes of affine closed sets)**.** *Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be an affine invariant linear code. Then there exists an affine closed set $S \subseteq \{0, \ldots, p^n - 1\}$ such that $\mathcal{C} = \mathcal{T}(S)$. Moreover, for any affine closed set S the code $\mathcal{T}(S)$ is linear and affine invariant.*

13.2.4 Weight distribution of affine invariant codes

Theorem 13.5 tells us that in order to study affine invariant codes, it suffices to study trace codes of affine closed sets. In this subsection we establish the following lemma, which gives a tight estimate on the number of codewords in $d\mathcal{T}(S)$ for affine closed sets S. For the statement of the lemma recall that $R = \{1 \leq e \leq p^n - 1 : (e, p) = 1\}$ is the set of elements co-prime to p.

Lemma 13.1 (Weight distribution of dual trace affine closed codes)**.** *There exist absolute constants $c, c' > 1$ such that the following is true. Let $S \subseteq \{0, \ldots, p^n - 1\}$ be affine closed of size $|S| \leq \frac{1}{c'} p^{n/c}$. Then there exists $\ell_{\min} = c'|S \cap R|^c$ and $\ell_{\max} = \frac{1}{c'} p^{n/c}$, such that for any $\ell_{\min} \leq \ell \leq \ell_{\max}$ the following holds. The number of codewords in $d\mathcal{T}(S)$ of weight exactly ℓ is given by*

$$\frac{C(p,\ell)}{\ell!} p^{n(\ell - |S \cap R|)}(1+\epsilon)$$

where $C(p,\ell)$ is defined as

$$C(p,\ell) = \left| \left\{ (v_1, \ldots, v_\ell) \in (\mathbb{F}_p \setminus \{0\})^\ell : v_1 + \ldots + v_\ell = 0 \right\} \right|.$$

and $|\epsilon| \leq p^{-n/2} \ll 1$. In particular, one can take $c = 8$ and $c' = 16$.

We start by showing a general bound on the weight degree of elements of affine closed sets, in terms of the size of the set.

Claim 13.6 (Weight degree bound on affine closed sets)**.** *Let $S \subseteq \{0, \ldots, p^n - 1\}$ such that S is affine closed. Then for any $e \in S$,*

$$\mathrm{wt}(e) \leq \log_p |S \cap R| + 1.$$

Proof. Let $S' = S \cap R$. Let $e \in S$ be of weight $k \geq 1$. By taking some shift of e we may assume $e \in R$ (that is, $0 \in \mathrm{support}(e)$), hence $e \in S' = S \cap R$. Consider the set

$$E' = \left\{ \sum_{i \in \mathrm{support}(e)} e'_i p^i : e'_i \in \mathbb{F}_p,\ e'_0 \neq 0 \right\}.$$

Note that as S is shadow closed, we have $E' \subseteq S$. Moreover since $e'_0 \neq 0$ we have $E' \subseteq R$, hence $E' \subseteq S' = S \cap R$. Thus $|E'| \leq |S'|$. On the other hand,

$$|E'| = (p-1)p^{\mathrm{wt}(e)-1}.$$

Hence we conclude that $\mathrm{wt}(e) \leq \log_p(\frac{p}{p-1}|S'|) \leq \log_p |S'| + 1$. □

We will need the following simple claim.

Claim 13.7 (Trace is not constant)**.** *Let $f(x) = \sum_{e \in R} \alpha_e x^e$ be a nonzero polynomial. Then $\mathrm{Tr}(f(x)) : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is not a constant function.*

Proof. Assume for contradiction that $\mathrm{Tr}(f(x)) = a$ for all $x \in \mathbb{F}_{p^n}$. Let $q(x) = \mathrm{Tr}(f(x)) - a$. We have

$$q(x) = -a + \sum_{i=0}^{n-1} (\sum_{e \in R} \alpha_e x^e)^{p^i} = -a + \sum_{i=0}^{n-1} \sum_{e \in R} (\alpha_e)^{p^i} x^{ep^i \pmod{p^n}}.$$

Since $e \in R$ all the degrees $ep^i \pmod{p^n}$ are distinct and different from 0. Thus $q(x)$ is not the zero polynomial. Since $\deg(q) \leq p^n - 1$ we have that there must be x such that $q(x) \neq 0$, hence $\mathrm{Tr}(f(x)) \neq a$. □

The next lemma is a general lemma, which estimates the number of elements in $d\mathcal{T}(S)$ where S is a relatively small set of elements of small weight degree. We will then show that the lemma can be applied to any affine invariant set S which is not too large.

Lemma 13.2 (Weight distribution of dual trace codes of reduced form sets)**.** *There exists an absolute constant $c > 1$ such that the following is true. Let $S \subseteq R$ be such that for any $e \in S$ its weight degree is at most $\mathrm{wt}(e) \leq d$. There exist $\ell_{\min} = c|S|^2 d^2 2^d$ and $\ell_{\max} = p^{n/c}$, such that for any $\ell_{\min} \leq \ell \leq \ell_{\max}$ the following holds.*

1. *The number of codewords in $d\mathcal{T}(S)$ of weight exactly ℓ is given by*

$$\frac{(p-1)^\ell}{\ell!} p^{n(\ell-|S|)}(1+\epsilon).$$

   where $|\epsilon| \leq p^{-n/2} \ll 1$.

2. *The number of codewords in $d\mathcal{T}(S \cup \{0\})$ of weight exactly ℓ is given by*

$$\frac{C(p,\ell)}{\ell!} p^{n(\ell-|S|)}(1+\epsilon).$$

   where $|\epsilon| \leq p^{-n/2}$ and $C(p,\ell)$ is defined as

$$C(p,\ell) = \left| \left\{ (v_1, \ldots, v_\ell) \in (\mathbb{F}_p \setminus \{0\})^\ell : v_1 + \ldots + v_\ell = 0 \right\} \right|.$$

In particular, one can take $c = 8$.

Proof. We start by proving the estimate for $d\mathcal{T}(S)$. For any $v = (v_1, \ldots, v_\ell) \in \{1, \ldots, p-1\}^\ell$ define the sets

$$A_\ell(v) = \{(\alpha_1, \ldots, \alpha_\ell) \in \mathbb{F}_{p^n}^\ell : \sum_{i=1}^{\ell} v_i \alpha_i^e = 0 \quad \forall e \in S\}$$

and

$$B_\ell(v) = \{(\alpha_1, \ldots, \alpha_\ell) \in A_\ell(v) : \alpha_1, \ldots, \alpha_\ell \text{ are all distinct}\}.$$

Let $f : \mathbb{F}_p^n \to \mathbb{F}_p$ be a function $f \in d\mathcal{T}(S)$, such that f has weight exactly ℓ. Equivalently, there are distinct points $\alpha_1, \ldots, \alpha_\ell \in \mathbb{F}_{p^n}$ such that $\sum f(\alpha_i)\alpha_i^e = 0$ for all $e \in S$. We can identify f uniquely by the list of points $(\alpha_1, \ldots, \alpha_\ell)$ and the evaluation of f on these points $v = (f(\alpha_1), \ldots, f(\alpha_\ell)) \in \{1, \ldots, p-1\}^\ell$. Since the order of $\alpha_1, \ldots, \alpha_\ell$ does not matter, and they are all distinct, there are $\ell!$ elements in $\cup B_\ell(v)$ which correspond to f, (i.e. these elements correspond to all orderings of $\alpha_1, \ldots, \alpha_\ell$). Thus we obtain the following identity,

$$\text{Number of codewords in } d\mathcal{T}(S) \text{ of weight } \ell = \frac{1}{\ell!} \sum_{v \in \{1,\ldots,p-1\}^\ell} |B_\ell(v)|.$$

Hence, to conclude the proof we will show that $|B_\ell(v)| \approx p^{n(\ell-|S|)}$. In fact, we will first show that $|A_\ell(v)| \approx p^{n(\ell-|S|)}$ and then deduce the estimate for $|B_\ell(v)|$.

Fix some $v \in \{1, \ldots, p-1\}^\ell$. We will now show an estimate on $|A_\ell(v)|$, where the main tool we use is Fourier analysis. Let $\alpha = (\alpha_e : e \in S) \in \mathbb{F}_{p^n}^S$, and define $\phi_\alpha : \mathbb{F}_{p^n} \to \mathbb{F}_p$ by

$$\phi_\alpha(x) = \mathrm{Tr}(\sum_{e \in S} \alpha_e x^e).$$

Take any tuple $(x_1, \ldots, x_\ell) \in \mathbb{F}_{p^n}^\ell$, and consider

$$\mu(x_1, \ldots, x_\ell) = \mathbb{E}_{\alpha \in \mathbb{F}_{p^n}^S} \left[\omega^{v_1 \phi_\alpha(x_1) + \ldots + v_\ell \phi_\alpha(x_\ell)} \right],$$

where $\omega = e^{\frac{2\pi i}{p}}$ is a p-root of unity. We claim that if $(x_1, \ldots, x_\ell) \in A_\ell(v)$ then $\mu(x_1, \ldots, x_\ell) = 1$, and if $(x_1, \ldots, x_\ell) \notin A_\ell(v)$ then $\mu(x_1, \ldots, x_\ell) = 0$. To see that,

$$\begin{aligned}
\mu(x_1, \ldots, x_\ell) &= \mathbb{E}_{\alpha \in \mathbb{F}_{p^n}^S} \left[\omega^{\mathrm{Tr}(\sum_{e \in S} \alpha_e (v_1 x_1^e + \ldots + v_\ell x_\ell^e))} \right] \\
&= \prod_{e \in S} \mathbb{E}_{\alpha_e \in \mathbb{F}_{p^n}} \left[\omega^{\mathrm{Tr}(\alpha_e (v_1 x_1^e + \ldots + v_\ell x_\ell^e))} \right] \\
&= \prod_{e \in S} \mathbf{1}_{v_1 x_1^e + \ldots + v_\ell x_\ell^e = 0} = \mathbf{1}_{(x_1, \ldots, x_\ell) \in A_\ell(v)}.
\end{aligned}$$

Hence we have

$$\begin{aligned}
|\mathbb{F}_{p^n}|^{-\ell} |A_\ell(v)| &= \mathbb{E}_{x_1, \ldots, x_\ell \in \mathbb{F}_{p^n}} \left[\mu(x_1, \ldots, x_\ell) \right] \\
&= \mathbb{E}_{x_1, \ldots, x_\ell \in \mathbb{F}_{p^n}} \mathbb{E}_{\alpha \in \mathbb{F}_{p^n}^S} \left[\omega^{\mathrm{Tr}(\sum_{e \in S} \alpha_e (v_1 x_1^e + \ldots + v_\ell x_\ell^e))} \right] \\
&= \mathbb{E}_{\alpha \in \mathbb{F}_{p^n}^S} \prod_{i=1}^{\ell} \mathbb{E}_{x_i \in \mathbb{F}_{p^n}} \left[\omega^{\mathrm{Tr}(\sum_{e \in S} \alpha_e v_i x_i^e)} \right]
\end{aligned}$$

We partition the expectation to the cases where $\alpha = 0^S$ and $\alpha \neq 0^S$. When $\alpha = 0^S$ then for all $i = 1, \ldots, \ell$ we have that

$$\mathbb{E}_{x_i \in \mathbb{F}_{p^n}} \left[\omega^{\mathrm{Tr}(\sum_{e \in S} \alpha_e v_i x_i^e)} \right] = 1.$$

Consider now any $\alpha \neq 0^S$ and any $i = 1, \ldots, \ell$. As $v_i \in \mathbb{F}_p \setminus \{0\}$ then also $\alpha v_i \neq 0^S$. We will show that $\mathrm{Tr}(\sum_{e \in S} \alpha_e v_i x_i^e)$ has small bias . To this end we apply Theorem 13.3. Let $f(x) = g(x) + h(x)$ for $g(x) = 0$ and $h(x) = \sum_{e \in S} \alpha_e v_i x_i^e$. As $S \subseteq R$ and not all $\alpha_e = 0$, we have by Claim 13.7 that $\mathrm{Tr}(f)$ is not constant. Our condition on the set S was that $\mathrm{wt}(e) \leq d$ for any $e \in S$. Hence we get by Theorem 13.3 (for $\delta = 1/2$) that

$$\left| \mathbb{E}_{x \in \mathbb{F}_{p^n}} \left[\omega^{\mathrm{Tr}(\sum_{e \in S} \alpha_e v_i x^e)} \right] \right| \leq |\mathbb{F}_{p^n}|^{-\frac{1}{4|S| d^2 2^d}}.$$

Hence we deduce that

$$|A_\ell(v)| = |\mathbb{F}_{p^n}|^{\ell - |S|} (1 + \epsilon)$$

where $|\epsilon| \leq |\mathbb{F}_{p^n}|^{|S| - \ell \cdot \frac{1}{4|S| d^2 2^d}}$. Thus, if we take $\ell \geq 8|S|^2 d^2 2^d$ we get that $|\epsilon| \leq p^{-n|S|} \leq p^{-n} \ll 1$.

To conclude, we need to derive an estimate on $|B_\ell(v)|$. Let $C_\ell(v) = A_\ell(v) \setminus B_\ell(v)$. We will show that $|C_\ell(v)| \ll |B_\ell(v)|$, and hence $|B_\ell(v)| \approx |A_\ell(v)|$. To derive this, note that if $(\alpha_1, \ldots, \alpha_\ell) \in C_\ell(v)$, then $\alpha_1, \ldots, \alpha_\ell$ are not all distinct, that is, $\alpha_i = \alpha_j$ for some distinct $i < j$. Define $v^{(i,j)} \in \{1, \ldots, p-1\}^{\ell-1}$ by "joining" α_i and α_j, i.e. $v_a^{(i,j)} = v_a$ for $1 \leq a < i$ and $i < a < j$, $v_i^{(i,j)} = v_i + v_j$, $v_a^{(i,j)} = v_{a+1}$ for $a > j$. Then we can identify uniquely $(\alpha_1, \ldots, \alpha_\ell) \in C_\ell(v)$ with $\alpha^{(i,j)} = (\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_\ell) \in A_{\ell-1}(v^{(i,j)})$. Hence we get

$$|C_\ell(v)| \leq \sum_{i<j} |A_{\ell-1}(v^{i,j})| \leq \binom{\ell}{2} |A_{\ell-1}(\cdot)| \leq \ell^2 |\mathbb{F}_{p^n}|^{\ell-1-|S|}(1+\epsilon) = \frac{\ell^2}{p^n}|A_\ell(v)|(1+\epsilon).$$

Hence we get that

$$|B(v)| = |\mathbb{F}_{p^n}|^{\ell-|S|}(1+\epsilon')$$

where $\epsilon' = \frac{\ell^2}{p^n} + \epsilon$. Thus if $\ell \leq p^{n/8}$ we get that $\frac{\ell^2}{p^n} \ll p^{-n/2}$. Hence we finished the proof of the first claim.

The proof of the second claim is completely analogous, except if we consider $d\mathcal{T}(S \cup \{0\})$, we have that additional requirement that $v_1 + \ldots + v_\ell = 0$. Thus one should not consider $A_\ell(v)$ for all $v \in (\mathbb{F}_p \setminus \{0\})^\ell$, but only those corresponding to $v \in C(p, \ell)$. Thus we have

$$\text{Number of codewords in } d\mathcal{T}(S \cup \{0\}) \text{ of weight } \ell = \frac{1}{\ell!} \sum_{v \in C(p,\ell)} |B_\ell(v)|.$$

and the proof follows by the estimates we proved on $|B_\ell(v)|$. □

We can now deduce Lemma 13.1 from Claim 13.6 and Lemma 13.2.

Proof of Lemma 13.1. Let $S \subseteq \{0, \ldots, p^n - 1\}$ be affine closed. We have that

$$d\mathcal{T}(S) = d\mathcal{T}((S \cap R) \cup \{0\}).$$

By Claim 13.6 the maximal weight of elements in S is at most

$$d \leq \log_p |S \cap R| + 1.$$

Applying Lemma 13.2, we get that for $\ell_{\min} = 16 \cdot |S \cap R|^4 \geq 8|S \cap R|^2 d^2 2^d$ and $\ell_{\max} = \frac{1}{16} p^{n/4}$, we get that for every $\ell_{\min} \leq \ell \leq \ell_{\max}$ the number of codewords of weight ℓ in $d\mathcal{T}(S) = d\mathcal{T}((S \cap R) \cup \{0\})$ is

$$\frac{C(p,\ell)}{\ell!} p^{n(\ell-|S\cap R|)}(1+\epsilon)$$

where $|\epsilon| \leq p^{-n/2}$. □

13.2.5 Trace codes of exponential size are generated by a single orbit

We prove in this subsection that any affine invariant linear code of up to exponential size is generated by a single orbit of a dual codeword. Combining this with Theorem 13.6 we get that any such code is locally testable, which prove our main result, Theorem 13.1. We now state the main theorem we prove in this subsection.

Theorem 13.7 (Affine invariant codes are generated by a single orbit)**.** *There exist absolute constants $0 < \alpha < 1$ and $c, c' \geq 1$ such that the following is true. Let $\mathcal{C} = \{f : \mathbb{F}_{p^n} \to \mathbb{F}_p\}$ be an affine invariant linear code, such that $\dim(\mathcal{C}) \leq \frac{1}{c'} p^{\alpha n}$. Then there exists $f \in \mathcal{C}^\perp$ such that*

$$\overline{\text{affine}}(f)^\perp = \mathcal{C}$$

and of weight

$$\text{wt}(f) \leq c'(\dim(\mathcal{C})/n)^c.$$

In particular, one may choose $\alpha = 1/16$, $c = 4$ and $c' = (2p)^8$.

Let $\mathcal{C} = \mathcal{T}(S)$ be an affine invariant code where $S \subseteq \{0, \ldots, p^n - 1\}$ is affine closed. We start by showing that if some $f \in \mathcal{C}^\perp = d\mathcal{T}(S)$ does not generate $d\mathcal{T}(S)$, then in fact $f \in d\mathcal{T}(S \cup \{e\})$ where $e \in \{1, \ldots, p^n - 1\} \setminus S$ has small weight (Corollary 13.1). From this and the exact estimates for the weight distribution for dual trace codes we derive Theorem 13.7. Before proving Corollary 13.1 we will require two technical claims.

Claim 13.8. *Let $S \subseteq \{0, \ldots, p^{n-1}\}$ be affine closed. Let $f \in d\mathcal{T}(S)$ be a codeword which does not affinely generate $d\mathcal{T}(S)$, i.e.*

$$\overline{\text{affine}}(f) \subsetneq d\mathcal{T}(S).$$

Then

$$\overline{\text{affine}}(f) = d\mathcal{T}(T)$$

for some affine closed $T \supsetneq S$.

Proof. The code $\overline{\text{affine}}(f)$ is an affine invariant code which is a proper subset of $d\mathcal{T}(S)$. By Theorem 13.5 we know that $\overline{\text{affine}}(f) = d\mathcal{T}(T)$ for some affine closed $T \subseteq \{0, \ldots, p^n - 1\}$. Since $d\mathcal{T}(T) \subsetneq d\mathcal{T}(S)$ we must have that $T \supsetneq S$. □

Claim 13.9. *Let $S \subsetneq T \subseteq \{0, \ldots, p^n - 1\}$ such that both S and T are affine closed. Then there exist an element $e \in (T \setminus S) \cap R$ such that*

$$\text{wt}(e) \leq \log_p |S \cap R| + 2.$$

Proof. Let $S' = S \cap R$ and $T' = T \cap R$. We have $S' \subsetneq T'$ as otherwise, if $S' = T'$, we would have $S = \overline{\text{affine}}(S') = \overline{\text{affine}}(T') = T$.

Let $k = \lfloor \log_p |S'| \rfloor + 2$. We argue there is $e \in T' \setminus S'$ of weight at most k. Otherwise, let $e \in S' \setminus T'$ such that $\text{wt}(e) > k$. Consider the set

$$E = \overline{\text{shadow}}(e) \cap R = \left\{ \sum_{i \in \text{support}(e)} e'_i p^i : e'_i \in \mathbb{F}_p,\ e'_0 \neq 0 \right\},$$

where we use the fact that since $e \in R$ then $0 \in \text{support}(e)$. Note that by definition, $E \subseteq T'$, since T is affine closed hence in particular shadow closed.

Let $e' \in E \subseteq T'$ such that $\mathrm{wt}(e') = k$ (by setting $\mathrm{wt}(e) - k$ digits of e in base p to zero). Consider the set

$$E' = \overline{\mathrm{shadow}}(e') \cap R = \left\{ \sum_{i \in \mathrm{support}(e')} e''_i p^i : e''_i \in \mathbb{F}_p,\ e''_0 \neq 0 \right\}.$$

Note that since $|E'| = (p-1)p^{\mathrm{wt}(e')-1} = (p-1)p^{k-1} > |S'|$ we cannot have that $e' \in S'$. Hence we found an element $e' \in T' \setminus S'$ such that $\mathrm{wt}(e') \leq k$. □

Corollary 13.1. *Let $S \subseteq \{0, \ldots, p^{n-1}\}$ be affine closed. Let $f \in d\mathcal{T}(S)$ be a codeword which does not affinely generate $d\mathcal{T}(S)$, i.e.*

$$\overline{\mathrm{affine}}(f) \subsetneq d\mathcal{T}(S).$$

Then there must exist $e \in R \setminus S$ of weight $\mathrm{wt}(e) \leq \log_p |S \cap R| + 2$ such that

$$f \in d\mathcal{T}(S \cup \{e\}).$$

Proof. By Claim 13.8 we have $\overline{\mathrm{affine}}(f) = d\mathcal{T}(T)$ where $T \supsetneq S$. By Claim 13.9 there is $e \in (T \setminus S) \cap R \subseteq R \setminus S$ such that $\mathrm{wt}(e) \leq \log_p |S \cap R| + 2$. Hence we conclude sicne

$$f \in d\mathcal{T}(T) \subseteq d\mathcal{T}(S \cup \{e\}).$$

□

We are now ready to prove Theorem 13.7.

Proof of Theorem 13.7. Let $\mathcal{C}$ be a linear affine invariant code. By theorem 13.5 we have $\mathcal{C} = \mathcal{T}(S)$ where $S \subseteq \{0, \ldots, p^n - 1\}$ is affine closed. By Claims 13.2, 13.3 and 13.4 we have that

$$|\mathcal{C}| = \mathcal{T}((S \cap R) \cup \{0\}) \leq |\mathcal{T}(S \cap R)| = p^{n|S \cap R|}.$$

Hence we need to prove there is a codeword $f \in d\mathcal{T}(S)$ of weight $|S \cap R|^c$ whose affine closure spans $d\mathcal{T}(S)$. Let ℓ be an appropriate weight to be determined later. We now count the number of codewords in $d\mathcal{T}(S)$ of weight exactly ℓ. To this end we apply Lemma 13.1. The number of codewords in $d\mathcal{T}(S)$ of weight ℓ (as long as ℓ is in the permissible range) is given by

$$W_\ell = \frac{C(p,\ell)}{\ell!} p^{n(\ell - |S \cap R|)}(1 + p^{-\Omega(n)}).$$

Let $f \in d\mathcal{T}(S)$ be such that $\overline{\mathrm{affine}}(f) \subsetneq d\mathcal{T}(S)$. By Corollary 13.1 we know that there exists some $e \in R \setminus S$ of weight $\mathrm{wt}(e) \leq k$, where $k \leq \log_p(|S \cap R|) + 2$, such that $f \in d\mathcal{T}(S \cup \{e\})$. Let E be the set of all such possible e,

$$E = \{e \in R \setminus S : \mathrm{wt}(e) \leq k\}.$$

Fix some $e \in E$. Let $S_e = \overline{\text{affine}}(S \cup \{e\})$. Note that as $e \in R \setminus S$ we have $|S_e \cap R| \geq |S \cap R| + 1$. Hence for ℓ in the permissible range for S_e we get that the number of codewords of weight ℓ in $d\mathcal{T}(S_e)$ is given by

$$\frac{C(p,\ell)}{\ell!} p^{n(\ell - |S_e \cap R|)}(1 + p^{-\Omega(n)}) \leq p^{-n} W_\ell (1 + p^{-\Omega(n)}),$$

So, as long as $|E| \ll p^n$, we can deduce that there must exist some $f \in d\mathcal{T}(S)$ of weight ℓ which is not in any of $d\mathcal{T}(S \cup \{e\})$ for any $e \in E$ (in fact, almost all $f \in d\mathcal{T}(S)$ of weight ℓ will do). This will establish the theorem. Thus, we need to bound $|E|$. The following is a simple bound which is sufficient for our needs.

$$|E| \leq \sum_{i=1}^{k} \binom{n}{i} p^i \leq p^{3n/4}$$

as long as $k \leq n/4$.

