On Elliptic Curves, the ABC Conjecture, and Polynomial Threshold Functions

A dissertation presented

by

Daniel Mertz Kane

to

The Department of Mathematics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Mathematics

Harvard University
Cambridge, Massachusetts

March, 2011

Dissertation Advisor: Professor Barry Mazur Daniel Mertz Kane

On Elliptic Curves, the ABC Conjecture, and Polynomial Threshold Functions

Abstract

We present a number of papers on topics in mathematics and theoretical computer science. Topics include: a problem relating to the ABC Conjecture, the ranks of 2-Selmer groups of twists of an elliptic curve, the Goldbach problem for primes in specified Chebotarev classes, explicit models for Deligne-Lusztig curves, constructions for small designs, noise sensitivity bounds for polynomials threshold functions, and pseudo-random generators for polynomial threshold functions.

Contents

Acknowledgements

I would like to thank those who have been influential in my mathematics education, namely Jonathan Kane, Ken Ono, the organizers of MOP, David Jerison, Victor Guillemin, Erik Demaine, Noam Elkies, Benedict Gross, Barry Mazur, and Henry Cohn.

I would also like to thank Jelani Nelson and Ilias Diakonikolas for getting me involved in much of my recent computer science work.

The material in Chapter E was developed while working as a summer intern at Microsoft Research New England.

I am also grateful for the NSF and NDSEG graduate fellowships which have helped support me financially during graduate school.

Introduction

My work in the past few years has been largely centered around problem solving. As a result I find myself with a number of results on separate topics, having little relation to one another. As such, this thesis consists of a number of largely independent results. Each Chapter in this thesis (with the exception of those in Part III, which share motivational background) is meant to stand on its own.

Part I contains several results in the field of number theory, mostly of the analytic variety.

Chapter A is based on [28] and studies a problem related to the ABC Conjecture. In particular I look at the number of solutions to the equation $A + B + C = 0$ for A, B, C pairwise relatively prime, highly divisible integers. The ABC Conjecture states that if "highly divisible" is defined strictly enough that there should be only finitely many such triples. In this Chapter, I obtain asymptotic results for the number of such solutions for a broad range of weaker meanings of "highly divisible". My major techniques are the use of lattice methods and bound do to Heath-Brown on the density of integer points on a conic.

Chapter B is based on [25], and deals with a variant of the Goldbach problem. In particular, I prove an asymptotic for the number of solutions to a linear equation in at least three prime numbers given that these primes are required to lie in specified Chebotarev classes. I make use of a number of analytic techniques including Hecke L-functions, the circle method and Vinogradov's method.

Chapter C is based on [29], and deals with the ranks of 2-Selmer groups of twists of an elliptic curve. In particular, I show that for a wide range of elliptic curves, Γ, that the ranks of the 2-Selmer groups of their twists satisfy a particular distribution. This extends a result of Swinnerton-Dyer ([58]), which gives the same asymptotics but with an unusual notion of density. I extend Swinnerton-Dyer's result to use the standard notion of density using a number of analytic techniques including L-functions and the large sieve.

Chapter D is based on [26], and produces explicit models for the Deligne-Lusztig varieties of dimension 1. The constructions attempt to be as canonical as possible, with a category-theoretic construction used at the core of my construction of the twisted Chevalley groups. In addition to finding models for the curves associated to ${}^2A_2$, ${}^2B_2$ and ${}^2G_2$, I also produce explicit isomorphisms $C/G^\sigma \to \mathbb{P}^1$.

Part II consists of Chapter E, which is based upon [30]. In it I deal with the construction

of designs of small size in a number of