To conclude we need to show that we can choose ℓ such that $\ell \leq |S \cap R|^c$ for some absolute constant $c > 0$, as long as $|S \cap R| \leq p^{\alpha n}$ for some absolute constant $\alpha > 0$. The bounds on $\ell_{\min}$ and $\ell_{\max}$ that are required for the application Lemma 13.1 are stricter for S_e than for S, and are given by

$$\begin{aligned} |S_e| &\leq \tfrac{1}{16} p^{n/4}, \\ \ell_{\min} &\geq 16 |S_e \cap R|^4, \\ \ell_{\max} &\leq \tfrac{1}{16} p^{n/4}. \end{aligned}$$

To verify them we need to give an upper bound on $|S_e|$ and $|S_e \cap R|$. As $S_e = S \cup \overline{\text{affine}}(\{e\})$ we have

$$\begin{aligned} |S_e| &\leq |S| + |\overline{\text{affine}}(\{e\})| = |S| + np^k, \\ |S_e \cap R| &\leq |S \cap R| + |\overline{\text{affine}}(\{e\}) \cap R| \leq |S \cap R| + p^k. \end{aligned}$$

Note that $p^k = p^2 |S \cap R|$. Thus, the bounds for applying Lemma 13.1 are satisfied if we make sure that

$$\begin{aligned} |S| &\leq \tfrac{1}{32p^2 n} p^{n/4}, \\ \ell_{\min} &\geq (2p)^8 |S \cap R|^4, \\ \ell_{\max} &\leq \tfrac{1}{16} p^{n/4}. \end{aligned}$$

Notice that as long as $|S| \leq \frac{1}{16p^3} p^{n/16}$ we have that all the conditions are satisfied (for large enough n) and that $\ell_{\min} \leq \ell_{\max}$. Hence we may choose $\ell = \ell_{\min}$ to conclude the proof. $\square$

13.3 Extension of the Weil bound

In this section we prove our new extension to the Weil bound for character sums, which is one of the key technical ingredients in our proof of the local testability of affine invariant codes. As this result may be of independent interest, this section is self-contained, and the interested reader may read this section without relying on Section 13.2.

We recall several definitions and theorems from the introduction, for the sake of self containment. Let $\mathbb{F} = \mathbb{F}_{p^n}$ be a finite field. An additive character $\chi : \mathbb{F} \to \mathbb{C}$ is a mapping such that $\chi(x+y) = \chi(x)\chi(y)$ and χ is not identically zero. The following is a classical result by Weil.

Theorem (Weil bound - Theorem 13.2)**.** *Let $f(x)$ be a univariate polynomial over $\mathbb{F}$ of degree $|\mathbb{F}|^{1/2-\delta}$. Let $\chi : \mathbb{F} \to \mathbb{C}$ be an additive character. Then either $\chi(f(x))$ is constant or*

$$|\mathbb{E}_{x\in\mathbb{F}}[\chi(f(x))]| \leq |\mathbb{F}|^{-\delta}.$$

The *weight degree* of a monomial x^t is defined as follows. Let $t = \sum_{i=0}^{n-1} a_i p^i$ be the representation of t in base p, where $0 \leq a_i \leq p-1$. The weight degree of x^t is defined to be $wt(x^t) = \sum a_i$. The weight degree of a polynomial $f(x)$ is the maximal weight of a monomial in f.

Note 13.1. *We note that the weight degree of a polynomial can be equivalently defined also as a* derivative degree*, defined as follows. The directional derivative of $f(x)$ in direction $y \in \mathbb{F}_{p^n}$ is defined as $f_y(x) = f(x+y) - f(x)$. Define iterative derivatives in directions $y_1, \ldots, y_k$ as $f_{y_1,\ldots,y_k} = (f_{y_1,\ldots,y_{k-1}})_{y_k}$. The* derivative degree *of f is the minimal d such that for any $d+1$ derivatives $y_1, \ldots, y_{d+1} \in \mathbb{F}$, $f_{y_1,\ldots,y_{d+1}}(x) \equiv 0$. It can be verified that the derivative degree of a polynomial is exactly its weight degree. We do not prove this here, and will not require this fact in the proof.*

We prove an extension of the Weil bound in case f is the sum of a low degree polynomial and a small number of monomials of bounded weight (but of arbitrary degree).

Theorem (Extension of the Weil bound - Theorem 13.3)**.** *Let $f(x) = g(x) + h(x)$ be a univariate polynomial over $\mathbb{F}_{p^n}$, where $g(x)$ is a polynomial of degree $|\mathbb{F}|^{1/2-\delta}$ and $h(x)$ is the sum of at most $k \geq 1$ monomials, each of weight degree at most d. Let $\chi : \mathbb{F}_{p^n} \to \mathbb{C}$ be an additive character. Then either $\chi(f(x))$ is constant or*

$$|\mathbb{E}_{x\in\mathbb{F}}[\chi(f(x))]| \leq |\mathbb{F}|^{-\frac{\delta}{2kd^2 2^d}}.$$

13.3.1 Technical claims

In this subsection we provide some technical claims that will be needed for the proof of Theorem 13.3.

The trace operator

The trace operator $Tr : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is defined as $Tr(x) = \sum_{i=0}^{n-1} x^{p^i}$. We give in this subsection some simple properties of the Trace operator.

Claim 13.10 (Characterization of additive characters)**.** *Let $\chi : \mathbb{F}_{p^n} \to \mathbb{C}$ be an additive character. Then there exists $a \in \mathbb{F}_{p^n}$ such that $\chi(x) \equiv \omega^{\mathrm{Tr}(ax)}$ where $\omega = e^{2\pi i/p}$.*

Proof. We first prove that $\chi(x) = \omega^{\ell(x)}$ where $\ell : \mathbb{F}_{p^n} \to \mathbb{F}_p$ is a linear map. Note that we must have $\chi(0) = 1$ since $\chi(0) = \chi(0+0) = \chi(0)^2$, and we cannot have $\chi(0) = 0$ as this will imply that $\chi \equiv 0$. Thus, we get that the image of χ is a p-th root of unity since $\chi(x)^p = \chi(px) = \chi(0) = 1$. Thus we can write $\chi(x) = \omega^{\ell(x)}$ for some mapping $\ell : \mathbb{F}_{p^n} \to \mathbb{F}_p$. The mapping ℓ is linear since

$$\omega^{\ell(x+y)} = \chi(x+y) = \chi(x)\chi(y) = \omega^{\ell(x)+\ell(y)}.$$

Now we argue that any linear mapping $\ell : \mathbb{F}_{p^n} \to \mathbb{F}_p$ can be represented as $\ell(x) \equiv \mathrm{Tr}(ax)$ for some $a \in \mathbb{F}_{p^n}$. This is proved by a counting argument. Each linear map $\ell : \mathbb{F}_{p^n} \to \mathbb{F}_p$ can be uniquely identified by its image on a basis for $\mathbb{F}_{p^n}$ as a linear space over $\mathbb{F}_p$. Thus, the number of such linear mappings is at most p^n. On the other hand, for each $a \in \mathbb{F}_{p^n}$ the mapping $x \to Tr(ax)$ is linear (since Trace is a linear mapping), and the total number of theses mappings is the number of distinct $a \in \mathbb{F}_{p^n}$, that is p^n. To conclude we just need to show that for any distinct $a \neq b \in \mathbb{F}_{p^n}$ the mappings $\mathrm{Tr}(ax)$ and $\mathrm{Tr}(bx)$ are distinct. Equivalently, since Trace is a linear mapping, we need to show that $Tr((a-b)x) \not\equiv 0$. This is clear however because the Trace mapping is not identically zero and $a - b \neq 0$ is invertible. □

Claim 13.11 (Trace of a p-power is unbiased)**.** *For every $c \neq 0$ and $0 \leq L \leq n-1$ we have*

$$\mathbb{E}_{x \in \mathbb{F}_{p^n}}[\omega^{\mathrm{Tr}(cx^{p^L})}] = 0.$$

Proof. We have $Tr(cx^{p^L}) = Tr(c^{p^{n-L}}x)$, so it suffices to prove the claim for $L = 0$. Let $\ell : \mathbb{F}_{p^n} \to \mathbb{F}_p$ defined as $\ell(x) = Tr(cx)$. The mapping ℓ is linear, and as it is not identically zero, its output is uniform over $\mathbb{F}_p$. Thus we have that $\mathbb{E}_{x \in \mathbb{F}_{p^n}}[\omega^{\ell(x)}] = 0$. □

Reduced forms

We define in this subsection reduced forms of polynomials. We show that for studying character sums it is the sufficient to restrict to reduced polynomials. We start by considering univariate polynomials, and then generalize the definitions and claims to multivariate polynomials.

Definition 13.13 (Reduced form: univariate polynomials)**.** *Let $m(x) = ax^t$ be a monomial. We say m is* reduced *if $p \nmid t$. If $t = p^k r$ for $p \nmid r$ we define the* reduced form *of $m(x)$ to be $m(x)^{p^{n-k}} \equiv a^{p^{n-k}} x^r$. A constant term $c \in \mathbb{F}_{p^n}$ is reduced if $c \in \mathbb{F}_p$, otherwise its reduced form is $\mathrm{Tr}(c) \in \mathbb{F}_p$. We say a polynomial is reduced if all its monomials are reduced, and the reduced form of a polynomial is the sum of the reduced forms of its monomials.*

Claim 13.12 (Equivalence of reduced form: univariate polynomials)**.** *Let $f(x)$ be a univariate polynomial over $\mathbb{F}$. Let $f'(x)$ be its reduced form. Then*

1. $\mathrm{Tr}(f(x)) \equiv \mathrm{Tr}(f'(x))$.
2. $deg(f') \leq deg(f)$.
3. $wt(f') \leq wt(f)$.

Proof. For a monomial $m(x) = ax^t$ with $t = p^k r$, $p \nmid r$, let $m'(x) = a^{p^{n-k}} x^r$ be its reduced form. Note that $m'(x) = m(x)^{p^{n-k}}$. Since $\mathrm{Tr}(x) = \mathrm{Tr}(x^p)$ we have that $\mathrm{Tr}(m(x)) = \mathrm{Tr}(m'(x))$ for all $x \in \mathbb{F}$. Note that $wt(m') = wt(m)$ and $\deg(m') = r \leq t = \deg(m)$. For a general polynomial $f(x) = \sum m_i(x)$ we have that $f'(x) = \sum m'_i(x)$. Hence we get that $\mathrm{Tr}(f) \equiv \mathrm{Tr}(f')$, and since cancelations among the m'_i can only reduce the degree and weight degree of f', we get that $deg(f') \leq deg(f)$ and $wt(f') \leq wt(f)$. □

Claim 13.13 (Trace of reduced non-constant polynomial is non-constant: univariate polynomials)**.** *Let $f(x)$ be a non-constant reduced univariate polynomial. Then $\mathrm{Tr}(f(x))$ is not constant.*

Proof. Assume for contradiction that $\mathrm{Tr}(f(x)) \equiv c$ for some $c \in \mathbb{F}_p$. Let $f(x) = a_0 + \sum_{i \in I} a_i x^i$ where $a_0 \in \mathbb{F}_p$, $a_i \in \mathbb{F}_{p^n}$ for $i \in I$ and $I \subseteq \{0, \ldots, p^n - 1\}$ is nonempty such that $p \nmid i$ for all $i \in I$. Define $g(x) = \mathrm{Tr}(f(x)) - c$. We have that

$$g(x) = -c + \mathrm{Tr}(f(x)) = (a_0 - c) + \sum_{i \in I} \sum_{j=0}^{n-1} a_i^{p^j} x^{ip^j} = (a_0 - c) + \sum_{i \in I} \sum_{j=0}^{n-1} a_i^{p^j} x^{ip^j} \pmod{p^n}.$$

Notice that all the monomials in this representation are distinct, since all $i \in I$ are not divisible by p. Thus this is a non-zero polynomial of degree at most $p^n - 1$, and so it cannot evaluate to zero on all elements of $\mathbb{F}_{p^n}$. □

We now generalize some of the definitions and claims to multivariate polynomials. When we refer to the degree of a multivariate polynomial we always mean is its total degree. The weight degree of a monomial $x_1^{e_1} \ldots x_s^{e_s}$ is the sum of the weight degrees of the variables, that is $wt(x_1^{e_1} \ldots x_s^{e_s}) = wt(x_1^{e_1}) + \ldots + wt(x_s^{e_s})$. The weight degree of a multivariate polynomial is the maximal weight degree of its monomials.

Note 13.2. *As in the univariate case, the weight degree of a multivariate degree is equivalent to its derivative degree, which is defined in an analogous way to the univariate case.*

Definition 13.14 (Reduced form: multivariate polynomials)**.** *Let $m(x_1, \ldots, x_s) = ax_1^{e_1} \ldots x_s^{e_s}$ be a monomial. We say m is* reduced *if $p \nmid gcd(e_1, \ldots, e_s)$ (that is, at least one e_i is co-prime to p). If $e_i = p^k r_i$ where $p \nmid gcd(r_1, \ldots, r_s)$ we define the* reduced form *of $m(x_1, \ldots, x_s)$ to be $a^{p^{n-k}} x_1^{r_1} \ldots x_s^{r_s}$. We say a polynomial is reduced if all its monomials are reduced, and the reduced form of a polynomial is the sum of the reduced forms of its monomial.*

Claim 13.14 (Equivalence of reduced form: multivariate polynomials)**.** *Let $f(x_1, \ldots, x_s)$ be a multivariate polynomial over $\mathbb{F}$. Let $f'(x_1, \ldots, x_s)$ be its reduced form. Then*

1. $\mathrm{Tr}(f(x_1, \ldots, x_s)) \equiv \mathrm{Tr}(f'(x_1, \ldots, x_s))$.
2. $deg(f') \leq deg(f)$.
3. $wt(f') \leq wt(f)$.

Proof. The proof is identical to the proof of Claim 13.14 for the univariate case. □

Claim 13.15 (Trace of reduced non-constant polynomial is non-constant: multivariate polynomials)**.** *Let $f(x_1, \ldots, x_s)$ be a non-constant reduced multivariate polynomial. Then $\mathrm{Tr}(f(x_1, \ldots, x_s))$ is not constant.*

Proof. The proof is very similar to the proof of Claim 13.13 for the univariate case. If f is not a constant polynomial, that is if I is not empty, then for any $c \in \mathbb{F}_p$ the polynomial $\mathrm{Tr}(f(x_1, \ldots, x_s)) - c$ is a non-zero polynomial of individual degree at most $p^n - 1$ in each variable, and such a polynomial cannot evaluate to zero on all points in $(\mathbb{F}_{p^n})^s$. □

Properties of derivatives

Let $f(x)$ be a univariate polynomial. For every $s \geq 1$ define the s-iterated derivative polynomial of f, $\Delta f(x; y_1, \ldots, y_s)$, to be the multivariate polynomial in variables $x, y_1, \ldots, y_s \in \mathbb{F}$ defined as

$$\Delta f(x; y_1, \ldots, y_s) = f_{y_1,\ldots,y_s}(x) = \sum_{I \subseteq [s]} (-1)^{|I|+s} f(x + \sum_{i \in I} y_i).$$

Derivatives play a crucial role in the proof of Theorem 13.3. We study in this subsection some of their properties, and prove some structural results on polynomials of the form $\Delta f(x; y_1, \ldots, y_s)$.

Claim 13.16 (Derivation maintains degree)**.** *Let $m(x) = x^t$ be a monomial. Then for any k, all the monomials appearing in $\Delta m(x; y_1, \ldots, y_k)$ have total degree t (or $\Delta m(x; y_1, \ldots, y_k) \equiv 0$).*

Proof. The polynomial $\Delta m(x; y_1, \ldots, y_k)$ is a linear combination of $(x + \sum_{i \in I} y_i)^t$ for subsets $I \subseteq [k]$, each of which is homogeneous of degree t. □

We show that the character sum of a polynomial can be bounded by a character sum of its iterated derivatives polynomial.

Claim 13.17 (Bias can be bounded by bias of derivatives)**.** *For any univariate polynomial $f(x)$ and $s \geq 1$*

$$\left|\mathbb{E}_{x \in \mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right| \leq \left(\mathbb{E}_{x, y_1, \ldots, y_s \in \mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_s))}]\right)^{1/2^s}$$

Proof. Consider first the case $s = 1$. We have

$$\left|\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right|^2 = \mathbb{E}_{x,x'\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}\overline{\omega^{\mathrm{Tr}(f(x'))}}] =$$
$$\mathbb{E}_{x,x'\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))-\mathrm{Tr}(f(x'))}] = \mathbb{E}_{x,y\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x+y))-\mathrm{Tr}(f(x))}] =$$
$$\mathbb{E}_{x,y\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x+y)-f(x))}] = \mathbb{E}_{x,y\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y))}].$$

Hence

$$\left|\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right| \leq \left(\mathbb{E}_{x,y\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y))}]\right)^{1/2}.$$

For $s > 1$ we prove the result by induction. By the base case of $s = 1$ and the Cauchy-Schwartz inequality, we have that

$$\left|\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right|^{2^s} \leq \left(\mathbb{E}_{x,y_1\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y_1))}]\right)^{2^{s-1}} \leq \mathbb{E}_{y_1\in\mathbb{F}}\left[\left(\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y_1))}]\right)^{2^{s-1}}\right].$$

For every value of $y_1 \in \mathbb{F}$ we have by the $s-1$ case that

$$\left(\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y_1))}]\right)^{2^{s-1}} \leq \mathbb{E}_{x,y_2,\ldots,y_s\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y_1,\ldots,y_s))}],$$

hence we get that

$$\left|\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right|^{2^s} \leq \mathbb{E}_{x,y_1,y_2,\ldots,y_s\in\mathbb{F}}[\omega^{\mathrm{Tr}(\Delta f(x;y_1,\ldots,y_s))}].$$

□

We now define a special family of multivariate polynomials that will play an important role in the proof. Such polynomials arise when taking d-iterated derivatives from a polynomial of weight degree d.

Definition 13.15 (p-multilinear polynomials)**.** *A multivariate polynomial $f(x_1,\ldots,x_s)$ over $\mathbb{F}_{p^n}$ is p-multilinear if all its monomials are of the form $x_1^{p^{i_1}}\ldots x_s^{p^{i_s}}$. In particular, if it is nonzero it has weight degree s.*

Claim 13.18 (Structure of derivatives of monomials)**.** *Let $m(x) = x^t$ be a monomial of weight degree d. The d-iterated derivatives polynomial $\Delta m(x;y_1,\ldots,y_d)$ of m is given as follows. Let $t = \sum_{j=1}^k a_{\ell_j}p^{\ell_j}$ where $1 \leq a_{\ell_1},\ldots,a_{\ell_k} \leq p-1$ and $\sum a_\ell = d$. Let $\mathcal{S}$ be the family of all partitions of $\{1,\ldots,d\}$ into k subsets of sizes $a_{\ell_1},\ldots,a_{\ell_s}$, that is*

$$\mathcal{S} = \{(S_1,\ldots,S_k) : S_1\cup\ldots\cup S_k = \{1,\ldots,d\}, |S_1| = a_{\ell_1},\ldots,|S_k| = a_{\ell_k}\}.$$

Then we have

$$\Delta m(x;y_1,\ldots,y_d) = c\sum_{(S_1,\ldots,S_k)\in\mathcal{S}}\prod_{j=1}^k\prod_{i\in S_j}(y_i)^{p^{\ell_j}}.$$

where $c = \prod_{j=1}^k a_{\ell_j}! \neq 0$ in $\mathbb{F}$. In particular, Δm is a non-zero p-multilinear polynomial in $y_1,\ldots,y_d$ which does not depend on x.

Proof. We have

$$\Delta m(x; y_1, \ldots, y_d) = \sum_{I \subseteq [d]} (-1)^{d+|I|} m(x + \sum_{i \in I} y_i) = \sum_{I \subseteq [d]} (-1)^{d+|I|} (x + \sum_{i \in I} y_i)^t.$$

Substituting $t = \sum a_{\ell_j} p^{\ell_j}$, and using the linearity of the Frobenius map $x \to x^{p^{\ell_j}}$ we get that

$$\Delta m(x; y_1, \ldots, y_d) = \sum_{I \subseteq [d]} (-1)^{d+|I|} \prod_{j=1}^{k} (x^{p^{\ell_j}} + \sum_{i \in I} (y_i)^{p^{\ell_j}})^{a_{\ell_j}}.$$

Since $\sum a_{\ell_j} = d$ we get that Δm is a degree-d polynomial in the Frobenius images of $x, y_1, \ldots, y_d$, i.e. in the monomials $\{x^{p^j}, (y_1)^{p^j}, \ldots, (y_d)^{p^j} : 0 \leq j \leq n-1\}$.

We first claim that Δm does not depend on x, and is p-linear in $y_1, \ldots, y_d$. That is, all the monomials of Δm consist of a product $(y_1)^{p^{j_1}} \ldots (y_d)^{p^{j_d}}$, where $0 \leq j_1, \ldots, j_d \leq n-1$. Otherwise, there exists some monomial in Δm which does not depend on at least one of $y_1, \ldots, y_d$. This is because all monomials of Δm are products of d Frobenius images of $x, y_1, \ldots, y_d$, and by the pigeonhole principle, if either a single variable y_i has two images appearing, or an image of x appears in the monomial, then there must exists a variable y_j not participating in the monomial.

Assume w.l.o.g that Δm contains monomials in which y_1 does not participate. Substituting $y_1 = 0$ in the definition of Δm, since $\Delta f(x; 0) = f(x) - f(x) \equiv 0$ for any polynomial f, we get that

$$\Delta m(x; 0, y_2, \ldots, y_d) \equiv 0.$$

Hence, if there exist monomials in $\Delta m(x; y_1, \ldots, y_d)$ which do not depend on y_1, they are left intact by the substitution $y_1 = 0$, while all monomials depending on y_1 vanish. Thus since $\Delta m(x; 0, y_2, \ldots, y_d) \equiv 0$ all the monomials in $\Delta m(x; y_1, \ldots, y_d)$ must depend on y_1.

We have thus proved that $\Delta m(x; y_1, \ldots, y_d)$ does not depend on x, and is p-linear in $y_1, \ldots, y_d$. To conclude we need to compute the exact form of $\Delta m(x; y_1, \ldots, y_d)$. Any monomial depending on all $y_1, \ldots, y_d$ must come from the term corresponding for $I = \{1, \ldots, d\}$,

$$(x + \sum_{i \in [d]} y_i)^t = \prod_{j=1}^{k} (x^{p^{\ell_j}} + \sum_{i \in [d]} (y_i)^{p^{\ell_j}})^{a_{\ell_j}}.$$

The individual degree of each y_i is some p^{ℓ_j}, and there are exactly a_{ℓ_j} variables among $y_1, \ldots, y_d$ which has individual degree p^{ℓ_j}. Since the number of variables d is exactly the sum $\sum a_{\ell_j}$, all the monomials depending on all of $y_1, \ldots, y_d$ must be of the form $\prod_{j=1}^{k} \prod_{i \in S_j} (y_i)^{p^{\ell_j}}$, where $(S_1, \ldots, S_k) \in \mathcal{S}$ is a partition of $\{1, \ldots, d\}$ into sets of sizes $a_{\ell_1}, \ldots, a_{\ell_k}$. The coefficient of the monomial $\prod_{j=1}^{k} \prod_{i \in S_j} (y_i)^{p^{\ell_j}}$ is equal to the number of times this monomial appears in the last term, which is exactly $\prod_{j=1}^{k} a_{\ell_j}!$. □

Claim 13.19 (Derivative of reduced monomial is nonzero)**.** *Let $m(x)$ be a nonzero reduced monomial of weight degree d. Then $\Delta m(x; y_1, \ldots, y_d)$ is a nonzero reduced polynomial.*

Proof. Let $m(x) = x^t$ for $t = \sum a_{\ell_j} p^{\ell_j}$. Since m is reduced we must have $a_0 \neq 0$. By Claim 13.18 we know that

$$\Delta m(x; y_1, \ldots, y_d) = c \sum_{(S_1,\ldots,S_k)\in\mathcal{S}} \prod_{j=1}^{k} \prod_{i\in S_j} (y_i)^{p^{\ell_j}}.$$

Thus any monomial of $\Delta m(x; y_1, \ldots, y_d)$ contains at least one variable of degree 1, thus it is reduced. □

Claim 13.20 (Derivative of distinct reduced monomials is distinct)**.** *Let $m'(x), m''(x)$ be two distinct monomials of weight degree d. Then $\Delta m'(x; y_1, \ldots, y_d)$ and $\Delta m''(x; y_1, \ldots, y_d)$ are nonzero polynomials which do not share any common monomial.*

Proof. Let $m'(x) = x^{t'}$ and $m''(x) = x^{t''}$ for $t' \neq t''$. By Claim 13.18 we have that $\Delta m'(x; y_1, \ldots, y_d)$ is a nonzero polynomial such that all its monomials have total degree exactly t'. Similarly $\Delta m''(x; y_1, \ldots, y_d)$ is a nonzero polynomial such that all its monomials have total degree exactly t''. Since $t' \neq t''$ the polynomials $\Delta m'(x; y_1, \ldots, y_d)$ and $\Delta m''(x; y_1, \ldots, y_d)$ contain no common monomial. □

Claim 13.21 (High derivative vanishes)**.** *Let $f(x)$ be a polynomial of weight degree at most $d-1$. Then $\Delta m(x; y_1, \ldots, y_d) \equiv 0$.*

Proof. It is enough to prove the claim for monomials. Let $m(x) = x^t$ be some monomial, and let $d' = wt(m) \leq d-1$ be its weight degree. By Claim 13.18 we have that $\Delta m(x; y_1, \ldots, y_{d'})$ does not depend on x, thus

$$\Delta m(x; y_1, \ldots, y_{d'}, y_{d'+1}) = \Delta m(x + y_{d'+1}; y_1, \ldots, y_{d'}) - \Delta m(x; y_1, \ldots, y_{d'}) \equiv 0.$$

□

Lemma 13.3 (Highest non-vanishing derivative)**.** *Let $f(x)$ be a nonzero reduced polynomial of weight degree d. Then $\Delta f(x; y_1, \ldots, y_d)$ is a nonzero reduced polynomial which does not depend on x and is p-linear in $y_1, \ldots, y_d$.*

Proof. Let $f(x) = \sum c_t x^t$. Let $m(x) = c_t x^t$ be some monomial of f. If $wt(m) \leq d-1$ then by Claim 13.21 we have $\Delta m(x; y_1, \ldots, y_d) \equiv 0$. Thus it is enough to consider just the monomials of weight degree exactly d. By Claim 13.19 the derivative polynomial of each reduced monomial of weight degree d is a reduced polynomial, and these polynomials for two distinct monomials contain no shared monomials, and so cannot cancel each other. Thus the derivative polynomial $\Delta f(x; y_1, \ldots, y_d)$ is a nonzero reduced polynomial. By Claim 13.18 is does not depend on x, and it is p-linear in $y_1, \ldots, y_d$. □

Lemma 13.4 (General non-vanishing derivatives)**.** *Let $f(x)$ be a nonzero reduced polynomial of weight degree d. For any $k \leq d$ the polynomial $\Delta f(x; y_1, \ldots, y_k)$ is a nonzero reduced polynomial in $x, y_1, \ldots, y_k$.*

Proof. Let $f(x) = \sum c_t x^t$. Let $m(x) = c_t x^t$ be some monomial of f. Observe that all monomials in the polynomial $\Delta m(x; y_1, \ldots, y_k)$ have the same total degree t. Thus, if $m(x)$ is reduced then so is $\Delta m(x; y_1, \ldots, y_k)$, since if $x^{e_0} y_1^{e_1} \ldots y_k^{e_k}$ is a monomial of $\Delta m(x; y_1, \ldots, y_k)$ which is not reduced, then $p \mid gcd(e_0, \ldots, e_k)$. However $t = e_0 + \ldots + e_k$ and since $m(x)$ is reduced we have that $p \nmid t$. Contradiction, hence $\Delta m(x; y_1, \ldots, y_k)$ must be reduced. Hence, we get that if $f(x)$ is a reduced polynomial, then $\Delta f(x; y_1, \ldots, y_k)$ is also reduced. To conclude we need to prove that $\Delta f(x; y_1, \ldots, y_k)$ is nonzero. Assume by contradiction it is zero; then so is $\Delta f(x; y_1, \ldots, y_d) = \sum_{I \subseteq \{k+1, \ldots, d\}} (-1)^{|I|+d-k} \Delta f(x + \sum_{i \in I} y_i; y_1, \ldots, y_k)$. However by Lemma 13.3 we know that if f is a nonzero reduced polynomial, then $\Delta f(x; y_1, \ldots, y_d)$ is nonzero. Hence also $\Delta f(x; y_1, \ldots, y_k)$ must be nonzero. □

Additional claims

We give in this subsection some more claims we will require. The first is the Schwarz-Zippel lemma.

Claim 13.22 (Schwarz-Zippel)**.** *Let $f(x_1, \ldots, x_s)$ be a polynomial over $\mathbb{F}$ of total degree e. Then*

$$\Pr_{x_1, \ldots, x_s \in \mathbb{F}}[f(x_1, \ldots, x_s) = 0] \leq \frac{e}{|\mathbb{F}|}.$$

The second result we will need is a theorem of Deligne [Del78] which is a multivariate analog of Weil's bound.

Theorem 13.8 (Deligne theorem [Del78])**.** *Let $f(x_1, \ldots, x_s)$ be a multivariate polynomial over $\mathbb{F}$ of degree $|\mathbb{F}|^{1/2-\delta}$. Let $\chi : \mathbb{F} \to \mathbb{C}$ be an additive character. Then either $\chi(f(x_1, \ldots, x_s))$ is constant or*

$$|\mathbb{E}_{x_1, \ldots, x_s \in \mathbb{F}}[\chi(f(x))]| \leq |\mathbb{F}|^{-\delta}.$$

13.3.2 The case of high weight g

In this subsection we prove Theorem 13.3 in the case that g has high weight degree, $\mathrm{wt}(g) \geq d + 1$. This is captured by the following lemma, which we prove in this subsection. This is the easier case for Theorem 13.3.

Lemma 13.5 (The case of high weight g)**.** *Let $f(x) = g(x) + h(x)$ be a nonzero reduced univariate polynomial over $\mathbb{F}_{p^n}$, where $g(x)$ is a polynomial of degree $|\mathbb{F}|^{1/2-\delta}$ and weight degree at least $d + 1$, and $h(x)$ has weight degree at most d. Then*

$$\left|\mathbb{E}_{x \in \mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right| \leq |\mathbb{F}|^{-\frac{\delta}{2^{d+1}}}.$$

Proof. The polynomial f is nonzero reduced and of weight degree at least $d + 1$. By Lemma 13.4 we know that $\Delta f(x; y_1, \ldots, y_{d+1})$ is nonzero and reduced. However, since $\mathrm{wt}(h) \leq d$ we have that $\Delta h(x; y_1, \ldots, y_{d+1}) \equiv 0$ by Claim 13.21, hence we get that $\Delta f(x; y_1, \ldots, y_{d+1}) = \Delta g(x; y_1, \ldots, y_{d+1})$. Also, since derivation cannot increase total degree, we have that $\deg(\Delta f(x; y_1, \ldots, y_{d+1})) \leq \deg(g) \leq |\mathbb{F}|^{1/2-\delta}$.

So, we have that $f'(x, y_1, \ldots, y_{d+1}) = \Delta f(x; y_1, \ldots, y_{d+1})$ is a nonzero reduced polynomial of degree at most $|\mathbb{F}|^{1/2-\delta}$. By Claim 13.15 we have that $\mathrm{Tr}(f')$ is a non-constant function. Thus by Deligne's Theorem (Theorem 13.8) we get that is must be highly unbiased, that is

$$\left|\mathbb{E}_{x,y_1,\ldots,y_{d+1}\in\mathbb{F}}[\omega^{\mathrm{Tr}(f'(x,y_1,\ldots,y_{d+1}))}]\right| \leq |\mathbb{F}|^{-\delta}.$$

To conclude we apply Claim 13.17 to get that

$$\left|\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right| \leq \left|\mathbb{E}_{x,y_1,\ldots,y_{d+1}\in\mathbb{F}}[\omega^{\mathrm{Tr}(f'(x,y_1,\ldots,y_{d+1}))}]\right|^{\frac{1}{2^{d+1}}} \leq |\mathbb{F}|^{-\frac{\delta}{2^{d+1}}}.$$

□

13.3.3 The case of low weight g

In this subsection we prove Theorem 13.3 in the case that g has low weight degree, $\mathrm{wt}(g) \leq d$. This is captured by the following lemma, which we prove in this subsection. This is the harder case for Theorem 13.3.

Lemma 13.6 (The case of low weight g)**.** *Let $f(x) = g(x) + h(x)$ be a nonzero reduced univariate polynomial over $\mathbb{F}_{p^n}$, where $g(x)$ is a polynomial of degree $|\mathbb{F}|^{1/2-\delta}$ and weight degree at most d, and $h(x)$ has weight degree d and is the sum of k monomials. Then*

$$\mathbb{E}_{x\in\mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}] \leq |\mathbb{F}|^{-\frac{\delta}{d^2 2^{d_k}}+O(1/n)}.$$

To prove Lemma 13.6 we require some claims.

Claim 13.23 (Structure of derivative of g)**.** *Let $g(x)$ be a polynomial of degree at most $|\mathbb{F}|^{1/2-\delta}$ and weight degree at most d. For $L = \lceil n(1/2-\delta)\rceil$ there exists a p-multilinear polynomial $u(y_2, \ldots, y_d)$ such that*

$$\mathrm{Tr}(\Delta g(x; y_1, \ldots, y_d)) \equiv Tr(y_1^{p^L} \cdot u(y_2, \ldots, y_d)).$$

and such that $\deg(u) \leq p^{2L} \leq |\mathbb{F}|^{1-2\delta+2/n}$.

Proof. By linearity, it suffices to show that for every monomial $m(x)$ appearing in g, there exists a p-multilinear polynomial $u_m(y_2, \ldots, y_d)$ such that $\mathrm{Tr}(\Delta m(x; y_1, \ldots, y_d)) \equiv Tr(y_1^{p^L} \cdot u_m(y_2, \ldots, y_d))$ and $\deg(u_m) \leq p^{2L}$.

Let $m(x) = cx^t$ be such a monomial. If $wt(m) < d$ we have by Claim 13.21 that $\Delta m(x; y_1, \ldots, y_d) \equiv 0$. Otherwise assume that $\mathrm{wt}(m) = d$. By Claim 13.18 we know that $\Delta m(x; y_1, \ldots, y_d)$ does not depend on x and is p-multilinear in $y_1, \ldots, y_d$. Moreover, if $t = \sum_{j=1}^{k} a_{\ell_j} p^{\ell_j}$ where $1 \leq a_{\ell_j} \leq p-1$ we know that

$$\Delta m(x; y_1, \ldots, y_d) = \sum_{j=1}^{k} y_1^{p^{\ell_j}} w_j(y_2, \ldots, y_d)$$

where $w_j(y_2,\ldots,y_d)$ is a homogeneous p-multilinear polynomial of total degree $t-p^{\ell_j}$. Since $t \le |\mathbb{F}|^{1/2-\delta}$ we have that $\ell_1,\ldots,\ell_k \le n(1/2-\delta) \le L$. Thus, taking $u_m(y_2,\ldots,y_d)$ to be

$$u_m(y_2,\ldots,y_d) = \sum_{j=1}^{k} w_j(y_2,\ldots,y_d)^{p^{L-\ell_j}}$$

we get that

$$Tr(y_1^{p^L}\cdot u_m(y_2,\ldots,y_d)) \equiv \sum_{j=1}^{k} Tr(y_1^{p^L} w_j(y_2,\ldots,y_d)^{p^{L-\ell_j}}) \equiv$$

$$\sum_{j=1}^{k} Tr(y_1^{p^{\ell_j}} w_j(y_2,\ldots,y_d)) = \mathrm{Tr}(\Delta m(x;y_1,\ldots,y_d)).$$

To conclude we need to bound $\deg(u_m)$. Since $\deg(w_j) \le \deg(m) \le p^{n(1/2-\delta)}$ and $L-\ell_j \le L$ we get that $\deg(u_m) \le \deg(m)\cdot p^L \le p^{2L}$. □

Claim 13.24 (Structure of derivative of h)**.** *Let $h(x)$ be a polynomial of weight degree d which is the sum of k monomials. For every $0 \le L \le n-1$ there exists a p-multilinear polynomial $v(y_2,\ldots,y_d)$ such that*

$$\mathrm{Tr}(\Delta h(x;y_1,\ldots,y_d)) \equiv Tr(y_1^{p^L}\cdot v(y_2,\ldots,y_d)).$$

and the number of distinct total degrees of monomials appearing in v is at most kd.

Proof. By linearity, it suffices to show that for every monomial $m(x)$ appearing in h, there exists a p-multilinear polynomial $v_m(y_2,\ldots,y_d)$ such that $\mathrm{Tr}(\Delta m(x;y_1,\ldots,y_d)) \equiv Tr(y_1^{p^L}\cdot v_m(y_2,\ldots,y_d))$ and the monomials appearing in v_m have at most d distinct total degrees.

Let $m(x) = cx^t$ be such a monomial. If $wt(m) < d$ we have by Claim 13.21 that $\Delta m(x;y_1,\ldots,y_d) \equiv 0$. Otherwise assume that $\mathrm{wt}(m) = d$. By Claim 13.18 we know that $\Delta m(x;y_1,\ldots,y_d)$ does not depend on x and is p-multilinear in $y_1,\ldots,y_d$. Moreover, if $t = \sum_{j=1}^{k} a_{\ell_j} p^{\ell_j}$ where $1 \le a_{\ell_j} \le p-1$ we know that

$$\Delta m(x;y_1,\ldots,y_d) = \sum_{j=1}^{k} y_1^{p^{\ell_j}} w_j(y_2,\ldots,y_d)$$

where $w_j(y_2,\ldots,y_d)$ is a homogeneous p-multilinear polynomial of total degree $t-p^{\ell_j}$. Let

$$v_m(y_2,\ldots,y_d) = \sum_{j=1}^{k} w_j(y_2,\ldots,y_d)^{p^{L-\ell_j+n}}$$

where we reduce individual powers of $y_2,\ldots,y_d$ modulo p^n (that is, we replace each y_i^e with $y_i^{e \bmod p^n}$, which are equivalent as functions over the field $\mathbb{F}_{p^n}$). Thus we get that

$$Tr(y_1^{p^L}\cdot v_m(y_2,\ldots,y_d)) \equiv \sum_{j=1}^{k} Tr(y_1^{p^L} w_j(y_2,\ldots,y_d)^{p^{L-\ell_j+n}}) \equiv$$

$$\sum_{j=1}^{k} Tr(y_1^{p^{\ell_j}} w_j(y_2,\ldots,y_d)) = \mathrm{Tr}(\Delta m(x;y_1,\ldots,y_d)).$$

To conclude we need to bound the number of distinct total degrees of monomials appearing in v_m. Each polynomial w_j is homogeneous, and so also $w_j^{p^{L-\ell_j+n}}$ is homogenous, hence contributing a unique total degree to monomials in v_m. As the number of distinct w_j is bounded by $k \leq d$ we get the required bound. □

Claim 13.25 (Covering argument for a single element)**.** *Let $0 \leq e \leq p^n - 1$ such that $\mathrm{wt}(e) = d$. For $0 \leq s \leq n-1$ define $e_s = e \cdot p^s \bmod p^n$, such that also $0 \leq e_s \leq p^n - 1$. For $a \leq n$ let*

$$S = \{0 \leq s \leq n-1 : e_s \geq p^{n-a}\}.$$

Then $|S| \leq a \cdot d$.

Proof. For every $0 \leq e \leq p^n - 1$ let $\vec{e} \in \{0, \ldots, p-1\}^n$ denote the vector corresponding to the base-p representation of e, that is $e = \sum_{i=0}^{n-1} \vec{e}(i) p^i$. Observe that $\vec{e}_s$ is just the cyclic shift of $\vec{e}$ by s coordinates, that is $\vec{e}_s(i) = \vec{e}(i - s \pmod n)$. Note that the weight of e is just the hamming weight of $\vec{e}$, and that $e_s \geq p^{n-a}$ if and only if the vector $\vec{e}_s$ contains some nonzero entry in the indices $n-a \leq i \leq n-1$. As $\vec{e}$ contains only d nonzero entries, there are at most $a \cdot d$ cyclic shift of $\vec{e}$ such that some of these entries moves to indices $i \in \{n-a, \ldots, n-1\}$. Thus we get that $|S| \leq a \cdot d$. □

Claim 13.26 (Covering argument for sum of monomials)**.** *Let $h(y_1, \ldots, y_b)$ be a polynomial over $\mathbb{F}_{p^n}$ of weight degree at most d, such that the number of distinct total degrees of its monomial is z. Let $h_s(y_1, \ldots, y_b) = h(y_1, \ldots, y_b)^{p^s}$ reducing each individual degree of $y_1, \ldots, y_b$ modulo p^n. Then for every a there exists $0 \leq s \leq a$ such that*

$$\deg(h_s) < p^{n - \lfloor \frac{a}{dz} \rfloor}.$$

Proof. Let $q = \lfloor \frac{a}{dz} \rfloor$. Let $\{e_1, \ldots, e_z\}$ be the set of total degrees occurring in monomials of h. The number of $0 \leq s \leq n-1$ such that $(e_i \cdot p^s \bmod p^n) \geq p^{n-q}$ is bounded by $d \cdot q \leq a/z$ by Claim 13.25. Thus, there are at most a values for s such that for some e_i we have $e_i \cdot p^s \bmod p^n \geq p^{n-q}$. Since there are $a+1$ possible values for $0 \leq s \leq a$, by the pigeonhole principle there exists a value for which for all $i = 1, \ldots, k$,

$$(e_i \cdot p^s \bmod p^n) < p^{n-q}$$

hence we get that $\deg(h_s) < p^{n-q}$. □

Claim 13.27 (Structure of derivative of f)**.** *Let $f(x) = g(x) + h(x)$ be a nonzero reduced univariate polynomial over $\mathbb{F}_{p^n}$, where $g(x)$ is a polynomial of degree $|\mathbb{F}|^{1/2-\delta}$ and weight degree at most d, and $h(x)$ has weight degree d and is the sum of k monomials. Then there exists $M \in \{0, \ldots, n-1\}$ and a p-multilinear polynomial $r(y_2, \ldots, y_d)$ such that*

$$\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_d)) \equiv \mathrm{Tr}(y_1^{p^M} \cdot r(y_2, \ldots, y_d))$$

and $\deg(r) \leq |\mathbb{F}|^{1 - \frac{2\delta}{d^2 k+1} + 3/n}$.

Proof. Let $L = \lceil n(1/2-\delta) \rceil$. By Claim 13.23 there is a p-multilinear polynomial $u(y_2, \ldots, y_d)$ such that $\mathrm{Tr}(\Delta g(x; y_2, \ldots, y_d)) \equiv Tr(y_1^{p^L} \cdot u(y_2, \ldots, y_d))$ and $\deg(u) \leq p^{2L}$. By Claim 13.24 there is a p-multilinear polynomial $v(y_2, \ldots, y_d)$ such that $\mathrm{Tr}(\Delta h(x; y_2, \ldots, y_d)) \equiv Tr(y_1^{p^L} \cdot v(y_2, \ldots, y_d))$ and the number of distinct total degrees of monomials in v is bounded by kd.

For s define $r_s(y_2, \ldots, y_d) = p^s(u(y_2, \ldots, y_d) + v(y_2, \ldots, y_d))$ where individual degrees of $y_2, \ldots, y_d$ are reduced modulo p^n, and set $a = \alpha n$ to be determined later. We will show there exists $0 \leq s \leq n - 2L - a$ such that $\deg(r_s) \leq p^{n-a}$. This will establish the result as for every s,

$$\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_d)) \equiv \mathrm{Tr}(y_1^{p^{L+s}} r_s(y_2, \ldots, y_d)).$$

First, notice that since $\deg(u) \leq p^{2L}$ we have that for any $0 \leq s \leq n - 2L - a$ we have that

$$\deg(u^{p^s}) \leq \deg(u) \cdot p^s \leq p^{2L+s} \leq p^{n-a}.$$

We now move to consider v. By Claim 13.26 we have that there exists $0 \leq s \leq n - 2L - a$ such that if we let $v_s(y_2, \ldots, y_d) = v(y_2, \ldots, y_d)^{p^s}$ reducing individual degrees modulo p^n, we have that

$$\deg(v_s) \leq p^{n - \lfloor \frac{n-2L-a}{d^2k} \rfloor}.$$

Combining the two bounds, we get that

$$\deg(r_s) \leq \max(p^{n-a}, p^{n - \lfloor \frac{n-2L-a}{d^2k} \rfloor}).$$

Setting $a = \lfloor \frac{n-2L-d^2k}{d^2k+1} \rfloor$ to optimize the bound we get that

$$\deg(r_s) \leq p^{n-a} \leq p^{n(1-\frac{2\delta}{d^2k+1})+3}.$$

□

We are now ready to prove Lemma 13.6.

Proof of Lemma 13.6. We will bound the bias of $\mathrm{Tr}(f(x))$ by the bias of $\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_d))$. By Claim 13.17 we have that

$$\left|\mathbb{E}_{x \in \mathbb{F}}[\omega^{\mathrm{Tr}(f(x))}]\right| \leq \left|\mathbb{E}_{x, y_1, \ldots, y_d \in \mathbb{F}}[\omega^{\mathrm{Tr}(f(x; y_1, \ldots, y_d))}]\right|^{1/2^d}.$$

To bound the bias of $\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_d))$, we apply Claim 13.27. We have

$$\mathrm{Tr}(\Delta f(x; y_1, \ldots, y_d)) \equiv \mathrm{Tr}(y_1^{p^M} \cdot r(y_2, \ldots, y_d))$$

where $\deg(r) \leq |\mathbb{F}|^{1 - \frac{2\delta}{d^2k+1} + 3/n}$. Moreover since f is nonzero and reduced, then by Lemma 13.3 $\Delta f(x; y_1, \ldots, y_d)$ is nonzero, hence $r(y_2, \ldots, y_d)$ must also be nonzero.

Whenever $y_2, \ldots, y_d$ are such that $r(y_2, \ldots, y_d) \neq 0$, we have that $\mathbb{E}_{y_1 \in \mathbb{F}}[\omega^{\mathrm{Tr}(y_1^{p^M} \cdot r(y_2, \ldots, y_d))}] = 0$ by Claim 13.11. The probability that $r(y_2, \ldots, y_d) = 0$ is bounded by Claim 13.22 by

$$\Pr_{y_2, \ldots, y_d \in \mathbb{F}}[r(y_2, \ldots, y_d) = 0] \leq \frac{\deg(r)}{|\mathbb{F}|} \leq |\mathbb{F}|^{-\frac{2\delta}{d^2k+1} + 3/n}.$$

Combining the results, we get that

$$\left|\mathbb{E}_{x\in\mathbb{F}}\left[\omega^{\mathrm{Tr}(f(x))}\right]\right| \leq |\mathbb{F}|^{-\frac{2\delta}{(d^2k+1)2^d}+\frac{3}{2^d n}} \leq |\mathbb{F}|^{-\frac{\delta}{d^2 2^d k}+O(1/n)}.$$

□

Part V

Property testing for polynomials

Chapter 14

The inverse conjecture for the Gowers norm is false

Let p be a fixed prime number, and N be a large integer. The 'Inverse Conjecture for the Gowers norm' states that if the "d-th Gowers norm" of a function $f : \mathbb{F}_p^N \to \mathbb{F}$ is non-negligible, that is larger than a constant independent of N, then f can be non-trivially approximated by a degree $d-1$ polynomial. The conjecture is known to hold for $d = 2, 3$ and for any prime p. In this paper we show the conjecture to be false for $p = 2$ and for $d = 4$, by presenting an explicit function whose 4-th Gowers norm is non-negligible, but whose correlation any polynomial of degree 3 is exponentially small.

Essentially the same result (with different correlation bounds) was independently obtained by Green and Tao [GT07]. Their analysis uses a modification of a Ramsey-type argument of Alon and Beigel [AB01] to show inapproximability of certain functions by low-degree polynomials.

We observe that a combination of our results with the argument of Alon and Beigel implies the inverse conjecture to be false for any prime p, for $d = p^2$.

Joint work with Roy Meshulam and Alex Samorodnitsky.

14.1 Introduction

We consider multivariate functions over finite fields. The main question of interest here would be whether these functions can be non-trivially approximated by a low-degree polynomial.

Fix a prime number p. Let $\mathbb{F} = \mathbb{F}_p$ be the finite field with p elements. Let $\xi = e^{\frac{2\pi i}{p}}$ be the primitive p-th root of unity. Denote by $e(x)$ the exponential function taking $x \in \mathbb{F}$ to $\xi^x \in \mathbb{C}$. For two functions $f, g : \mathbb{F}^N \to \mathbb{F}$, let $\langle f, g \rangle := \mathbb{E}_x e(f(x) - g(x))$.

Definition 14.1. *A function f is non-trivially approximable by a degree-d polynomial if*

$$| \langle f, g \rangle | > \epsilon$$

for some polynomial g of degree at most d in $\mathbb{F}[x_1 ... x_N]$.

More precisely, in this definition we are looking at a sequence f_N of functions and of approximating low-degree polynomials g_N in N variables, and let N grow to infinity. In this paper, the remaining parameters, that is the field size p, the degree d and the offset ϵ are fixed, independent of N.

A counting argument shows that a generic function cannot be approximated by a polynomial of low degree. The problems of showing a specific given function to have no non-trivial approximation and of constructing an explicit non-approximable function have been extensively investigated, since solutions to these problems have many applications in complexity (cf. discussion and references in [AB01, VW08, BV07])

This paper studies a technical tool that measures distance from low-degree polynomials. This is the Gowers norm, introduced in [Gow01]. For a function $f : \mathbb{F}^N \to \mathbb{F}$ and a vector $y \in \mathbb{F}^n$, we take f_y to be the directional derivative of f in direction y by setting

$$f_y(x) = f(x+y) - f(x)$$

For a k-tuple of vectors $y_1 ... y_k$ we take the iterated derivative in these directions to be

$$f_{y_1 ... y_k} = \left(f_{y_1 ... y_{k-1}}\right)_{y_k}$$

It is easy to see that this definition does not depend on the ordering of $y_1 ... y_k$.

The k-th Gowers "norm" $\|f\|_{U^k}$ of f is

$$\left(\mathbb{E}_{x, y_1 ... y_k} \left[e\left(f_{y_1 ... y_k}(x)\right)\right]\right)^{1/2^k}$$

More accurately, as shown in [Gow01], this is indeed a norm of the associated complex-valued function $e(f)$ (for $k \geq 2$).

It is easy to see that $\|f\|_{U^{d+1}}$ is 1 iff f is a polynomial of degree at most d. This is just another way of saying that all order-$(d+1)$ iterative derivatives of f are zero if and only if f is a polynomial of degree at most d. It is also possible to see [Gow01] that $|\langle f, g \rangle| > \epsilon$ for g of degree at most d, implies $\|f\|_{U^{d+1}} > \epsilon$. That is to say, if f is non-trivially close to a degree-d polynomial, this can be detectable via an appropriate Gowers norm.

This discussion naturally leads to the inverse conjecture [GT08, Sam07, Tao05], that is if $(d+1)$-th Gowers norm of f is non-trivial, then f is non-trivially approximable by a degree-d polynomial. This conjecture is easily seen to hold for $d = 1$ and has been proved also for $d = 2$ [GT08, Sam07]. It is of interest to prove this conjecture for higher values of d.

In this paper we show this conjecture, which we will refer to as the 'Inverse Conjecture for the Gowers norm', or, informally, as ICGN, to be false. Let S_n be the elementary symmetric polynomial of degree n in N variables, that is

$$S_n(x) = \sum_{S \subseteq [N],\ |S|=n} \prod_{i \in S} x_i$$

We prove two claims about symmetric polynomials. Note that here and below a constant is *absolute* if it does not depend on N.

First, we show Gowers norms of some symmetric polynomials to be non-trivial.

Theorem 14.1. *There is an absolute positive constant ϵ such that for any prime p*

$$\|S_{2p}\|_{U^{p+2}} > \epsilon,$$

Here S_{2p} is viewed as a function over $\mathbb{F} = \mathbb{F}_p$.

Two versions of this result will be useful later.

- A special case $p = 2$.
$$\|S_4\|_{U^4} > \epsilon \tag{14.1}$$
- An easy generalization: for any $n \geq 2p$,
$$\|S_n\|_{U^{n-p+2}} > \epsilon \tag{14.2}$$

In the second claim we show a specific symmetric polynomial to have no non-trivial approximation by polynomials of lower degree.

Theorem 14.2. *Let $p = 2$. For any polynomial g of degree 3 holds*

$$|\langle S_4, g\rangle| < \exp\{-\alpha N\} \tag{14.3}$$

We conjecture the second claim of the theorem to be true for any prime number p, replacing 3 with $p + 1$ and 4 with $2p$.

The combination of (14.1) and (14.3) shows ICGN to be false for $p = 2$ and $d = 4$.

14.1.1 Related work

Our results have a large overlap with a recent work of Green and Tao [GT07].

The paper of Green and Tao has two parts. In the first part ICGN is shown to be true when f is itself a polynomial of degree less than p and $d < p$. In the second part, the conjecture is shown to be false in general. In particular the symmetric polynomial S_4 is shown to be a counterexample for $p = 2$ and $d = 4$.

To proof of non-approximability of S_4 by lower-degree polynomials in [GT07] uses a modification of a Ramsey-type argument due to Alon and Beigel [AB01]. Very briefly, this argument shows that if a function over $\mathbb{F}_2$ has a non-trivial correlation with a multilinear polynomial of degree d, then its restriction to a subcube of smaller dimension has a non-trivial correlation with a symmetric polynomial of degree d. The problem of inapproximability by symmetric polynomials turns out to be easier to analyze.

This argument gives a somewhat weaker bounds for non-inapproximability of S_4, in that it shows $\langle S_4, g\rangle < \log^{-c}(N)$ for any degree-3 polynomial g and for an absolute constant $c > 0$.

On the other hand, this argument is more robust than our inapproximability argument. We observe below that it can be readily extended to the case of general prime p and, combined with (14.2), show ICGN to be false for all p.

14.1.2 The case of a general prime field

We briefly observe here that a minor adaptation of the Alon-Beigel argument, together with (14.2), show the symmetric polynomial S_{p^2} to have a non-negligible (p^2)-nd Gowers norm over $\mathbb{F}_p$ and to have no good approximation by lower-degree polynomials. In that, S_{p^2} provides a counterexample to ICGN for any prime p.

Indeed, by monotonicity of the Gowers norms ([GT08]), and since $p \geq 2$, a direct implication of (14.2) gives

$$\|S_{p^2}\|_{U^{p^2}} > \epsilon$$

On the other hand, let g be a polynomial of degree less than p^2 in N variables such that $\langle S_{p^2}, g\rangle > \epsilon$. Note that the Alon-Beigel argument (as given in [AB01] and in [GT07]) does not seem to be immediately applicable in this case, since g does not have to be multilinear. A way around this obstacle, is to observe, via an averaging argument, that there is a copy of an N'-dimensional boolean cube $\{0,1\}^{N'}$, such that restrictions S' and g' of S_{p^2} and of g on this subcube satisfy $\langle S', g'\rangle > \epsilon'$, and N', ϵ' depend linearly on N, ϵ'. Without loss of generality assume the coordinates of the boolean cube to be $\{1...N'\}$ and consider the functions S', g' as functions in variables $x_1, ..., x_{N'}$ (with some fixed assignment of values to variables $x_i,\ i > N'$). Now, $S' = \sum_{i=0}^{p^2} a_i S_i$ is a symmetric polynomial of degree p^2 over $\mathbb{F}^{N'}$, with $a_i = 1$, and g' is a polynomial of a degree smaller than p^2. Our gain is in that now g' can be replaced by a multilinear polynomial coinciding with g' on the boolean cube, and hence having a non-trivial correlation with S' on the boolean cube.

Now, the Alon-Beigel argument can be applied to show that the symmetric polynomial S_{p^2} has a non-trivial correlation with a symmetric polynomial h of a smaller degree over the boolean cube $\{0,1\}^{N'}$ viewed as a subset of $\mathbb{F}^{N'}$. This, however, couldn't be true due to a theorem of Lucas, which implies that for a boolean vector x with Hamming weight $w = \sum_{i=1}^{N'} x_i$, the value $S_{p^2}(x)$ depends only on the 3-rd digit in the representation of w in base p, while the value of h depends only on the first 2 digits.

This completes the argument. We conclude with an observation that this argument directly extends to S_{p^k} for any $k > 1$.

Here is a brief overview of the rest of the paper. Section 14.2 defines relevant notions and contains proofs of several technical claims. Theorem 14.1 is proved in Section 14.3. Theorem 14.2 is proved in Section 14.4.

14.2 Some useful notions and claims

14.2.1 Some multilinear polynomials and their properties

In this sub-section we introduce and discuss certain polynomials over the finite field $\mathbb{F}$. These polynomials can be conveniently viewed as multi-linear functions on matrices whose entries are elements of $\mathbb{F}$, or formal variables with values in the field. A basic object we consider is a rectangular $n \times N$ matrix, $N \geq n$. A matrix M with rows $r_1...r_n$ will be denoted by $M[r_1...r_n]$. Sometimes there will be repeated rows. In such a case we consider a partition $\lambda = (\lambda_1...\lambda_k)$ of $[n]$, that is λ_i are (possibly empty) subsets of $[n]$, whose disjoint union is $[n]$. We denote by $M_\lambda[r_1...r_k]$ the matrix whose rows in positions indexed by elements of λ_i

equal r_i. Note that the partition λ is ordered, in that the ordering of the sets λ_i is relevant. We use the notation $\{\lambda_1...\lambda_k\}$ for an unordered partition.

First, we introduce the "symmetric" function $\mathcal{S}$. We define $\mathcal{S}(M)$ to be the sum of all the permanental minors of M, that is

$$\mathcal{S}(M) := \sum_{C\subseteq[N],|C|=n} Per\left(M_C\right),$$

where M_C is an $n \times n$ submatrix of M which is obtained by deleting all the columns of M except these with indices in C.

Let $\lambda = (\lambda_1...\lambda_k)$ be a partition of $[n]$, and set $\ell_i = |\lambda_i|$. Clearly $\mathcal{S}\left(M_\lambda\right)$ depends only on the cardinalities ℓ_i of λ_i. This leads to the notation $M\left[r_1^{(\ell_1)}...r_k^{(\ell_k)}\right]$ which denotes the matrix in which the row r_1 appears ℓ_1 times, followed by ℓ_2 appearances of the row r_2 and so on. In this notation, therefore

$$\mathcal{S}\left(M_{(\lambda_1...\lambda_k)}[r_1...r_k]\right) = \mathcal{S}\left(M\left[r_1^{(|\lambda_1|)}...r_k^{(|\lambda_k|)}\right]\right)$$

The second matrix function we consider is the "forward" function $\mathbb{F}$, with

$$\mathbb{F}(M\left[r_1...r_n\right]) = \sum_{C\subseteq[N],|C|=\{j_1<j_2<...<j_n\}} \prod_{i=1}^{n} r_i\left(j_i\right)$$

Here $r_i(j)$ denote the j-th coordinate of the vector r.

To connect the two notions, observe that

$$\mathcal{S}(M[r_1...r_n]) = \sum_{\sigma} \mathbb{F}(M[r_{\sigma_1}...r_{\sigma_n}])$$

where σ runs over all permutations on n items.

The last function we consider is a "hybrid" function $\mathcal{H}$ which has some 'symmetric' and some 'forward' properties. Let $\lambda = (\lambda_1...\lambda_k)$ be an ordered partition of $[n]$ with k terms. For another such partition $\theta = (\theta_1...\theta_k)$ of $[n]$ write $\theta \sim \lambda$ if $|\theta_1| = |\lambda_1|,...,|\theta_k| = |\lambda_k|$. We define

$$\mathcal{H}\left(\mathcal{M}_\lambda[\nabla_\infty...\nabla_\parallel]\right) = \sum_{\mathcal{C}\subseteq[\mathcal{N}],|\mathcal{C}|=\{|_\infty<|_\in<...<|_\backslash\}} \sum_{\theta\sim\lambda} \prod_{\sqcup=\infty}^{\parallel} \prod_{\rangle\in\theta_\sqcup} \nabla_\sqcup\left(|_\rangle\right)$$

An alternative view of the functions $\mathcal{S}, \mathbb{F}$ and $\mathcal{H}$ might be helpful at this point. Consider the set of *paths* which are one-to-one functions from $[n]$ to $[N]$. Let us call a path ρ monotone on a subset $\{i_1 < i_2 < ... < i_\ell\}$ of $[n]$ if $\rho\left(i_1\right) < \rho\left(i_2\right) < ... < \rho\left(i_\ell\right)$. A path is (fully) monotone if it is monotone on $[n]$. Then, for a partition $\lambda = (\lambda_1...\lambda_k)$ of $[n]$ and an $n \times N$ matrix $M = M_\lambda$,

$$\mathcal{S}(M) = \sum_{\text{all } \rho} \prod_{i=1}^{n} M_{i,\rho(i)}$$

$$\mathbb{F}(M) = \sum_{\text{monotone } \rho} \prod_{i=1}^{n} M_{i,\rho(i)}$$

$$\mathcal{H}(\mathcal{M}) = \sum_{\rho \text{ monotone on } \lambda_1 ... \lambda_k} \prod_{\rangle=\infty}^{\setminus} \mathcal{M}_{\rangle,\rho(\langle)}$$

Note that for the function $\mathcal{H}$, similarly to the symmetric function $\mathcal{S}$, holds

$$\mathcal{H}\left(\mathcal{M}_{(\lambda_\infty ... \lambda_\|)}[\nabla_\infty ... \nabla_\|]\right) = \mathcal{H}\left(\mathcal{M}\left[\nabla_\infty^{(|\lambda_\infty|)} ... \nabla_\|^{(|\lambda_\||)}\right]\right)$$

Observe also that if $\lambda = (\{1\}...\{n\})$ then $\mathcal{S}(M) = \mathcal{H}(\mathcal{M})$. If $\lambda = (\{[n]\})$ then $\mathbb{F}(M) = \mathcal{H}(\mathcal{M})$ and $\mathcal{S}(M) = n! \cdot \mathbb{F}(M) = n! \cdot \mathcal{H}(\mathcal{M})$. For a general $\lambda = (\lambda_0 ... \lambda_k)$

$$\mathcal{S}(M) = \left(\prod_{t=1}^{k} |\lambda_t|!\right) \cdot \mathcal{H}(\mathcal{M}) \tag{14.4}$$

Note that this is an identity in $\mathbb{F}$. In particular, if one of the terms λ_i has cardinality at least p then $\mathcal{S}(M) = 0$ and (14.4) provides no information.

To simplify the notation we will usually write $\mathcal{S}(r_1 ... r_n)$ for $\mathcal{S}(M[r_1 ... r_n])$, $\mathbb{F}_\lambda(r_1 ... r_k)$ for $\mathbb{F}\left(M_\lambda[r_1 ... r_k]\right)$ and so on.

14.2.2 Directional derivatives of symmetric polynomials

The functions we have defined are relevant to the discussion here for two reasons. First, the elementary symmetric polynomial $S_n(x)$ in N variables can be viewed as the forward function $\mathbb{F}$ applied to the matrix $M[x...x]$, where M has n identical rows equal to x. In our notation,

$$S_n(x) = \mathbb{F}_{\{[n]\}}(x)$$

Second, it is possible to write a directional derivative $(S_n)_{y_1 ... y_k}$ of S_n of any order as a combination of values of $\mathbb{F}$ on explicitly defined matrices M whose rows are either the indeterminate x or the directions y_i.

The basic observation here is the following lemma which is straightforward from the definition of directional derivative.

Lemma 14.1. *Let a polynomial $P(x)$ in N variables be given by*

$$P(x) = \mathbb{F}_{(\lambda_0 ... \lambda_k)}(x, y_1 ... y_k)$$

Then

$$P_z(x) = \sum_{A \subset \lambda_0} \mathbb{F}_{(A, \lambda_0 \setminus A, \lambda_1 ... \lambda_k)}(x, z, y_1 ... y_k)$$

In words, when we take the derivative of such a polynomial in direction z, we replace some of the rows which contained x with z.

As a corollary we have a following expression for higher order derivatives of a symmetric polynomial.

Proposition 14.1. *Let $k \leq n$, then*

$$(S_n)_{y_1 \dots y_k}(x) = \sum_{m=0}^{n-k} \sum_{\ell_1 \dots \ell_k \geq 1,\ \sum_i \ell_i = n-m} \mathcal{H}\left(\S^{(\Updownarrow)}, \dagger_{\infty}^{(\ell_\infty)} \dots \dagger_{\|}^{(\ell_\|)}\right)$$

Proof. Iterating Lemma 14.1,

$$(S_n)_{y_1 \dots y_k}(x) = \sum_{\lambda = (\lambda_0, \lambda_1 \dots \lambda_k)} \mathbb{F}_\lambda(x, y_1 \dots y_k)$$

where the summation is over partitions λ such that λ_i are not empty for $i = 1 \dots k$. Rearranging, this is

$$\sum_{m=0}^{n-k} \sum_{\ell_1 \dots \ell_k \geq 1,\ \sum_i \ell_i = n-m} \sum_{\lambda:\ |\lambda_0| = m, |\lambda_1| = \ell_1 \dots |\lambda_k| = \ell_k} \mathbb{F}_\lambda(x, y_1 \dots y_k) =$$

$$\sum_{m=0}^{n-k} \sum_{\ell_1 \dots \ell_k \geq 1,\ \sum_i \ell_i = n-m} \mathcal{H}\left(\S^{(\Updownarrow)}, \dagger_{\infty}^{(\ell_\infty)} \dots \dagger_{\|}^{(\ell_\|)}\right)$$

□

We can give explicit expressions for the coefficients of $(S_n)_{y_1 \dots y_k}(x)$. Fix m indices $j_1 < j_2 < \dots < j_m$ for $0 \leq m \leq n-k$, and let a be the coefficient of $x_{j_1} \cdots x_{j_m}$ in $(S_n)_{y_1 \dots y_k}$.

Corollary 14.1. •

$$a = \sum_{\ell_1 \dots \ell_k \geq 1,\ \sum_i \ell_i = n-m} \mathcal{H}^{\{|_\infty \cdots |_\Updownarrow\}}\left(\dagger_{\infty}^{(\ell_\infty)} \dots \dagger_{\|}^{(\ell_\|)}\right)$$

- *If $k + m + p > n + 1$ then*

$$a = \sum_{\ell_1 \dots \ell_k \geq 1,\ \sum_i \ell_i = n-m} \left(\prod_{i=1}^{k} \ell_i!\right)^{-1} \cdot \mathcal{S}^{\{j_1 \dots j_m\}}\left(y_1^{(\ell_1)} \dots y_k^{(\ell_k)}\right)$$

Here, for a subset of indices $T \subseteq [N]$, $\mathcal{H}^T(\mathcal{M})$ returns the value of the matrix function $\mathcal{H}$ applied to the $n \times (N - |T|)$ matrix obtained from M by deleting columns in T. The function $\mathcal{S}^T(M)$ is defined similarly.

Proof. The first claim is immediate from Proposition 14.1. The second claim follows from the first claim, from (14.4), and from the simple observation that if $k + m + p > n + 1$ then $\ell_i < p$ for $i = 1 \dots k$ in the above summation, which means $\ell_i!$ is invertible in $\mathbb{F}_p$. □

Example 14.1. *The following "toy" example will be relevant for the case of the binary field. It is sufficiently simple to illustrate what's going on behind the cumbersome formulas. Consider* $P = (S_4)_{y,z}$. *Then* P *is a quadratic polynomial and for* $1 \leq i < j \leq N$

$$coef_{x(i)x(j)}(P) = \sum_{k \neq l,\ k,l \notin \{i,j\}} y(k)z(l) = \mathcal{S}^{\{i,j\}}(y,z)$$

Continuing with the same example, note that it convenient to express the symmetric function $\mathcal{S}(y,z)$ via inner products of vectors $y, z, \mathbf{1}$, where $\mathbf{1}$ is the all-1 vector of length N.

$$\mathcal{S}(y,z) = \sum_{k \neq l} y(k)z(l) = \langle y, \mathbf{1} \rangle \cdot \langle z, \mathbf{1} \rangle - \langle yz, \mathbf{1} \rangle$$

Here we take yz to be the vector whose coordinates are point-wise inner products of the coordinates of y and z, that is $(yz)(i) = y(i)z(i)$. Of course, $\langle yz, \mathbf{1} \rangle$ is the same as $\langle y, z \rangle$.

Similarly, we can express the 'incomplete' symmetric function $\mathcal{S}^{\{i,j\}}(y,z)$ via the complete symmetric function $\mathcal{S}(y,z)$ minus forbidden terms, as follows

$$\mathcal{S}^{\{i,j\}}(y,z) = \mathcal{S}(y,z) - \Big(z(i) + z(j)\Big)\langle y, \mathbf{1} \rangle - \Big(y(i) + y(j)\Big)\langle z, \mathbf{1} \rangle + \Big(y(i)z(j) + y(j)z(i)\Big)$$

Note the "inclusion-exclusion" structure in the two expressions above. (To make it even clearer we use "+" and "-" notation, though in the binary field both are, of course, the same.) This structure becomes more evident as we pass to our next order of business, which is expressing, for general n and k, the coefficients of $(S_n)_{y_1 \ldots y_k}$ via inner products of vectors $y_1 \ldots y_k, \mathbf{1}$.

14.2.3 Inclusion-Exclusion formulas for symmetric functions

Some notation: Given m vectors $y_1 \ldots y_m$ and a subset $\tau \subseteq [m]$, let y_τ to be vector whose coordinates are point-wise products of the corresponding coordinates of y_i, $i \in \tau$. Let $\mathcal{S}(y[\tau])$ for the value of the function $\mathcal{S}$ on a matrix with $|\tau|$ rows y_i, $i \in \tau$. Let $\langle y_\tau \rangle$ be the polynomial $\langle y_\tau, \mathbf{1} \rangle = \sum_{j=1}^{N} \prod_{i \in \tau} y_i(j)$.

We start with an auxiliary lemma expressing the incomplete symmetric function $\mathcal{S}^{\{k\}}(r_1 \ldots r_n)$ as a polynomial in the k-th coordinate of the vectors r_i and in complete symmetric functions applied to sub-matrices of $M[r_1 \ldots r_n]$.

Lemma 14.2.

$$\mathcal{S}^{\{k\}}(r_1 \ldots r_n) = \sum_{\tau \subseteq [n]} (-1)^{|\tau|}(|\tau|)! \cdot r_\tau(k) \cdot \mathcal{S}\left(r\Big[[n] \setminus \tau\Big]\right)$$

From now on we assume $r_\emptyset$ *to be the all-*1 *vector, and* $\mathcal{S}(r[\emptyset])$ *to equal* 1.

Proof. The proof is by induction on n. For $n = 1$ both sides equal $\sum_{j=1}^{N} r_1(j) - r_1(k)$.

For $n > 1$, observe that

$$\mathcal{S}^{\{k\}}(r_1 \ldots r_n) = \mathcal{S}(r_1 \ldots r_n) - \sum_{i=1}^{n} r_i(k) \cdot \mathcal{S}^{\{k\}}\left(r\Big[[n] \setminus \{i\}\Big]\right)$$

and the claim is easily verified using the induction hypothesis. □

Now we can state two main claims of this section. The first expresses the complete symmetric function $\mathcal{S}(r_1...r_n)$ via inner products $\langle r_T \rangle$.

Proposition 14.2.

$$\mathcal{S}(r_1...r_n) = \sum_{\lambda=\{\lambda_1...\lambda_m\}} \prod_{t=1}^{m} \left((-1)^{|\lambda_t|-1} (|\lambda_t|-1)! \cdot \langle r_{\lambda_t} \rangle \right)$$

In this summation $\lambda = \{\lambda_1...\lambda_m\}$ runs over all unordered partitions of $[n]$ with non-empty λ_i.

Proof. Again, the proof is by induction on n. For $n = 1$ both sides equal $\sum_{j=1}^{N} r_1(j)$. For $n > 1$ we have

$$\mathcal{S}(r_1...r_n) = \sum_{k=1}^{N} r_n(k) \cdot \mathcal{S}^{\{k\}}(r_1...r_{n-1})$$

Using Lemma 14.2 and the induction hypothesis,

$$\mathcal{S}(r_1...r_n) = \sum_{k=1}^{N} r_n(k) \cdot \sum_{\tau \subseteq [n-1]} (-1)^{|\tau|} (|\tau|)! \cdot r_\tau(k) \cdot \mathcal{S}\left(r\left[[n-1] \setminus \tau \right] \right) =$$

$$\sum_{\tau \subseteq [n-1]} (-1)^{|\tau|} (|\tau|)! \cdot \left\langle r_{\tau \cup [n]} \right\rangle \cdot \mathcal{S}\left(r\left[[n-1] \setminus \tau \right] \right)$$

Consider the summand corresponding to $\tau = [n-1]$. Recall the boundary assumption $\mathcal{S}(r[\emptyset]) = 1$. Hence this summand is $(-1)^{n-1}(n-1)! \cdot \left\langle r_{[n]} \right\rangle$. This summand therefore corresponds to the partition $\lambda = \{[n]\}$ in the claim of the proposition.

For τ a proper subset of $[n-1]$, we use the induction hypothesis to obtain

$$\mathcal{S}(r_1...r_n) = \sum_{\tau \subseteq [n-1]} (-1)^{|\tau|} (|\tau|)! \cdot \left\langle r_{\tau \cup [n]} \right\rangle \cdot \sum_{\theta=\{\theta_1...\theta_l\}} \prod_{t=1}^{l} \left((-1)^{|\theta_t|-1} (|\theta_t|-1)! \cdot \langle r_{\theta_t} \rangle \right) +$$

$$(-1)^{n-1}(n-1)! \cdot \left\langle r_{[n]} \right\rangle$$

Here θ runs over all the unordered partitions of $[n-1] \setminus \tau$ with non-empty θ_i. Observe that each pair (τ, θ) leads to a unique partition $\lambda = \{\lambda_1...\lambda_{l+1}\} = \{\theta_1...\theta_l, \tau \cup [n]\}$ of $[n]$. Rearranging the terms, the last summation can be written as

$$\sum_{\lambda=(\lambda_1...\lambda_m)} \prod_{t=1}^{m} \left((-1)^{|\lambda_t|-1} (|\lambda_t|-1)! \cdot \langle r_{\lambda_t} \rangle \right)$$

completing the proof of the proposition. □

The second claim expresses the incomplete symmetric function $\mathcal{S}^{\{j_1...j_k\}}(r_1...r_n)$ as a polynomial in the missing coordinates $j_1...j_k$ of the vectors r_i and in complete symmetric functions applied to sub-matrices of $M[r_1...r_n]$. Note that Lemma 14.2 is a special case $k = 1$ of this claim.

Proposition 14.3.

$$\mathcal{S}^{\{j_1...j_k\}}(r_1...r_n) = \sum_{\tau=(\tau_1...\tau_k)} \prod_{t=1}^{k} \left((-1)^{|\tau_t|}(|\tau_t|)! \cdot r_{\tau_t}(j_t)\right) \cdot \mathcal{S}\left(r\left[[n] \setminus \cup_t \tau_t\right]\right)$$

Here the summation is on all ordered set systems τ such that the terms τ_t are disjoint subsets of $[n]$. The terms may also be empty.

Proof. The proof is by induction on k and n. The case $k = 1$ is treated in Lemma 14.2.

Consider the case $n = 1$. On one hand $\mathcal{S}^{\{j_1...j_k\}}(r_1) = \sum_{j=1}^{N} r_1(j) - \sum_{t=1}^{k} r_1(j_t)$. We claim that this value can be also represented as

$$\sum_{\tau=(\tau_1...\tau_k)} \prod_{t=1}^{k} \left((-1)^{|\tau_t|}(|\tau_t|)! \cdot r_{\tau_t}(j_t)\right) \cdot \mathcal{S}\left(r\left[[1] \setminus \cup_t \tau_t\right]\right)$$

Here τ_i are disjoint subsets of $[1]$. Observe that there are $k+1$ summands in this expression, corresponding to different set systems τ. Let $\tau^{(0)}$ denote the set system with k empty terms, and let $\tau^{(t)}$, for $t = 1...k$ denote the set system with $\tau_t = \{1\}$ and all the remaining terms are empty. The summand corresponding to $\tau^{(0)}$ is $\mathcal{S}(r_1) = \sum_{j=1}^{N} r_1(j)$. The summand corresponding to $\tau^{(t)}$ is $(-r_1(j_t)) \cdot \mathcal{S}(r_\emptyset) = -r_1(j_t)$, and we are done in this case.

For $k, n > 1$, we have

$$\mathcal{S}^{\{j_1...j_k\}}(r_1...r_n) = \mathcal{S}^{\{j_1...j_{k-1}\}}(r_1...r_n) - \sum_{i=1}^{n} r_i(j_k) \cdot \mathcal{S}^{\{j_1...j_k\}}\left(r\left[[n] \setminus \{i\}\right]\right)$$

By the induction hypothesis, this is

$$\sum_{\theta=(\theta_1...\theta_{k-1})} \prod_{t=1}^{k-1} \left((-1)^{|\theta_t|}(|\theta_t|)! \cdot r_{\theta_t}(j_t)\right) \cdot \mathcal{S}\left(r\left[[n] \setminus \cup_t \theta_t\right]\right) -$$

$$\sum_{i=1}^{n} r_i(j_k) \cdot \sum_{\mu^{(i)}=\left(\mu_1^{(i)}...\mu_k^{(i)}\right)} \prod_{u=1}^{k} \left((-1)^{|\mu_u^{(i)}|}(|\mu_u^{(i)}|)! \cdot r_{\mu_u^{(i)}}(j_u)\right) \cdot \mathcal{S}\left(r\left[[n] \setminus \cup_t \mu_t^{(i)} \setminus \{i\}\right]\right)$$

Here the summation is on all ordered set systems θ such that the terms θ_t are disjoint subsets of $[n]$ and on ordered set systems $\mu^{(i)}$, $i = 1...n$ such that the terms $\mu_u^{(i)}$ are disjoint subsets of $[n] \setminus \{i\}$.

Given a set system $\theta = (\theta_1...\theta_{k-1})$ we define a set system $\tau = (\tau_1...\tau_k)$ by setting $\tau_t = \theta_t$, $t = 1...k-1$ and $\tau_k = \emptyset$. Given a set system $\mu^{(i)} = \left(\mu_1^{(i)}...\mu_k^{(i)}\right)$ we define a set system $\tau = (T_1...\tau_k)$ by setting $\tau_u = \mu_u^{(i)}$, $u = 1...k-1$ and $\tau_k = \mu_k^{(i)} \cup \{i\}$. In both cases we have obtained a set system of the type we want, that is an ordered family of k disjoint subsets of $[n]$. Moreover, each such system with empty k-th term is obtained exactly once, from the corresponding θ-system, and each system with non-empty k-th term τ_k is obtained exactly

$|\tau_k|$ times, from systems $\mu^{(i)}$ with $i \in \tau_k$. Rearranging the terms and the signs, the last expression is precisely

$$\sum_{\tau=(\tau_1 \ldots \tau_k)} \prod_{t=1}^{k} \left((-1)^{|\tau_t|}(|\tau_t|)! \cdot r_{\tau_t}(j_t)\right) \cdot \mathcal{S}\left(r\left[[n] \setminus \cup_t \tau_t\right]\right),$$

completing the proof. □

14.2.4 Some properties of Gowers' norms

The main result in this subsection shows that if a function from $\mathbb{F}^N$ to $\mathbb{F}$ is fixed on a subset of $\mathbb{F}^N$ defined by low-degree polynomial constraints, then it has a non-trivial Gowers norm of an appropriate order.

Recall that for a vector $x \in \mathbb{F}^N$, x^i stands for a vector in $\mathbb{F}^N$ whose coordinates are i-th powers of the coordinates of x.

Proposition 14.4. *Let K be an absolute constant. Let $y_{i,j}$, $i = 1...p-1$, $j = 1...K$, be $K(p-1)$ vectors in $\mathbb{F}^N$. Let M be a subset of $\mathbb{F}^N$ defined by the constraints $\langle x^i, y_{i,j}\rangle = 0$ for all i, j.*

Let f be a function from $\mathbb{F}^N$ to $\mathbb{F}$. Assume that f is fixed on M. Then

$$\|f\|_{U^p} > \left(\frac{|M|}{2^N}\right)^2 =: Pr^2\{M\}$$

Proof. Let $f_{|M} \equiv c_0$.

Consider a subspace V of polynomials of degree at most $p-1$ in $\mathbb{F}[x_1 ... x_N]$ spanned by the polynomials $\langle x^i, y_{i,j}\rangle$, for all i, j. We will first find a polynomial $g \in V$ such that $|\langle f, g\rangle| \geq Pr\{M\}$. This, combined with a lemma from [GT08], will imply the claim of the proposition.

Let $\mathbf{b} = (b_{i,j})$, $i = 1...p-1$, $j = 1...K$, be a matrix with entries in $\mathbb{F}$. Let $c \in \mathbb{F}$. Set

$$\mu(\mathbf{b}, c) = Pr\left\{x : \ f(x) = c \ \wedge \ \langle x^i, y_{i,j}\rangle = b_{i,j} \text{ for all } i, j\right\}$$

Note that, by assumption, for a zero matrix $\mathbf{b}$ holds $\mu(\mathbf{b}, c_0) = Pr\{M\}$. In other words, $\mu(\mathbf{b}, c) = 0$ and for $\mathbf{b} = 0$ any $c \neq c_0$.

Now, for any $g(x) = \sum_{i,j} a_{i,j} \langle x^i, y_{i,j}\rangle$ in V holds

$$\langle f, g\rangle = \mathbb{E}e(f - g) = \sum_{\mathbf{b}, c} \mu(\mathbf{b}, c) \cdot e(c - \langle \mathbf{a}, \mathbf{b}\rangle)$$

where $\mathbf{a} = (a_{i,j})_{i,j}$ and $\langle \mathbf{a}, \mathbf{b}\rangle = \sum_{i,j} a_{i,j} b_{i,j}$. Averaging over V, we have

$$\mathbb{E}_{g \in V} \langle f, g\rangle = \frac{1}{|V|} \sum_{\mathbf{a}} \sum_{\mathbf{b}, c} \mu(\mathbf{b}, c) \cdot e(c - \langle \mathbf{a}, \mathbf{b}\rangle) = \frac{1}{|V|} \sum_{\mathbf{b}, c} \mu(\mathbf{b}, c) \cdot e(c) \sum_{\mathbf{a}} e(-\langle \mathbf{a}, \mathbf{b}\rangle) =$$

$$\sum_{c} \mu(0, c) \cdot e(c) = \mu(0, c_0) \cdot e(c_0) = Pr\{M\} \cdot e(c_0)$$

This means, there is $g \in V$ with $|\langle f, g\rangle| \geq Pr\{M\}$. We conclude the proof of the proposition by quoting a lemma from [GT08], which states that $|\langle f, g\rangle| \geq \epsilon$ implies $\|f\|_{U^p} \geq \epsilon$. □

14.2.5 Asymptotic uniformity and independence of some random variables

In this subsection we deal with another property of multiviarite polynomials. Let n be fixed integer and let N be an integer parameter growing to infinity. Let $r_1...r_n$ be n vectors in $\mathbb{F}^N$. Let $\kappa = (k_1...k_n)$ be a non-zero sequence of integers $0 \leq k_i < p$. For each such sequence define a polynomial $X_\kappa(r_1, ..., r_n) = \sum_{j=1}^{N} \prod_{i=1}^{n} r_i^{k_i}(j)$.

Now,let $r_1...r_n$ be chosen uniformly and independently from $\mathbb{F}^N$. We claim that for a large N the random variables $X_\kappa(r_1, ..., r_n)$ are nearly independent and uniformly distributed over $\mathbb{F}$. Let $X = (X_\kappa)_\kappa$, and let $K = p^n$.

Proposition 14.5. *Let U be the uniform distribution on $\mathbb{F}^K$. Let P be distribution of X on $\mathbb{F}^K$. Let $\|\cdot\|$ denote the statistical (l_1) distance between distributions.*

Then there is a constant $c > 0$ depending on n, p but not on N such that

$$\|P - U\| \leq \exp\{-cN\}$$

Proof. We start from a simple observation that Fourier transform of a uniform distribution is the delta function at 0. In addition, the two following statements are equivalent up to constants: 'a distribution is exponentially close to uniform' and 'all non-zero Fourier coefficients of the distribution are exponentially close to zero'. Accordingly, we will show that all the non-zero Fourier coefficients of P tend exponentially fast in N to zero.

Consider a character $\chi(y) = \xi^{\langle y,a\rangle}$, corresponding to a non-zero vector $a = (a_\kappa)_\kappa \in \mathbb{F}^K$. (Recall that $\xi = e^{2\pi i/p}$ is the p-th primitive root of unity.) Then, normalizing appropriately,

$$\widehat{P}(\chi) = \sum_y P(y)\bar{\chi}(y) = \sum_y Pr\{X = y\} \cdot \xi^{-\sum_\kappa a_\kappa y_\kappa} = \mathbb{E}\xi^{-\sum_\kappa a_\kappa X_\kappa}$$

Let P_a denote the distribution of the random variable $X_a = \sum_\kappa a_\kappa X_\kappa$. Then we have shown $\widehat{P}(\chi) = \widehat{P_a}(1)$. We will show the non-zero Fourier coefficients of P_a to be exponentially small, completing the proof of the proposition.

We have

$$X_a(r_1, ..., r_n) = \sum_\kappa a_\kappa P_\kappa(r_1, ..., r_n) = \sum_{j=1}^{N} \sum_{\kappa=(k_1...k_n)} a_\kappa \prod_{i=1}^{n} r_i^{k_i}(j)$$

Let x_i be elements of the field $\mathbb{F}$. Consider an n-variate polynomial

$$Q(x_1...x_n) = \sum_{\kappa=(k_1...k_n)} a_\kappa \prod_{i=1}^{n} x_i^{k_i}$$

Since not all of the coefficients a_κ are zero, and since all κ are non-zero sequences, Q is a multi-variate polynomial of degree at least 1 in $\mathbb{F}[x_1...x_n]$, and therefore attains at least two values with probability bounded away from zero. Now, $X_a = \sum_{j=1}^{N} Q(r_1(j)...r_n(j))$ is a sum of N independent copies of Q. Let μ denote the distribution of Q on $\mathbb{F}$. Then the distribution P_a of X_a is μ^{*N}, the N-wise convolution of μ with itself. Since p is prime, $\widehat{\mu}(0) = 1$, and $|\widehat{\mu}| < 1$ everywhere else. Therefore, $\widehat{P_a} = (\widehat{\mu})^N$ tends to the delta function at 0 exponentially fast in N, completing the proof. □

14.2.6 Estimates on the number of common zeroes of some families of polynomials

The main claim of this subsection is the following proposition.

Proposition 14.6. *Let M be the ring of $\mathbb{F}$-valued functions on $\mathbb{F}^N$, that is $M = \mathbb{F}[x_1...x_N]/I$, where I is the ideal $(x_1^p - x, ..., x_N^p - x)$. Let $f_1...f_K$ be polynomials in M. Let S be the set of common zeroes of $f_1...f_K$, that is*

$$S = \left\{ u \in \mathbb{F}^N : \ f_1(u) = ... = f_K(u) = 0 \right\}$$

Then

$$|S| \leq dim\,(M/J)$$

where J is the ideal generated by $\{f_i\}$, and $dim\,(M/J)$ denotes the dimension of $dim\,(M/J)$, viewed as a vector space over $\mathbb{F}$.

Proof. For each $u \in S$, let $q_u \in M$ be defined by $q_u(u) = 1$ and $q_u(v) = 0$ for all $v \neq u$. We will show that the family $\{q_u + J\}_{u \in S}$ is linearly independent in M/J. This will immediately imply the claim of the proposition.

Consider a linear combination $q = \sum_{u \in S} \lambda_u q_u$ such that $q \in J$. Let $v \in S$. We compute $q(u)$ in two ways. First, since $q \in J$, we have $q(v) = 0$. On the other hand, $q(v) = \sum_{u \in S} \lambda_u q_u(v) = \lambda_v$. This shows $\lambda_v = 0$ for all $v \in S$, completing the proof. □

In some cases, the dimension of M/J is easy to estimate.

Lemma 14.3. *Let $p = 2$, let $K = \binom{N}{k}$, and let $\{f_I\}$ be indexed by k-subsets I of $[N]$. Assume that for any such subset I holds*

$$deg\left(f_I(x) - \prod_{i \in I} x_i \right) \leq k - 1 \tag{14.5}$$

Then,

$$dim\,(M/J) \leq \sum_{j=0}^{k-1} \binom{N}{j}$$

Proof. We will construct a generating subset of the vector space M/J of cardinality at most $\sum_{j=0}^{k-1} \binom{N}{j}$. We start from a trivial generating set $\{m + J\}$, where m runs through all the 2^N multi-linear monomials in N variables. Now, in the factor space M/J, we can replace any product of k variables, $\prod_{i \in I} x_i$, by a polynomial of degree smaller than k. Iterating this procedure, we arrive to a generating set spanned by $\{s + J\}$, where s now runs through $\sum_{j=0}^{k-1} \binom{N}{j}$ monomials of degree at most $k - 1$. □

14.3 Proof of Theorem 14.1

We need to show that

$$\|S_{2p}\|_{U^{p+2}} > \epsilon$$

for an absolute constant ϵ.

We remark that (14.2) can be shown exactly in the same way, replacing $2p$ with n and $p+2$ with $n-p+2$ throughout.

Recall ([GT08]) that $\|f\|_{U^{p+2}} = \mathbb{E}_{y,z}^{1/2^{p+2}} \|f_{y,z}\|_{U^p}^{2^p}$. Since the Gowers' norms are nonnegative, it will suffice to show that $\|f_{y,z}\|_{U^p}$ is non-negligible for a non-negligible fraction of directions y, z.

Let

$$A = \left\{(y,z) : \ \left\langle y^a, z^b \right\rangle = 0 \text{ for all } 0 \le a, b < p\right\}$$

By Proposition 14.5, for uniformly and independently chosen directions y, z, and for a sufficiently large N, the probability of A is very close to p^{-p^2}. Therefore, A is a non-negligible event. We will now show that for any $(y,z) \in A$ holds $\|f_{y,z}\|_{U^p} > \epsilon'(y,z)$, for an appropriate function ϵ'.

Fix (y,z) in A. Let $f = (S_{2p})_{y,z}$. Let

$$M = M(y,z) = \left\{x : \ \left\langle x^i, y^a z^b \right\rangle = 0 \text{ for all } 1 \le i \le p-1, 0 \le a, b < p\right\}$$

We will show that f is fixed on M. Assuming this, by Proposition 14.4, we have $\|f_{y,z}\|_{U^p} > Pr^2\{M\}$, and therefore

$$\|f\|_{U^{p+2}}^{2^{p+2}} = \mathbb{E}_{y,z}\|f_{y,z}\|_{U^p}^{2^p} \ge Pr\{A\} \cdot \mathbb{E}_{(y,z)\in A} Pr^{2^{p+1}}\{M(y,z)\} \ge$$

$$Pr\{A\} \cdot \mathbb{E}_{(y,z)\in A}^{2^{p+1}} Pr\{M(y,z)\} \ge \left(Pr\{A\} \cdot \mathbb{E}_{(y,z)\in A} Pr\{M(y,z)\}\right)^{2^{p+1}} =$$

$$Pr^{2^{p+1}}\left\{x : \ \left\langle x^i y^a z^b \right\rangle = 0 \text{ for all } 0 \le a, b, i \le p-1\right\} \ge \Omega\left(p^{-p^3 \cdot 2^{p+1}}\right)$$

The last inequality follows from Proposition 14.5, since random variables $\left\langle x^i y^a z^b \right\rangle$ are asymptotically uniform and independent.

It remains to prove the following fact.

Lemma 14.4. *Let x, y, z be three vectors in $\mathbb{F}^N$ satisfying $\left\langle x^i y^a z^b \right\rangle = 0$ for all $0 \le a, b, i \le p-1$. Then*

$$(S_{2p})_{y,z}(x) = \mathcal{H}\left(\dagger^{(\surd)}, \ddagger^{(\surd)}\right)$$

Proof. By Proposition 14.1,

$$(S_{2p})_{y,z}(x) = \sum_{m=0}^{2p-2} \sum_{a,b\ge 1,\ a+b=2p-m} \mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right)$$

We claim that all of the summands on the right, except (possibly) $\mathcal{H}\left(\dagger^{(\surd)}, \ddagger^{(\surd)}\right)$ are 0.

There are two possible cases to consider. The easier case is when $a, b, m < p$. In such a case, by (14.4), $\mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right)$ is proportional to $\mathcal{S}\left(x^{(m)}, y^{(a)}, z^{(b)}\right)$. By Proposition 14.2, the symmetric function $\mathcal{S}\left(x^{(m)}, y^{(a)}, z^{(b)}\right)$ is a polynomial in $\left\langle x^i y^a z^b \right\rangle$, which vanishes when all of these inner products are 0.

In the second case, one of the indices a, b, m is at least p. Note, that there could be at most one such index (barring the case $a = b = p$). We may assume this index is m. We claim that in this case $\mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right)$ can be written as a linear combination of hybrid functions $\mathcal{H}\left(\S^{(\ell)}, \nabla_\infty, ..., \nabla_{\Updownarrow - \ell}\right)$, where $\ell < m$ and the vectors r_i are of the form $x^\alpha y^\beta z^\gamma$. Note that this will suffice to prove the lemma, since iterating this step will express $\mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right)$ as a linear combination of symmetric functions in r_i, and these functions vanish.

Consider $\mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right)$. For notational convenience, let $w_1 ... w_{a+b}$ stand for the vectors $y...y, z...z$ (y taken a times and z taken b times). Note that both a and b are smaller than p. Using Corollary 14.1 and Proposition 14.3,

$$\mathcal{H}\left(\S^{(\Updownarrow)}, \dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right) = \left(\dashv! \cdot \lfloor!\right)^{-\infty} \cdot \sum_{\rangle_\infty < \rangle_\in < ... < \rangle_\Updownarrow} \S_{\rangle_\infty} \S_{\rangle_\in} \cdots \S_{\rangle_\Updownarrow} \mathcal{S}^{\{\rangle_\infty ... \rangle_\Updownarrow\}}\left(\dagger^{(\dashv)}, \ddagger^{(\lfloor)}\right) =$$

$$(a! \cdot b!)^{-1} \cdot \sum_{i_1 < i_2 < ... < i_m} x_{i_1} x_{i_2} \cdots x_{i_m} \cdot \sum_{\tau = (\tau_1 ... \tau_m)} \prod_{t=1}^{m} \left((-1)^{|\tau_t|} (|\tau_t|)! \cdot w_{\tau_t}(i_t)\right) \cdot \mathcal{S}\left(w\left[[a+b] \setminus \cup_t \tau_t\right]\right)$$

Here the inner summation is on all ordered set systems τ such that the terms τ_t are disjoint subsets of $[a+b]$. The terms may also be empty.

Let us attempt to simplify the double summation we obtained. First, we may disregard the constant term $(a! \cdot b!)^{-1}$. Next, observe that, as before, all symmetric functions of the form $\mathcal{S}(w[T])$ vanish, unless T is empty, in which case they equal 1. Therefore, we may consider the double summation

$$\sum_{i_1 < i_2 < ... < i_m} x_{i_1} x_{i_2} \cdots x_{i_m} \cdot \sum_{\tau = (\tau_1 ... \tau_m)} \prod_{t=1}^{m} \left((-1)^{|\tau_t|} (|\tau_t|)! \cdot w_{\tau_t}(i_t)\right)$$

Here the inner summation is on all ordered partitions τ of $[a+b]$. The terms τ_t may also be empty. Changing the order of summation, and ignoring the constant term $(-1)^{a+b}$, we get

$$\sum_{\tau = (\tau_1 ... \tau_m)} \prod_{t=1}^{m} (|\tau_t|)! \cdot \sum_{i_1 < i_2 < ... < i_m} \prod_{t=1}^{m} (x \cdot w_{\tau_t})(i_t) = \sum_{\tau = (\tau_1 ... \tau_m)} \left(\prod_{t=1}^{m} (|\tau_t|)!\right) \cdot \mathbb{F}\left(x w_{\tau_1}, x w_{\tau_2}, ..., x w_{\tau_m}\right)$$

Consider the last expression. Let us use some more notation. For an ordered partition $\tau = (\tau_1 ... \tau_m)$, let $n = n(\tau)$ be the number of empty terms. Let $\{\tau_1 ... \tau_m\}$ denote the unordered version of this partition, where the first $n(\tau)$ terms are taken, by agreement, to be the empty ones. Then we can rewrite this expression as

$$\sum_{\tau = \{\tau_1 ... \tau_m\}} \left(\prod_{t=1}^{m} (|\tau_t|)!\right) \cdot \mathcal{H}\left(\S^{(\setminus)}, \S_{\sqsupseteq \tau_{\setminus + \infty}}, ..., \S_{\sqsupseteq \tau_\Updownarrow}\right)$$

Now, clearly not all the terms in the partition are empty and, therefore, $n(\tau) < m$ for all τ, completing the proof of our last claim, of the lemma, and of the theorem. $\square$

14.4 Proof of Theorem 14.2

Let $p = 2$. We will show there is an absolute constant $\alpha > 0$ such that for any polynomial g of degree at most 3 in N variables holds

$$\langle S_4, g\rangle < \exp\{-\alpha N\}$$

A first step is to observe that there is a relation between the inner product of two functions and the average inner product of their derivatives.

Lemma 14.5. *For any two functions f and g holds*

$$\langle f, g\rangle^4 \leq \mathbb{E}_y \langle f_y, g_y\rangle^2$$

Proof. This is an immediate corollary of a lemma in [Sam07], but we give the elementary proof for completeness. By the Cauchy-Schwarz inequality,

$$\mathbb{E}_y \langle f_y, g_y\rangle^2 \geq \mathbb{E}_y^2 \langle f_y, g_y\rangle = \mathbb{E}_{x,y}^2 (-1)^{f(x)+f(x+y)+g(x)+g(x+y)} = \mathbb{E}^4 (-1)^{f(x)+g(x)} = \langle f, g\rangle^4$$

□

Corollary 14.2.

$$\langle f, g\rangle^8 \leq \mathbb{E}_{y,z} \langle f_{y,z}, g_{y,z}\rangle^2$$

We will show that for any polynomial g of degree at most 3 holds $\mathbb{E}_{y,z} \left\langle (S_4)_{y,z}, g_{y,z} \right\rangle^2 \leq \exp\{-\alpha N\}$. First, here is a brief overview of the argument.

The point is that taking second derivatives makes life easier, since a second derivative of g is a linear function, and a second derivative of S_4 is a quadratic. We therefore need to show that for the large majority of directions y, z, the quadratic function $(S_4)_{y,z}$ has a small inner product with the linear function $(-1)^{g_{y,z}}$. In this we will be helped by a theorem of Dixon giving a structural description of quadratic polynomials, which, in particular, characterizes the Fourier transform of functions of the type $(-1)^Q$, where Q is a quadratic. In fact, setting $Q = (S_4)_{y,z}$ we will see that for many of the directions y, z the Fourier coefficients of $(-1)^Q$ will be exponentially small. For the remaining directions, these Fourier coefficients will be supported on an explicit easy to describe 3-dimensional affine subspace depending on y, z. We will then argue that for any fixed polynomial g of lower degree, the support of the character $(-1)^{g_{y,z}}$ lies in this affine subspace with exponentially small probability over y, z.

We proceed with computing the second derivative $Q = (S_4)_{y,z}$.

14.4.1 Second derivatives of S_4

Write $Q(x) = \sum_{i<j} q_{i,j} x(i)x(j) + \sum_i \ell_i x(i) + c$.

By Proposition 14.1 or by Example 14.1.

$$q_{i,j} = \mathcal{S}(y, z) - \langle y, \mathbf{1}\rangle \cdot \Big(z(i) + z(j)\Big) + \langle z, \mathbf{1}\rangle \cdot \Big(y(i) + y(j)\Big) + \Big(y(i)z(j) + y(j)z(i)\Big)$$

At this point we invoke (a corollary of) a theorem of Dixon [MS83]:

Theorem 14.3. *Let $Q(x) = \sum_{i<j} q_{i,j}x(i)x(j) + \sum_i \ell_i x(i) + c$ be a quadratic polynomial over $\mathbb{F}_2$. Consider the symmetric matrix with zeros on the diagonal and off-diagonal entries given by $S_{i,j} = S_{j,i} = q_{i,j}$. Let the rank of $B = 2h$ (it is always even). Then the function $(-1)^Q$ has 2^{2h} non-zero Fourier coefficients of absolute value 2^{-h}. Moreover, all these coefficients lie in an $2h$-dimensional affine subspace of $\mathbb{F}_2^n$.*

Consider the matrix B in our case. Some notation: let J be the matrix with 0 on the diagonal and 1 off the diagonal. Let $u \otimes v$ denote the outer product uv^t. Then,

$$B = \mathcal{S}(y,z) \cdot J + \langle y, \mathbf{1}\rangle \cdot \Big(z \otimes \mathbf{1} + \mathbf{1} \otimes z\Big) + \langle z, \mathbf{1}\rangle \cdot \Big(y \otimes \mathbf{1} + \mathbf{1} \otimes y\Big) + \Big(y \otimes z + z \otimes y\Big)$$

Since the rank of J is at least $N-1$ and the rank of the remaining matrices is at most 2, the matrix B is almost of full rank if $\mathcal{S}(y,z) = 1$. In this case, by Theorem 14.3, the Fourier coefficients of $(-1)^Q$ are exponentially small.

We therefore may assume $\mathcal{S}(y,z) = 0$. In this case the quadratic part of Q may be written as

$$\sum_{i<j} q_{i,j}x(i)x(j) = \langle y, \mathbf{1}\rangle \cdot \langle x, \mathbf{1}\rangle \langle x, z\rangle + \langle z, \mathbf{1}\rangle \cdot \langle x, \mathbf{1}\rangle \langle x, y\rangle + \Big(\langle x, y\rangle \langle x, z\rangle + \langle x, yz\rangle \Big)$$

Recall that yz denotes the pointwise product of vectors y and z.

This implies the non-zero Fourier coefficients of $\sum_{i<j} q_{i,j}x(i)x(j)$ lie in a 3-dimensional affine subspace of $\mathbb{F}_2^n$. The linear part of this subspace is spanned by the vectors $y, z, \mathbf{1}$ and it is shifted by a vector yz.

Next, consider the linear part $\sum_i \ell(i)x(i)$ of Q. By Proposition 14.1,

$$\ell(i) = \mathcal{H}^{\{\}}\left(\dagger^{(\in)}, \ddagger\right) + \mathcal{H}^{\{\}}\left(\dagger, \ddagger^{(\in)}\right) =$$

$$\sum_{j<k<l\neq i} \Big(y(k)y(l)z(j) + y(j)y(l)z(k) + y(j)y(k)z(l)\Big) + \Big(y(j)z(k)z(l) + y(k)z(j)z(l) + y(l)z(j)z(k)\Big)$$

This can be directly verified to be equal to

$$\Big(\mathcal{S}(y,z) + \mathcal{S}(z,z) + \langle z, \mathbf{1}\rangle\Big) \cdot y(i) + \Big(\mathcal{S}(y,z) + \mathcal{S}(y,y) + \langle y, \mathbf{1}\rangle\Big) \cdot z(i) +$$

$$\Big(\mathcal{S}(y,y) \cdot \langle z, \mathbf{1}\rangle + \mathcal{S}(z,z) \cdot \langle y, \mathbf{1}\rangle + \langle y, z\rangle \cdot \langle y + z, \mathbf{1}\rangle\Big)$$

By assumption, $\mathcal{S}(y,z) = \langle y, \mathbf{1}\rangle \cdot \langle z, \mathbf{1}\rangle + \langle y, z\rangle = 0$. Note that this also implies $\langle y, z\rangle \cdot \langle y + z, \mathbf{1}\rangle = 0$, implying

$$\ell(i) = \Big(\mathcal{S}(z,z) + \langle z, \mathbf{1}\rangle\Big) \cdot y(i) + \Big((S(y,y) + \langle y, \mathbf{1}\rangle\Big) \cdot z(i) + \Big(\mathcal{S}(y,y) \cdot \langle z, \mathbf{1}\rangle + \mathcal{S}(z,z) \cdot \langle y, \mathbf{1}\rangle\Big)$$

Consequently, the linear part of Q may be written as

$$\sum_i \ell(i)x(i) =$$

$$\Big(\mathcal{S}(z,z) + \langle z, \mathbf{1}\rangle\Big) \cdot \langle x, y\rangle + \Big((S(y,y) + \langle y, \mathbf{1}\rangle\Big) \cdot \langle x, z\rangle + \Big(\mathcal{S}(y,y) \cdot \langle z, \mathbf{1}\rangle + \mathcal{S}(z,z) \cdot \langle y, \mathbf{1}\rangle\Big) \cdot \langle x, \mathbf{1}\rangle$$

This means that the non-zero Fourier coefficients of the polynomial $Q = \sum_{i<j} q_{i,j}x(i)x(j) + \sum_i \ell(i)x(i) + c$ lie in the affine subspace $AF_{y,z} = yz + \text{Span}\,(y, z, \mathbf{1})$.

14.4.2 Second derivatives of a fixed polynomial of degree 3

Let

$$g(x) = \sum_{i<j<k} a_{i,j,k} x(i)x(j)x(k)$$

be a polynomial of degree 3. For directions $y, z \in \mathbb{F}^N$, consider the second derivative $g_{y,z} = \sum_i v_{y,z}(i)x(i) + c_{y,z}$. We need to show that the probability of the vector $v_{y,z}$ falling in the affine space $AF_{y,z} = yz + \text{Span}\,(y, z, \mathbf{1})$ is exponentially small.

First, some notation. For $1 \le i \le N$, let G_i be a symmetric $N \times N$ matrix over $\mathbb{F}$ with $(G_i)_{j,k} = (G_i)_{k,j} = a_{i,j,k}$ for all $j \neq k$. (Here we think about $\{i, j, k\}$ as an unordered subset of $[N]$.) The diagonal entries of G_i are set to 0. For future use note the important property $(G_i)_{j,k} = (G_j)_{i,k} = (G_k)_{i,j}$.

These matrices are relevant because they describe the vector $v_{y,z}$.

Lemma 14.6.

- $$v_{y,z}(i) = coef_{x(i)}\left(g_{y,z}(x)\right) = \langle y, G_i z\rangle$$
- *An alternative representation of $v_{y,z}$ will be more convenient for us. For $z \in \mathbb{F}^N$, let $G(z) = \sum_{i=1}^N z(i)G_i$. Then*
$$v_{y,z} = G(z) \cdot y$$

Proof. For the first claim of the lemma, by linearity of the derivative, it suffices to consider the monomial $g(x) = x(i)x(j)x(k)$. This case can be easily verified directly.

For the second claim, note that

$$(G(z)\cdot y)(l) = \sum_{k=1}^N \left(G(z)\right)_{k,l} y(k) = \sum_{k=1}^N y(k)\cdot \sum_{i=1}^N z(i)\left(G_i\right)_{k,l} = \sum_{k=1}^N y(k)\cdot \sum_{i=1}^N \left(G_l\right)_{k,i} z(i) = \langle y, G_l z\rangle$$

□

Consider the event $\{v_{y,z} \in AF_{y,z}\}$. This means $v_{y,z} = yz + u_{y,z}$, for some vector $u_{y,z} \in \text{Span}(y, z, \mathbf{1})$. There are only 8 possible choices for $u_{y,z}$. For convenience, let us assume, without loss of generality (as can be easily seen from the proof), that $u_{y,z} = y + z + \mathbf{1}$ is the most popular one. By the lemma, the event $\{v_{y,z} = yz + u_{y,z}\}$ is the same as $\{G(z) \cdot y = yz + u_{y,z}\}$. To simplify things some more, let $A_i = G_i + e_i \otimes e_i$, $i = 1...N$. That is, $A_i = G_i$ but for $(A_i)_{i,i} = 1$. Let $A(z) = \sum_{i=1}^N z(i)A_i$. Note that $A(z) \cdot y = G(z) \cdot y + yz$. Hence $\{G(z) \cdot y = yz + u_{y,z}\}$ is the same as $\{A(z) \cdot y = u_{y,z} = y + z + \mathbf{1}\}$

We conclude the proof by a technical claim.

Proposition 14.7. *Let $\{A_i\}$, $i = 1...N$ be a family of symmetric $N \times N$ matrices over $\mathbb{F}$ with $A_i(k, k) = \delta_{ik}$. Then, for y, z uniformly at random and independently from $\mathbb{F}^N$,*

$$Pr_{y,z}\Big\{ \left(A(z)\right) \cdot y = y + z + \mathbf{1}\Big\} \le \left(\frac{3}{4}\right)^N$$

The proof of the proposition is based on the claim that the rank of a matrix $A(z)$ is typically large.

Lemma 14.7. *Let matrices $\{A_i\}$ be as in the proposition. Let C be any fixed symmetric $N \times N$ matrix. Then*

$$Pr_z\Big\{ rank(A(z)+C) \leq k-1 \Big\} \leq \frac{1}{2^N} \cdot \sum_{i=0}^{k-1} \binom{N}{i}.$$

Proof. Consider a family of $\binom{N}{k}$ polynomials f_I on $\mathbb{F}^N$. These polynomials are indexed by k-subsets of $[N]$. For a k-subset I, let $f_I(z)$ be the determinant of the $I \times I$ minor of $A(z)+C$. Clearly, rank of $A(z)+C$ is smaller than k if and only if z is a joint zero of $\{f_I\}$.

We now claim that the coefficient of $\prod_{i \in I} z_i$ in $f_I(z)$ is 1. If this is true, $\deg(f_I - \prod_{i \in I} z_i) \leq k-1$, and the claim of the lemma will follow from Lemma 14.7.

Let $B(z) = A(z) + C$. Since we are working in characteristic two, the symmetry of $B(z)$ implies that

$$\det B(z) = \sum_{\sigma \in S_N:\ \sigma = \sigma^{-1}} \prod_{i=1}^{N} B_{i\sigma(i)}(z) =$$

$$\sum_{\sigma \in S_N:\ \sigma = \sigma^{-1}} \prod_{\{i:\sigma(i)=i\}} (z_i + C_{i,i}) \cdot \prod_{\{i:i<\sigma(i)\}} B_{i\sigma(i)}(z) = \prod_{i \in I} z_i + \text{lower order terms.}$$

In the second equality we use the identity $B^2_{i\sigma(i)}(z) = B_{i\sigma(i)}(z)$ in $\mathbb{F}$. □

Let I denote the identity $N \times N$ matrix.

Let $p(z) = Pr_y\Big\{ A(z) \cdot y = y + z + \mathbf{1} \Big\}$. Clearly $p(z) \leq 2^{-\text{rank}(A(z)+I)}$. By Lemma 14.7,

$$Pr_{y,z}\Big\{ (A(z)) \cdot y = y + z + \mathbf{1} \Big\} = \mathbb{E}_z p_z \leq \mathbb{E}_z 2^{-\text{rank}(A(z)+I)} \leq \frac{1}{2^N} \sum_{k=0}^{N} \binom{N}{k} 2^{-k} = \left(\frac{3}{4}\right)^N$$

This concludes the proof of the proposition, and of Theorem 14.2.

Chapter 15

Lower bound for adaptive linearity tests

Linearity tests are randomized algorithms which have oracle access to the truth table of some function f, and are supposed to distinguish between linear functions and functions which are far from linear. Linearity tests were first introduced by Blum, Luby and Rubenfeld [BLR93], and were later used in the PCP theorem, among other applications. The quality of a linearity test is described by its correctness c - the probability it accepts linear functions, its soundness s - the probability it accepts functions far from linear, and its query complexity q - the number of queries it makes.

Linearity tests were studied in order to decrease the soundness of linearity tests, while keeping the query complexity small (for one reason, to improve PCP constructions). Samorodnitsky and Trevisan [ST00] constructed the Complete Graph Test, and prove that no Hyper Graph Test can perform better than the Complete Graph Test. Later [ST06] they prove, among other results, that no non-adaptive linearity test can perform better than the Complete Graph Test. Their proof uses the algebraic machinery of the Gowers Norm. A result by Ben-Sasson, Harsha and Raskhodnikova [BSHR05] allows to generalize this lower bound also to adaptive linearity tests.

We also prove the same optimal lower bound for adaptive linearity test, but our proof technique is arguably simpler and more direct than the one used in [ST06]. We also study, like [ST06], the behavior of linearity tests on quadratic functions. However, instead of analyzing the Gowers Norm of certain functions, we provide a more direct combinatorial proof, studying the behavior of linearity tests on random quadratic functions. This proof technique also lets us prove directly the lower bound also for adaptive linearity tests.

15.1 Introduction

We study the relation between the number of queries and soundness of adaptive linearity tests. A linearity test (over the field $\mathbb{F}_2$ for example) is a randomized algorithm which has oracle access to the truth table of a function $f : \{0,1\}^n \to \{0,1\}$, and needs to distinguish between the following two extreme cases:

1. f is linear

2. f is far from linear functions

A function f is called *linear* if it can be written as $f(x_1, ..., x_n) = a_1x_1 + ... + a_nx_n$, with $a_1, ..., a_n \in \mathbb{F}_2$. The agreement of two functions $f, g : \{0,1\}^n \to \{0,1\}$ is defined as $d(f,g) = |\mathbb{P}_{\mathbf{x}}[f(\mathbf{x}) = g(\mathbf{x})] - \mathbb{P}_{\mathbf{x}}[f(\mathbf{x}) \neq g(\mathbf{x})]|$. f is far from linear functions if it has small agreement with all linear functions (we make this definition precise in Section 15.2).

Linearity tests were first introduced by Blum, Luby and Rubenfeld in [BLR93]. They presented the following test (coined the BLR test), which makes only 3 queries to f:

1. Choose $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ at random

2. Verify that $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$.

Bellare et al. [BCH$^+$95] gave a tight analysis of the BLR test. It is obvious that the BLR test always accepts a linear function. They have shown that if the test accepts a function f with probability $1/2 + \epsilon$, then f has agreement at least 2ϵ with some linear function.

For a linearity test, we define that it has *completeness* c if it accepts any linear function with probability of at least c. A test has *perfect completeness* if $c = 1$. A linearity test has *soundness* s if it accepts any function f with agreement at most ϵ with all linear functions, with probability of at most $s + \epsilon'$, where $\epsilon' \to 0$ when $\epsilon \to 0$. We define the *query complexity* q of a test as the maximal number of queries it performs. In the case of the BLR test, it has perfect completeness, soundness $s = 1/2$ (with $\epsilon' = 2\epsilon$) and query complexity $q = 3$.

If one repeats a linearity test with query complexity q and soundness s independently t times, the query complexity grows to $q' = qt$ while the soundness reduces to $s' = s^t$. So, it makes sense to define the *amortized query complexity* $\bar{q}$ of a test as $\bar{q} = q/\log_2(1/s)$. Independent repetition of a test doesn't change it's amortized query complexity. Notice that the BLR test has amortized query complexity $\bar{q} = 3$.

Linearity tests are a key ingredient in the PCP theorem, started in the works of Arora and Safra [AS98] and Arora, Lund, Motwani, Sudan and Szegedy [ALM$^+$98]. In order to improve PCP constructions, linearity tests were studied in order to improve their amortized query complexity.

Samorodnitsky and Trevisan [ST00] have generalized the basic BLR linearity test. They introduced the *Complete Graph Test*. The Complete Graph Test (on k vertices) is:

1. Choose $\mathbf{x}_1, ..., \mathbf{x}_k \in \{0,1\}^n$ independently

2. Verify $f(\mathbf{x}_i + \mathbf{x}_j) = f(\mathbf{x}_i) + f(\mathbf{x}_j)$ for all i, j

This test has perfect completeness and query complexity $q = \binom{k}{2} + k$. They show that all the $\binom{k}{2}$ tests that the Complete Graph Test performs are essentially independent, i.e. that the test has soundness $s = 2^{-\binom{k}{2}}$. This makes this test have amortized query complexity $\bar{q} = 1 + \theta(1/\sqrt{q})$. They show that this test is optimal among the family of Hyper-Graph Tests (see [ST00] for definition of this family of linearity tests), and raise the question of whether the Complete Graph Test is optimal among all linearity tests, i.e. does a test with the same query complexity but with better soundness exist?

They partially answer this question in [ST06], where (among many other results) they show that no non-adaptive linearity test can perform better than the Complete Graph Test. A test is called *non-adaptive* if it first chooses q locations in the truth table of f, then queries them, and based on the results accept or rejects f. Otherwise, a test is called *adaptive*. An adaptive test may decide on its query locations based on the values of f in previous queries.

The proof technique of [ST06] uses the algebraic analysis of the Gowers Norm of certain functions. The Gowers Norm is a measure of local closeness of a function to a low degree polynomial. For more details regarding the definition and properties of the Gowers Norm, see [GT08] and [Sam07].

Ben-Sasson, Harsha and Raskhodnikova prove in [BSHR05] that any adaptive linearity test with completeness c, soundness s and query complexity q can be transformed into a non-adaptive linearity test with the same query complexity, perfect completeness and soundness $s' = s + 1 - c$. Combining their result with the result of [ST06] proves the lower bound also for adaptive linearity tests.

We also prove the same optimal lower bound for adaptive linearity test, but our proof technique is arguably simpler and more direct than the one used in [ST06]. We also study, like [ST06], the behavior of linearity tests on quadratic functions. However, instead of employing algebraic analysis of the Gowers Norm of certain functions, we provide a more direct combinatorial proof, studying the behavior of linearity tests on random quadratic functions. This proof technique also lets us prove directly the lower bound also for adaptive linearity tests.

15.1.1 Our techniques

We model adaptive tests using test trees. A test tree T is a binary tree, where in each inner vertex v there is some label $\mathbf{x}(v) \in \{0,1\}^n$, and the leaves are labeled with either *accept* or *reject*. Running a test tree on a function f is done by querying at each stage f on the label of the current vertex (starting at the root), and following one of the two edges leaving the vertex, depending on the query response. When reaching a leaf, its label (*accept* or *reject*) is the value of that f gets in T . An adaptive test $\mathbb{T}$ can always be modeled as first randomly choosing a test tree from some set $\{T_i\}$, according to some distribution on the test trees, then running the test tree on f.

It turns out that in order to prove a lower bound which matches the upper bound of the Complete Graph Test, it is enough to consider functions f which are quadratic. Actually, it's enough to consider f which is a random quadratic function.

A function f is quadratic if it can be presented as $f(x_1, ..., x_n) = \sum_{i,j} a_{i,j} x_i x_j + \sum_i b_i x_i + c$ for some values $a_{i,j}, b_i, c \in \mathbb{F}_2$. We study the behavior of running test trees on a random linear function, and on a random quadratic function.

The main idea is as follows. Let v be some inner vertex in a test tree T, with the path from the root of T to v being $v_0, ..., v_{k-1}, v$. If $\mathbf{x}(v)$ is linearly dependent on $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, then when running T on any linear function, the value of $f(\mathbf{x}(v))$ can be deduced from the already known values of $f(\mathbf{x}(v_0)), ..., f(\mathbf{x}(v_{k-1}))$. Therefore, if the vertex v is reached, then the same edge leaving v will always be taken by any linear function. Additionally, if $\mathbf{x}(v)$ is

linearly independent of $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, then either v is never reached running T on linear functions, or the two edges leaving v are taken with equal probability when running T on a random linear function. A similar analysis can be made when running T on quadratic functions, replacing *linear dependence* with a corresponding notion of *quadratic dependence*.

Using this observation, we can define the *linear rank* of a leaf v, marked $l(v)$, as the linear rank of labels on the path from the root to v. We prove that running the test tree T on a random linear function reaches v with probability $2^{-l(v)}$. Similarly, we define the *quadratic rank* of a leaf v, marked $q(v)$, as the quadratic rank of those labels, and we proving that running T on a random quadratic function reaches v with probability $2^{-q(v)}$. We prove that the quadratic rank of any set cannot be much larger than its linear rank, and in particular that $q(v) \leq \binom{l(v)}{2} + l(v)$ for all leaves v. We use this inequality to prove that a test which has completeness c and query complexity q accepts a random quadratic function with a probability of at least $c - 1 + 2^{-q+\phi(q)}$, where $\phi(q)$ is defined as the unique non-negative solution to $\binom{\phi(q)}{2} + \phi(q) = q$.

We use this to show that any linearity test with completeness c and query complexity q must have $s \geq 2^{-q+\phi(q)}$. In particular, the Complete Graph Test on k vertices has perfect completeness, soundness $s = 2^{-\binom{k}{2}}$ and query complexity $q = \binom{k}{2} + k$. Since $\phi(q) = k$ the Complete Graph Test is optimal among all adaptive tests with the same query complexity.

In fact, we prove a stronger claim. We say that a test $\mathbb{T}$ has *average query complexity* q if for any function f, the average number of queries performed is at most q. In particular any test with query complexity q also has average query complexity q. We prove that for any test with completeness c and average query complexity q, the soundness is at least $s \geq 2^{-q+\phi(q)}$.

We present and analyze linearity tests over $\mathbb{F}_2$. Linearity tests can also be considered over larger fields or groups. Our lower bound actually generalizes easily to any finite field, but for ease of presentation, and since the techniques are exactly the same, we present everything over $\mathbb{F}_2$. We comment further on the modifications required for general finite fields in Section 15.2.

15.2 Preliminaries

15.2.1 Linearity tests

We call a function $f : \{0,1\}^n \to \{0,1\}$ linear if it can be written as $f(x_1, ..., x_n) = a_1 x_1 + ... + a_n x_n$ for some $a_1, ..., a_n \in \{0,1\}$ where addition and multiplication are in $\mathbb{F}_2$.

A linearity test is a randomized algorithm with oracle access to the truth table of f, which is supposed to distinguish the following two extreme cases:

1. f is linear (accept)
2. f is ϵ-far from linear functions (reject)

where the agreement of two functions $f, g : \{0,1\} \to \{0,1\}$ is defined as $d(f,g) = |Pr_{\mathbf{x}}[f(\mathbf{x}) = g(\mathbf{x})] - Pr_{\mathbf{x}}[f(\mathbf{x}) \neq g(\mathbf{x})]|$, and f is ϵ-far from linear functions if the agreement it has with any linear function is at most ϵ.

We now follow with some standard definition regarding linearity tests (or more generally, property tests). We say a test has *completeness* c if for any linear function f the test accepts with probability at least c. A test has *perfect completeness* if $c = 1$. We say a test has *soundness* s if for any f which is ϵ-far from linear the test accepts with probability at most $s+\epsilon'$, where $\epsilon' \to 0$ when $\epsilon \to 0$ (in fact, we talk about a family of linearity tests, for $n \to \infty$, but we ignore this subtle point).

A test is said to have *query complexity* q if it accesses the truth-table of f at most q times (for any choice of it's internal randomness). A test is said to have *average query complexity* q if for any function f, the average number of accesses (over the internal randomness of the test) done to the truth table of f is at most q. Obviously, any test with query complexity q is also a test with average query complexity q.

We say a test is *non-adaptive* if it chooses all the locations it's going to query in the truth table of f before reading any of their values. Otherwise, we call the test *adaptive*.

We now turn to model adaptive tests in a way that will be more convenient for our analysis. We first define a test tree and running a test tree on a function.

Definition 15.1. *A* test tree *on functions* $\{0,1\}^n \to \{0,1\}$ *is a rooted binary tree* T. *On each inner vertex of the tree* v *there is a label* $x(v) \in \{0,1\}^n$. *On each leaf there is a label of either* accept *or* reject.

Definition 15.2. Running a test tree T on a function f *is done as follows. We start at the root of the tree* v_0, *read the value of* $f(x(v_0))$, *and according to the value take the left or the right edge leaving* v_0. *We continue in this fashion on inner vertices of* T *until we reach a leaf of* T. *The* value of f in T *is the value of the end leaf (i.e.* accept *or* reject*), and the* depth of f in T *is the depth of the end vertex of* f *in* T.

Using these definitions, we can now model adaptive tests. We identify an adaptive test $\mathbb{T}$ on functions $\{0,1\}^n \to \{0,1\}$ with a distribution of binary trees $\{T_i\}$ (also on functions $\{0,1\}^n \to \{0,1\}$). Running the test $\mathbb{T}$ on a function f is done by randomly choosing one of the trees T_i (according to their distribution), and then running the test tree T_i on f. The result of the function f in the test tree T_i is the result the test $\mathbb{T}$ returns on f.

Notice that a test has query complexity q iff all trees T_i has depth at most q, and has average query complexity q iff for any function f, the average depth reached in a random tree from $\{T_i\}$ is at most q.

In order to define our main theorem, we will define the following function. For $x > 0$ define $\phi(x)$ as the unique real positive solution to $\phi(x)^2/2 + \phi(x)/2 = x$. Notice that for positive integer $\phi(x)$, this is the same as $\binom{\phi(x)}{2} + \phi(x) = x$. The following is the main theorem of this paper:

Theorem 15.1. *(main theorem) Let* $\mathbb{T}$ *be an adaptive test with completeness* c, *soundness* s *and average query complexity* $q \geq 1$. *Then* $s + 1 - c \geq 2^{-q+\phi(q)}$.

Notice that for large q, $\phi(q) \approx \sqrt{2q}$, also $\sqrt{q} \leq \phi(q) \leq \sqrt{2q}$, so we get that in particular, $s + 1 - c \geq 2^{-q+\theta(\sqrt{q})}$.

The Complete Graph Test was presented in [ST00]. The test (on a graph with k vertices) can be described as choosing $\mathbf{x}_1, ..., \mathbf{x}_k$ at random, and querying f at $\mathbf{x}_i$ (for $i = 1..k$) and

on $\mathbf{x}_i + \mathbf{x}_j$ (for $1 \leq i < j \leq k$). The test accepts f if for any i, j

$$f(x_i) + f(x_j) + f(x_i + x_j) = 0$$

In [ST00] it is proven that the Complete Graph Test has perfect completeness and soundness $s = 2^{-\binom{k}{2}}$. The total number of queries performed is $q = k + \binom{k}{2}$, so by our definitions, $k = \phi(q)$ and $s = 2^{-q+\phi(q)}$. We have the following corollary:

Corollary 15.1. *The Complete Graph Test is optimal among all adaptive linearity tests.*

Remark. *We state and prove all results for functions $f : \{0,1\}^n \to \{0,1\}$. In fact, the lower bound result on adaptive linearity tests holds for functions $f : \mathbb{F}^n \to \mathbb{F}$ for any finite field $\mathbb{F}$, and not just $\mathbb{F}_2$, with only minor adjustments to the definitions and proofs. We need to make the following modifications:*

1. *Define "ϵ-far from linear functions" for general fields*
2. *Test trees should have $|F|$ edges leaving each edge instead of 2*
3. *The proof that random quadratic functions are far from linear, proved in Section 15.5, should be slightly modified*

Since the results follow simply for any finite field, we chose to present the results over $\mathbb{F}_2$ to make the presentation simpler and clearer.

15.3 Quadratic functions

We will see that in order to prove Theorem 15.1, it will be enough to limit the functions f to be quadratic. We say a function f is quadratic if it can be written as:

$$f(x_1, ..., x_n) = \sum_{i,j} a_{i,j} x_i x_j + \sum_i b_i x_i + c$$

for some $a_{i,j}, b_i, c \in \mathbb{F}_2$.

In fact, for our usage, we will force our quadratic functions f to have $f(0) = 0$ (equivalently, $c = 0$ in the above description). So, throughout this paper, when speaking of quadratic functions, we actually speak of quadratic functions f with the added condition $f(0) = 0$.

We will study the dynamics of a test tree T in a linearity test $\mathbb{T}$, in two cases - when applied to a uniformly random linear function, and when applied to a uniformly random quadratic function.

The following technical lemma is the key ingredient to the proof of the Theorem 15.1.

Lemma 15.1. *Let $\mathbb{T}$ be an adaptive linearity test with completeness c and average query complexity q. Then running $\mathbb{T}$ on a random quadratic function returns* accept *with probability at least $c - 1 + 2^{-q+\phi(q)}$.*

In order to prove Theorem 15.1, we will also need the following simple lemma:

Lemma 15.2. *Let f be a random quadratic function. Then the probability that f is not $2^{-\Omega(n)}$-far from linear functions is $2^{-\Omega(n)}$.*

Theorem 15.1 now follows directly from Lemmas 15.1 and 15.2. We sketch now it's proof following the two lemmas.

Proof. (of the main theorem) The average probability that $\mathbb{T}$ returns *accept* on a random quadratic function which is $2^{-\Omega(n)}$-far from linear functions is at least $c-1+2^{-q+\phi(q)}-2^{-\Omega(n)}$. So, there exists some quadratic function f which is $2^{-\Omega(n)}$-far from linear and on which $\mathbb{T}$ returns *accept* with probability at least $c-1+2^{-q+\phi(q)}-2^{-\Omega(n)}$. Taking $n \to \infty$ shows that $s+1-c \geq 2^{-1+\phi(q)}$. □

The remainder of the paper is organized as follows. Lemma 15.1 is proved in Section 15.4, and Lemma 15.2 is proved in Section 15.5.

15.4 Linearity test applied to a random quadratic function

We study tests and test trees applied to linear and quadratic functions, in order to prove Lemma 15.1. Let $\mathbb{T}$ be an adaptive test with completeness c and average query complexity q. Let T be a some test tree which is a part of the test $\mathbb{T}$.

We start by studying the dynamics of applying T to linear functions. Assume we know that f is a linear function, and we are at some vertex $v \in T$, where the path from the root to v is $v_0, .., v_{k-1}, v$. Assume $\mathbf{x}(v)$ is linearly dependant on $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$. Since we know f is linear, we can deduce the value of $\mathbf{x}(v)$ from $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, and so we will always follow the same edge leaving v when we apply T to any linear function. On the other hand, if $\mathbf{x}(v)$ is linearly independent of $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, we know that when we apply T to a random linear function, either we never reach v, or we have equal chances of taking any of the two edges leaving v.

This gives rise to the following formal definition:

Definition 15.3. *Let v be a leaf in T, where the path from the root to v is $v_0, v_1, ..., v_{k-1}, v$. We define the* linear degree *of v, marked $l(v)$, to be the linear rank of $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$.*

We define L_T to be the set of leaves of T to which linear functions can arrive. i.e, $v \in L$ if the path from the root to v, $v_0, ..., v_{k-1}, v$ always takes the "correct" edge leaving any vertex v_i with $\mathbf{x}(v_i)$ linearly dependent on $\mathbf{x}(v_0), ..., \mathbf{x}(v_{i-1})$.

The following lemma formalizes the discussion above:

Lemma 15.3. *For any test tree T:*

1. *For any $v \in L_T$, the probability that a random linear function will arrive to v is $2^{-l(v)}$*
2. $\sum_{v \in L_T} 2^{-l(v)} = 1$

For $v \in L_T$, we define $c(v)$ to be 1 if the value of v is *accept*, and $c(v) = 0$ otherwise. Since the completeness of $\mathbb{T}$ is c, we have that the probability that $\mathbb{T}$ returns *accept* on a random linear function is at least c. On the other hand, for any test tree T in $\mathbb{T}$, the probability that a random linear function will return *accept* is exactly $\sum_{v \in L_T} c(v)2^{-l(v)}$. So, the following lemma follows:

Lemma 15.4. $\mathbb{E}_T \sum_{v \in L_T} c(v)2^{-l(v)} \geq c$

where by $\mathbb{E}_T$ here and throughout the paper we mean the average value of a random test tree T in $\mathbb{T}$.

We now generalize the concept of linear dependence to quadratic functions.

Definition 15.4. *Let* $\mathbf{x}_1, ..., \mathbf{x}_k \in \{0,1\}^n$.

1. *We say* $\mathbf{x}_1, ..., x_k$ *are* quadratically dependent *if there are constants* $a_1, ..., a_k \in \mathbb{F}_2$, *not all zero, s.t. for any quadratic function* f *we have:* $a_1 f(\mathbf{x}_1) + ... + a_k f(\mathbf{x}_k) = 0$. *otherwise will call* $x_1, ..., x_k$ quadratically independent.

2. *We say* $\mathbf{x}_k$ *is* quadratically dependent *on* $\mathbf{x}_1, ..., \mathbf{x}_{k-1}$ *if there are constants* $a_1, ..., a_{k-1} \in \mathbb{F}_2$ *s.t. for any quadratic function* f *we have:* $f(\mathbf{x}_k) = a_1 f(\mathbf{x}_1) + ... + a_{k-1} f(\mathbf{x}_{k-1})$. *Otherwise we say* $\mathbf{x}_k$ *is* quadratically independent *of* $\mathbf{x}_1, ..., \mathbf{x}_{k-1}$.

3. *We define the* quadratic dimension *of* $\mathbf{x}_1, ..., \mathbf{x}_k$ *to be the size of the largest subset of* $\{\mathbf{x}_1, ..., \mathbf{x}_k\}$ *which is quadratically independent.*

This definition may seem obfuscated, but the following alternative yet equivalent definition will clarify it. The space of quadratic functions over $\{0,1\}^n$ is a linear space over $\mathbb{F}_2$. Let M be it's generating matrix, i.e. the rows of M are a base for the linear space (in particular, the dimensions of M are $(\binom{n}{2} + n) \times 2^n$). A column of M corresponds to an input $\mathbf{x} \in \{0,1\}^n$. Now, $\mathbf{x}_1, ..., \mathbf{x}_k$ are quadratically dependent iff the columns corresponding to them are linearly dependent, and similarly for the other definitions.

Notice that the usual definition of linear dependence is equivalent to this more complex definition, when applied to the linear space of all linear functions.

We now can repeat the informal discussion at the start of this section, except this time for quadratic functions, with all the reasoning left intact. Let $v \in T$ be a vertex, with path from the root being $v_0, ..., v_{k-1}, v$. Assume $\mathbf{x}(v)$ is quadratically dependent on $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, and f is any quadratic function. The value of $f(\mathbf{x}(v))$ can be deduced from the already known values of $f(\mathbf{x}(v_0)), ..., f(\mathbf{x}(v_{k-1}))$, and so only one edge leaving v will be taken on all quadratic functions. Alternatively, if $x(v)$ is quadratically independent on $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$, then a random quadratic function either never reaches v, or has equal chances of taking each of the two edges leaving v.

This leads to the following definition and lemma for quadratic degree of a vertex $v \in T$, similar to the ones for linear degree.

Definition 15.5. *Let* v *be a leaf in* T, *where the path from the root to* v *is* $v_0, v_1, ..., v_{k-1}, v$. *We define the* quadratic degree *of* v, *marked* $q(v)$, *to be the quadratic rank of* $\mathbf{x}(v_0), ..., \mathbf{x}(v_{k-1})$.

We define Q_T to be the set of leaves of T to which quadratic functions can arrive. Naturally $L_T \subseteq Q_T$. The following lemma on quadratic degree follows from the discussion above:

Lemma 15.5. *For any test tree T:*

1. *For any $v \in Q_T$, the probability that a random quadratic function will arrive to v is $2^{-q(v)}$*
2. $\sum_{v \in Q} 2^{-q(v)} = 1$
3. *For any $v \in L_T$ we have $q(v) \geq l(v)$*

Last, we mark the depth of a vertex $v \in T$ by $d(v)$. Since $\mathbb{T}$ has average query complexity q, we know that for any function f, the average depth of running a random tree T of $\mathbb{T}$ on f is at most q. So, this also holds for a random linear function. However, the average depth a random linear function arrives on a tree T is exactly $\sum d(v)2^{-l(v)}$, so the following lemma follows.

Lemma 15.6. $\mathbb{E}_T \sum_{v \in L_T} d(v)2^{-l(v)} \leq q$

We now wish to make a connection between $q(v)$ and $l(v)$ for vertices $v \in L_T$.
First, we prove that following lemma:

Lemma 15.7. *For any $\mathbf{x}_1, ..., \mathbf{x}_k \in \{0,1\}^n$ there are coefficients $a_{i,j}, b_i \in \mathbb{F}_2$ s.t. for any quadratic function f we have:*

$$f(\mathbf{x}_1 + ... + \mathbf{x}_k) = \sum_{i,j} a_{i,j} f(\mathbf{x}_i + \mathbf{x}_j) + \sum_i b_i f(\mathbf{x}_i)$$

Proof. Let $f(\mathbf{x})$ by some polynomial of degree d. It's derivative in the $\mathbf{y}$ direction is defined to be $f_{\mathbf{y}}(\mathbf{x}) = f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x})$. It's easy to see that the degree of $f_{\mathbf{y}}$ as a function of $\mathbf{x}$ is at most $d-1$. So, taking 3 derivatives from a quadratic function makes it the zero function, and so in particular for any quadratic function f, we we take it's derivatives in directions $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, and evaluate the result at 0, we get that

$$(((f_{\mathbf{x}})_{\mathbf{y}})_{\mathbf{z}}(0) = 0$$

Opening this expression yields:

$$f(\mathbf{x}+\mathbf{y}+\mathbf{z}) - f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}+\mathbf{z}) - f(\mathbf{y}+\mathbf{z}) + f(\mathbf{x}) + f(\mathbf{y}) + f(\mathbf{z}) - f(0) = 0$$

Since $f(0) = 0$, we can express $f(\mathbf{x}+\mathbf{y}+\mathbf{z})$ as a sum of application of f on an element, or sum of two elements in $\{x, \mathbf{y}, \mathbf{z}\}$. This proves the lemma for $k = 3$. For $k > 3$ we use simple induction. □

Now we can bound $l(v)$ in term of $q(v)$. We first prove a result bounding in general the linear rank of a set by it's quadratic rank.

Lemma 15.8. *Let $\{\mathbf{x}_1, ..., \mathbf{x}_k\}$ be elements in $\{0,1\}^n$. Let l be the their linear rank, and q their quadratic rank. Then*

$$q \leq \binom{l}{2} + l$$

Proof. Let $S \subset \{\mathbf{x}_1, ..., \mathbf{x}_k\}$ be a maximal quadratic independent set. $|S| = q$. The linear rank of S is also l. Let $S' \subset S$ be a maximal set of linearly independent elements of S. $|S'| = l$. Assume w.l.o.g that $S' = \{\mathbf{x}_1, ..., \mathbf{x}_l\}$. Since every $x \in S$ is linearly dependent on S', it can be written as a sum of some of the elements of S'. Using Lemma 15.7, we get that for any $\mathbf{x} \in S$ there exists coefficients $a_{i,j}^{(\mathbf{x})}, b_i^{(\mathbf{x})} \in \mathbb{F}_2$ s.t for any quadratic function f:

$$f(\mathbf{x}) = \sum_{1 \leq i < j \leq l} a_{i,j}^{(\mathbf{x})} f(\mathbf{x}_i + \mathbf{x}_j) + \sum_{1 \leq i \leq l} b_i^{(\mathbf{x})} f(\mathbf{x}_i)$$

We have assumed that all the elements of S are quadratically independent. For this to hold, the above equations in the symbolic variables $f(\mathbf{x}_i + \mathbf{x}_j)$ and $f(\mathbf{x}_i)$ must be linearly independent. So the number of equations q must be at most the number of variables, which is $\binom{l}{2} + l$. So, we get that:

$$q = |S| \leq \binom{l}{2} + l$$

□

Lemma 15.9. *For any leaf $v \in L_T$, $l(v) \geq \phi(q(v))$*

Proof. Let $v_0, ..., v_{k-1}, v$ be the path in T from the root to v. Let $\mathbf{x}_i = \mathbf{x}(v_i)$ for $i = 0..k-1$. Apply lemma 15.8 on $\{\mathbf{x}_0, ..., \mathbf{x}_{k-1}\}$ to get that $q(v) \leq \binom{l(v)}{2} + l(v)$. Reversing this formula, since $\phi(x)$ is monotone, we get that $l(v) \geq \phi(q(v))$. □

We can now prove our main technical lemma (Lemma 15.1). We start with some technical lemmas. We define $\psi(x)$ to be $x - \phi(x)$ for $x \geq 1$, and 0 for $x < 1$. Notice that ψ is continuous, and $\psi(x) = x - \phi(x)$ for any non-negative integer x. Hence, using Lemma 15.9 we get that:

Lemma 15.10. *For any vertex v in a tree T, $q(v) - l(v) \leq \psi(q(v))$.*

Lemma 15.11. *ψ is increasing and convex.*

Proof. Since ψ is continuous and constant for $x \leq 1$, it's enough to prove the claim for $x > 1$ (for increasing it's clear, and once we've proved ψ is increasing, it shows it's enough to prove convexity for $x > 1$). We first show ψ is increasing.

For $x > 1$, define $y = \phi(x)$, so $x = y^2/2 + y/2$ and $\psi(y) = y^2/2 - y/2$.

$$\frac{d\psi}{dx} = \frac{d\psi}{dy}\frac{dy}{dx} = \frac{\frac{d\psi}{dy}}{\frac{dx}{dy}} = \frac{y - 1/2}{y + 1/2}$$

If $x > 1$ then $y = \phi(x) > 1$, hence $\frac{d\psi}{dx} > 0$ for $x > 1$, and so ψ is increasing.

To show that ψ in convex,

$$\frac{d^2\psi}{dx^2} = \frac{d\left(\frac{y-1/2}{y+1/2}\right)}{dy}\frac{dy}{dx} = \frac{1}{(y + 1/2)^3} > 0$$

□

We are now finally ready to prove Lemma 15.1.

Proof. (of Lemma 15.1) We need to prove that any test $\mathbb{T}$ with completeness c and average query complexity $q \geq 1$ accepts a random quadratic function with probability at least $c - 1 + 2^{-\psi(q)}$. Let us mark the probability the test accepts a random quadratic function by p. Let p_T mark the probability that a tree T accepts a random quadratic function. p_T is at least the probability that a random quadratic function reaches a leaf in L_T which is labeled *accept*. So:

$$p_T \geq \sum_{v \in L_T} c(v) 2^{-q(v)}$$

We now follow to analyze $p = \mathbb{E}_T[p_T]$.

$$p \geq \mathbb{E}_T[\sum_{v \in L_T} c(v) 2^{-q(v)}] = \mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} c(v) 2^{-q(v)+l(v)}]$$

.

We divide the sum in the right side into two parts, $p_0 - p_1$, with $p_0, p_1 \geq 0$, where:

$$p_0 = \mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} 2^{-q(v)+l(v)}]$$

. and

$$p_1 = \mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} (1 - c(v)) 2^{-q(v)+l(v)}]$$

.

We start by analyzing p_1. Since for any v always $q(v) \geq l(v)$ we have:

$$p_1 \leq \mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} (1 - c(v))]$$

Recall that by Lemma 15.5 for any tree T we have

$$\sum_{v \in L_T} 2^{-l(v)} = 1$$

and by Lemma 15.4 we have

$$\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} c(v)] \geq c$$

so we conclude that:

$$p_1 \leq 1 - c$$

We move to analyze p_0. Since $\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)}] = 1$ and since the function $X \to 2^X$ is concave, we have by Jensen's inequality that:

$$p_0 \geq 2^{\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)} (-q(v) + l(v))]}$$

Now, we have that $q(v) - l(v) \leq \psi(q(v))$ by Lemma 15.11, and also by the same lemma, since $q(v) \leq d(v)$, we get $\psi(q(v)) \leq \psi(d(v))$. So we get:

$$\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)}(q(v) - l(v))] \leq \mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)}\psi(d(v))]$$

.

Since by Lemma 15.11 ψ is convex, we get that again by Jensen's inequality we get that this is at most $\psi(\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)}d(v)])$. By Lemma 15.6

$$\mathbb{E}_T[\sum_{v \in L_T} 2^{-l(v)}d(v)] \leq q$$

where q is the average query complexity of $\mathbb{T}$. So, we conclude that $p_0 \geq 2^{-\psi(q)}$, and in total

$$p \geq p_0 - p_1 \geq 2^{-\psi(q)} + c - 1$$

□

15.5 Random quadratic function is far from linear

In this section we prove Lemma 15.2, i.e. that a random quadratic function is far from linear. We will use commonly known facts about quadratic functions.

Any quadratic function can be written as:

$$f(\mathbf{x}) = \mathbf{x}^t A\mathbf{x} + < \mathbf{x}, b >$$

The correlation of f with some linear function g is the g-th Fourier coefficient of f. The Fourier coefficients of quadratic functions are well studied. In particular, it is known that all the Fourier coefficients of f have the same absolute value, and that the number of non-zero Fourier coefficients is $2^{rank(A+A^t)}$. So, in order to show that f has no large correlation with some linear function, it's enough to show that $B = A + A^t$ has high rank. In particular, in order to show that f is $2^{-\Omega(n)}$-far from linear functions, we need to show that B has rank $\Omega(n)$. We will show that the probability that a random quadratic function has rank less than $n/4$ is $2^{-\Omega(n)}$. We will use the following lemma:

Lemma 15.12. *The number of matrices of rank at most k is at most $n^k 2^{nk}$.*

Using Lemma 15.12, it's easy to prove Lemma 15.2. The number of matrices of rank at most $n/4$ is at most $2^{n^2/4(1+o(1))}$. For a random quadratic function, B is a random symmetric matrix with zero diagonal, and so the probability that B has rank less than $n/4$ is $2^{-n^2/4(1+o(1))} = 2^{-\Omega(n)}$.

Now we finish by proving Lemma 15.12.

Proof. Let B be a matrix of rank at most k. There are $\binom{n}{k}$ options to choose k rows which span the row span of the matrix, each other row have at most 2^k options since it must be in the row span of k specific rows. So, the number of possibilities for rank k matrices is at most:

$$\binom{n}{k}(2^k)^{n-k} \leq n^k 2^{nk}$$

□

Bibliography

[AB01] N. Alon and R. Beigel. Lower bounds for approximations by low degree polynomials over z_m. In *SCT: Annual Conference on Structure in Complexity Theory*, 2001.

[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[ABEK08] Noga Alon, Ido Ben-Eliezer, and Michael Krivelevich. Small sample spaces cannot fool low degree polynomials. In *APPROX '08 / RANDOM '08: Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization*, pages 266–275, Berlin, Heidelberg, 2008. Springer-Verlag.

[ABFR94] J. Aspnes, R. Beigel, M. Furst, and S. Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):1–14, 1994.

[AGHP90] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple construction of almost k-wise independent random variables. *Symposium on Foundations of Computer Science*, 2:544–553, 1990.

[AIK+90] Miklós Ajtai, Henryk Iwaniec, János Komlós, János Pintz, and Endre Szemerédi. Construction of a thin set with small Fourier coefficients. *Bull. London Math. Soc.*, 22(6):583–590, 1990.

[AKS] Miklós Ajtai, János Komlós, and E. Szemerédi. Deterministic simulation in LOGSPACE. pages 132–140.

[AKT76] Saburo Azumi, Tadao Kasami, and Nobuki Tokura. On the weight enumeration of weights less than $2.5d$ of Reed-Muller codes. *Information and Control*, 30(4):380–395, 1976.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998.

[Alo83] Noga Alon. On the density of sets of vectors. *Discrete Mathematics*, 46:199–202, 1983.

[AM95] Noga Alon and Yishay Mansour. ϵ-discrepancy sets and their application for interpolation of sparse polynomials. *Information Processing Letters*, 54(6):337–342, 1995.

[AR94] Noga Alon and Yuval Roichman. Random Cayley graphs and expanders. *Random Structures Algorithms*, 5(2):271–284, 1994.

[ARKS] Ryan O'Donnell Adam R. Klivans and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. In *Proceedings of the* 43^{rd} *Annual Symposium on Foundations of Computer Science (FOCS'02)*.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: a new characterization of NP. *J. ACM*, 45(1):70–122, 1998.

[AS00] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley, 2 edition, 2000.

[Ax64] James Ax. Zeroes of polynomials over finite fields. *American Journal of Mathematics*, 86(2):255–261, 1964.

[Baz07] Louay M. J. Bazzi. Polylogarithmic independence can fool dnf formulas. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*, pages 63–73. IEEE Computer Society, 2007.

[BBR94] David A. Barrington, Richard Beigel, and Steven Rudich. Representing Boolean functions as polynomials modulo composite numbers. *Computational Complexity*, 4:367–382, 1994.

[BCH+95] M. Bellare, D. Coppersmith, J. Hastad, M. Kiwi, and M. Sudan. Linearity testing in characteristic two. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS '95)*, pages 432–441, 1995.

[BEHL09] Ido Ben-Eliezer, Rani Hod, and Shachar Lovett. Random low degree polynomials are hard to approximate. In *Proceedings of the* 12^{th} *International Workshop and* 13^{th} *International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX '09 / RANDOM '09)*, pages 366–377, 2009.

[Bei93] Richard Beigel. The polynomial method in circuit complexity. *Structures in Complexity Theory:* 8^{th} *Annual Conference*, pages 82–95, 1993.

[BGL06] Nayantara Bhatnagar, Parikshit Gopalan, and Richard J. Lipton. Symmetric polynomials over $\mathbb{Z}_m$ and simultaneous communication protocols. *Journal of Computer and System Sciences*, 72:252–285, 2006.

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *J. Comput. Syst. Sci.*, 47(3):549–595, 1993.

[BM65] G. Birkhoff and S. MacLane. *A Survey of Modern Algebra, third edition*. MacMillan, 1965.

[BNS] László Babai, Noam Nisan, and Márió Szegedy. Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs. *Journal of Computer and System Sciences*, pages 204–232.

[Bog05] Andrej Bogdanov. Pseudorandom generators for low degree polynomials. In *Proceedings of the* 37^{th} *annual ACM symposium on Theory of computing (STOC '05)*, pages 21–30, 2005.

[Bou05] J. Bourgain. Mordell's exponential sum estimate revisited. *J. Amer. Math. Soc.*, 18(2), 2005.

[BR] Mihir Bellare and John Rompel. Randomness-efficient oblivious sampling. pages 276–287.

[Bra09] Mark Braverman. Poly-logarithmic independence fools AC_0 circuits. In *Proceedings of the* 24^{th} *Conference on Computational Complexity (CCC '09)*, 2009.

[BRS] R. Beigel, N. Reingold, and D. Spielman. The perceptron strikes back. In *Proceedings of the* 6^{th} *Conference on Structure in Complexity Theory.*

[BSHR05] E. Ben-Sasson, P. Harsha, and S. Raskhodnikov. Some 3CNF properties are hard to test. *SIAM Journal on Computing*, 35(1):1–21, 2005.

[BSK08] Eli Ben-Sasson and Swastik Kopparty, 2008. Unpublished manuscript.

[BSS05] Eli Ben-Sasson and Madhu Sudan. Simple PCPs with poly-log rate and query complexity. In *In Proceedings of the* 37*th Annual ACM Symposium on Theory of Computing (STOC' 2005)*, pages 266–275, 2005.

[BSS09] Eli Ben-Sasson and Madhu Sudan. Limits on the rate of locally testable affine-invariant codes. Manuscript, 2009.

[BSSVW03] Eli Ben-Sasson, Madhu Sudan, Salil Vadhan, and Avi Wigderson. Randomness-efficient low degree tests and short PCPs via epsilon-biased sets. In *Proceedings of the* 35^{th} *annual ACM symposium on Theory of computing (STOC '03)*, pages 612–621, 2003.

[BTZ09] Vitaly Bergelson, Terence Tao, and Tamar Ziegler. An inverse theorem for the uniformity seminorms associated with the action of $\mathcal{F}^{\omega}$. Submitted, 2009.

[BV07] Andrej Bogdanov and Emanuele Viola. Pseudorandom bits for polynomials. In *Proceedings of the* 48^{th} *Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*, pages 41–51, 2007.

[Cho61] C.K. Chow. On the characterization of threshold functions. In *Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS '61)*, pages 34–38, 1961.

[CHSL97] G. Cohen, I. Honkala, S.Litsyn, and A. Lobstein. *Covering Codes*. North–Holland, Amsterdam, 1997.

[Del78] P. Deligne. Aplications de la formule des traces aux sommes trigonometriques. *SGA* $4\frac{1}{2}$ *Springer Lecture Notes in Math*, (569), 1978.

[DGJ+09] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. To be published in FOCS '09, 2009.

[DGW70] P. Delsarte, J.-M. Goethals, and F. J. M. Williams. On Generalized Reed-Muller codes and their relatives. *Information and Control*, 16(5):403–442, 1970.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3), 2007.

[DLM+07] Ilias Diakonikolas, Homin K. Lee, Kevin Matulef, Krzysztof Onak, Ronitt Rubinfeld, Rocco A. Servedio, and Andrew Wan. Testing for concise representations. In *FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 549–558, Washington, DC, USA, 2007. IEEE Computer Society.

[Efr09] Klim Efremenko. 3-query locally decodable codes of exponential codes. In *Proceedings of the* 41^{st} *Annual Symposium on the Theory of Computing (STOC '09)*. ACM, 2009.

[EGL+98] Guy Even, Oded Goldreich, Michael Luby, Noam Nisan, and Boban Veličković. Efficient approximation of product distributions. *Random Structures Algorithms*, 13(1):1–16, 1998.

[Eve91] Guy Even. Construction of small probabilistic spaces for deterministic simulation. Master's thesis, The Technion - Israel Intitute of Technology, 1991.

[Fra83] P. Frankl. On the trace of finite sets. *J. of Combinatorial Theory, Series A*, 34(1):41–45, 1983.

[GKS09] Elena Grigorescu, Tali Kaufman, and Madhu Sudan. Succinct representation of codes with applications to testing. manuscript, 2009.

[GKZ08] Parikshit Gopalan, Adam R. Klivans, and David Zuckerman. List-decoding Reed-Muller codes over small fields. In *Proceedings of the* 40^{th} *Annual ACM Symposium on Theory of Computing (STOC '08)*, pages 265–274, 2008.

[GL89] Oded Goldreich and Leonid A. Levint. A hard-core predicate for all one-way functions. In *In Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing*, pages 25–32, 1989.

[GLS09] Parikshit Gopalan, Shachar Lovett, and Amir Shpilka. On the degree of boolean functions in different characteristics. In *Proceedings of the 24rd Annual CCC*, pages 173–183, 2009.

[Gol08] Oded Goldreich. *Computational Complexity: A Conceptual Perspective*. Cambridge University Press, 2008.

[Gop06a] Parikshit Gopalan. *Computing with Polynomials over Composites*. PhD thesis, Georgia Institute of Technology, 2006.

[Gop06b] Parikshit Gopalan. Constructing Ramsey graphs from Boolean function representations. In *Proceedings of the* 21^{st} *IEEE Conference on Computational Complexity (CCC '06)*, 2006.

[Gop09] Parikshit Gopalan. A Fourier-analytic approach to Reed-Muller decoding. Submitted, 2009.

[Gow01] W.T. Gowers. A new proof of Szemerédi's theorem. *Geometric And Functional Analysis*, 11(3):465–588, 2001.

[Gro00] Vince Grolmusz. Superpolynomial size set-systems with restricted intersections mod 6 and explicit Ramsey graphs. *Combinatorica*, 20(1):71–86, 2000.

[Gro02] Vince Grolmusz. Constructing set systems with prescribed intersection sizes. *Journal of Algorithms*, 44(2):321–337, 2002.

[GS06] Oded Goldreich and Madhu Sudan. Locally testable codes and PCPs of almost-linear length. *J. ACM*, 53(4):558–655, 2006.

[GT07] Ben Green and Terence Tao. The distribution of polynomials over finite fields, with applications to the Gowers norms. Submitted, 2007.

[GT08] Ben Green and Terence Tao. An inverse theorem for the Gowers U^3 norm. 2008.

[Gur04] V. Guruswami. *List decoding of Error-Correcting Codes*. Springer, 2004. Vol 3282 of Lecture notes in Computer Science.

[GW97] Oded Goldreich and Avi Wigderson. Tiny families of functions with random properties: A quality–size trade–off for hashing. *Journal of Random structures and Algorithms*, 11(4):315–343, 1997.

[Hås86] Johan Håstad. *Computational limitations for small-depth circuits*. PhD thesis, MIT, 1986.

[HPS93] Johan Håstad, Steven Phillips, and Shmuel Safra. A well-characterized approximation problem. *Information Processing Letters*, 47(6):301–305, 1993.

[HS10] Elad Haramaty and Amir Shpilka. On the structure of cubic and quartic polynomials. In *Proceedings of the 42nd STOC*, pages 331–340, 2010.

[INW94] Russell Impagliazzo, Noam Nisan, and Avi Wigderson. Pseudorandomness for network algorithms. In *Proceedings of the* 26^{th} *annual ACM symposium on Theory of computing (STOC '94)*, pages 356–364, 1994.

[JCJS02] Adam R. Klivans Jeffrey C. Jackson and Rocco A. Servedio. Learnability beyond AC^0. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC '02)*, 2002.

[JPRZ04] Charanjit S. Jutla, Anindya C. Patthak, Atri Rudra, and David Zukcerman. Testing low-degree polynomials over prime fields. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science (FOCS' 04)*, pages 423–432, 2004.

[Juk01] S. Jukna. *Extremal Combinatorics with Applications in Computer Science*. Springer–Verlag, 2001.

[Kat89] Nicholas M. Katz. An estimate for character sums. *Journal of the American Mathematical Society*, 2(2):197–200, 1989.

[KKMS05] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *Proceedings of the* 46^{th} *IEEE Symp. on Foundations of Computer Science (FOCS '05)*, 2005.

[KL05] Tali Kaufman and Simon Litsyn. Almost orthogonal linear codes are locally testable. In *Proceedings of the 46th Annual Symposium on Foundations of Computer Science (FOCS' 05)*, pages 317–326, 2005.

[KL08] Tali Kaufman and Shachar Lovett. Worst case to average case reductions for polynomials. In *Proceedings of the* 49^{th} *Annual IEEE Symposium on Foundations of Computer Science (FOCS '08)*, pages 166–175, 2008.

[KL19] Tali Kaufman and Shachar Lovett. Testing of exponentially large codes, by a new extension to weil bound for character sums. Manuscript, 2019.

[KLP10] Tali Kaufman, Shachar Lovett, and Ely Porat. The list-decoding size of Reed-Muller codes. In *Proceedings of the* 1^{st} *conference on Innovations in Computer Science (ICS' 10)*, 2010.

[KM93] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. on Computing*, 22(6):1331–1348, 1993.

[KR04] Tali Kaufman and Dana Ron. Testing polynomials over general fields. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science (FOCS' 04)*, pages 413–422, 2004.

[KS01] A. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{o}(n^{1/3})}$. In *Proceedings of the* 33^{rd} *Annual Symposium on Theory of Computing (STOC '01)*, pages 258–265, 2001.

[KS05] Peter Keevash and Benny Sudakov. Set systems with restricted cross-intersections and the minimum rank of inclusion matrices. *SIAM J. Discret. Math.*, 18(4):713–727, 2005.

[KS07] Tali Kaufman and Madhu Sudan. Sparse random linear codes are locally decodeable and testable. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science (FOCS' 07)*, pages 590–600, 2007.

[KS08] Tali Kaufman and Madhu Sudan. Algebraic property testing: The role of invariance. In *Proceedings of the 40th ACM Symposium on Theory of Computing (STOC' 08)*, 2008.

[KS10] Swastik Kopparty and Shubhangi Saraf. Local list-decoding and testing of random linear codes from high-error. In *to appear in the Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC' 10)*, 2010.

[KT70] T. Kasami and N. Tokura. On the weight structure of Reed–Muller codes. *IEEE Transactions on Information Theory*, 16(6):752–759, 1970.

[KTA76] Tadao Kasami, Nobuki Tokura, and Saburo Azumi. On the weight enumeration of weights less than 2.5d of reed-muller codes. *Information and Control*, 30(4):380–395, 1976.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[LMS08] Shachar Lovett, Roy Meshulam, and Alex Samorodnitsky. Inverse conjecture for the Gowers norm is false. In *Proceedings of the* 40^{th} *annual ACM symposium on Theory of computing (STOC '08)*, pages 547–556, 2008.

[LMS10] Shachar Lovett, Partha Mukhopadhyay, and Amir Shpilka. Pseudorandom generators for $\mathrm{CC}_0[p]$ and the Fourier spectrum of low-degree polynomials over finite fields. ECCC:TR10-033, 2010.

[LN90] Nati Linial and Noam Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10:349–365, 1990.

[Lov08] Shachar Lovett. Unconditional pseudorandom generators for low degree polynomials. In *Proceedings of the* 40^{th} *annual ACM symposium on Theory of computing (STOC '08)*, pages 557–562, 2008.

[Lov09] Shachar Lovett. Holes in generalized Reed-Muller codes. Accepted to the IEEE transactions on Information Theory, 2009.

[LRTV09] Shachar Lovett, Omer Reingold, Luca Trevisan, and Salil Vadhan. Pseudorandom bit generators that fool modular sums. In *Proceedings of the* 12^{th} *International Workshop and* 13^{th} *International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX '09 / RANDOM '09)*, pages 615–630, 2009.

[LT09] Shachar Lovett and Yoav Tzur. Explicit lower bound for fooling polynomials by the sum of small-bias generators. Electronic Colloquium on Computational Complexity, TR 09-088, 2009.

[LVW93] Michael Luby, Boban Velickovic, and Avi Wigderson. Deterministic approximate counting of depth-2 circuits. In *Proceedings of the* 2^{nd} *ISTCS*, pages 18–24, 1993.

[Mei08] Or Meir. Combinatorial construction of locally testable codes. In *In Proceedings of the* 40*th Annual ACM Symposium on Theory of Computing (STOC' 2008)*, pages 285–294, 2008.

[MNN94] Rajeev Motwani, Joseph Naor, and Moni Naor. The probabilistic method yields deterministic parallel algorithms. *Journal of Computer and System Sciences*, 49(3):478–516, 1994.

[MOS03] Elchannan Mossel, Ryan O'Donnell, and Rocco Servedio. Learning juntas. In *Proceedings of the* 35^{th} *Annual ACM Symposium on the Theory of Computing (STOC '03).*, 2003.

[MP68] Marvin Minsky and Seymour Papert. *Perceptrons: an Introduction to Computational Geometry*. MIT Press, 1968.

[MS83] Florence J. MacWilliams and Neil J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1983.

[MST06] Elchanan Mossel, Amir Shpilka, and Luca Trevisan. On ϵ-biased generators in NC^0. *Random Structures Algorithms*, 29(1):56–81, 2006.

[Mur71] Saburo Muroga. *Threshold logic and its applications*. Wiley-Interscience, 1971.

[MZ09] Raghu Meka and David Zuckerman. Small-bias spaces for group products. In *APPROX-RANDOM*, pages 658–672, 2009.

[Nao92] Moni Naor. Constructing Ramsey graphs from small probability spaces. Technical Report RJ 8810, IBM Research Report, 1992.

[NAR03] M. Krivelevich S. Litsyn N. Alon, T. Kaufman and D. Ron. Testing low-degree polynomials over $gf(2)$. In *RANDOM-APPROX 2003*, pages 188–199, 2003.

[Nis91] Noam Nisan. Pseudorandom bits for constant depth circuits. *Combinatorica*, 11(1):63–70, 1991.

[Nis92] Noam Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. on Computing*, 22(4):838–856, 1993.

[NS92] Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. In *Proceedings of the* 24^{th} *Annual ACM Symposium on the Theory of Computing (STOC '92)*, pages 462–467, 1992.

[NW94] Noam Nisan and Avi Wigderson. Hardness vs. randomness. *J. Comput. Syst. Sci.*, 49(2):149–167, 1994.

[Pat92] R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions. In *Proceedings of the* 24^{th} *Symposium on Theory of Computing*, pages 468–474, 1992.

[Raz87] A. A. Razborov. Lower bounds for the size of circuits of bounded depth with basis $\{\&, \oplus\}$. *Math. Notes Acad. Sci. USSR*, 41(4):333–338, 1987.

[RS93] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials and their applications to program testing. Technical report, Ithaca, NY, USA, 1993.

[RS09] Yuval Rabani and Amir Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. In Michael Mitzenmacher, editor, *STOC*, pages 649–658. ACM, 2009.

[RSW93] A. Razborov, E. Szemerédi, and A. Wigderson. Constructing small sets that are uniform in arithmetic progressions. *Combinatorics, Probability and Computing*, 2(4):513–518, 1993.

[RTV06] Omer Reingold, Luca Trevisan, and Salil Vadhan. Pseudorandom walks in regular digraphs and the RL vs. L problem. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC '06)*, 21–23 May 2006. Preliminary version on *ECCC*, February 2005.

[RV05] Eyal Rozenman and Salil Vadhan. Derandomized squaring of graphs. In *Approximation, Randomization and Combinatorial Optimization*, pages 436–447, 2005.

[Sam07] Alex Samorodnitsky. Low-degree tests at large distances. In *Proceedings of the* 39^{th} *annual ACM symposium on Theory of computing (STOC '07)*, pages 506–515, 2007.

[Smo87] R. Smolensky. Algebraic methods in the theory of lower bounds for Boolean circuit complexity. In *Proceedings of the* 19^{th} *annual ACM symposium on Theory of computing (STOC '87)*, pages 77–82, 1987.

[ST00] A. Samorodnitsky and L. Trevisan. A PCP characterization of NP with optimal amortized query complexity. In *Proceedings of the 32nd ACM symposium on Theory of Computation (STOC' 00)*, pages 191–199, 2000.

[ST06] A. Samorodnitsky and L. Trevisan. Gowers uniformity, influence of variables, and PCPs. In *Proceedings of the 38nd ACM symposium on Theory of Computation (STOC' 06)*, pages 11–20, 2006.

[Šte00] D. Štefankovič. Fourier transforms in computer science. Master's thesis, University of Chicago, Department of Computer Science, 2000. Available at http://www.cs.rochester.edu/ stefanko/Publications/Fourier.ps.

[STV99] Madhu Sudan, Luca Trevisan, and Salil Vadhan. Pseudorandom generators without the xor lemma. In *COCO '99: Proceedings of the Fourteenth Annual IEEE Conference on Computational Complexity*, page 4, Washington, DC, USA, 1999. IEEE Computer Society.

[Sud97] Madhu Sudan. Decoding of reed solomon codes beyond the error-correction bound. *Journal of Complexity*, 13:180–193, 1997.

[Sud00] Madhu Sudan. List decoding: Algorithms and applications. *SIGACT News*, 31:2000, 2000.

[Sud10] Madhu Sudan. Invariance in property testing. ECCC, TR10-051, 2010.

[SZ99] Michael Saks and Shiyu Zhou. Rspace(s) $\subseteq$ dspace($s^{3/2}$). *J. Comput. Syst. Sci.*, 58(2):376–403, 1999.

[Tao05] Terence Tao. Title: The dichotomy between structure and randomness, arithmetic progressions, and the primes. ICM lecture, 2005.

[TB98] Gabor Tardos and David A. Mix Barrington. A lower bound on the mod 6 degree of the OR function. *Computational Complexity*, 7:99–108, 1998.

[TZ09] Terence Tao and Tamar Ziegler. The inverse conjecture for the Gowers norm over finite fields via the correspondence principle. Submitted, 2009.

[Tzu] Yoav Tzur. $gf(2^n)$-linear tests versus $gf(2)$-linear tests. Electronic Colloquium on Computational Complexity, TR 09-018.

[Vio08] Emanuele Viola. The sum of d small-bias generators fools polynomials of degree d. In *Proceedings of the IEEE* 23^{rd} *Annual Conference on Computational Complexity (CCC '08)*, pages 124–127. IEEE Computer Society, 2008.

[Vio09] Emanuele Viola. Guest column: correlation bounds for polynomials over 0 1. *SIGACT News*, 40(1):27–44, 2009.

[VW08] Emanuele Viola and Avi Wigderson. Norms, xor lemmas, and lower bounds for polynomials and protocols. *Theory of Computing*, 4(1):137–168, 2008.

[Wei48] A. Weil. Sur les courbes algebriques et les varietes qui s'en deduisent. *Actualities Sci. et Ind.*, (1041), 1948.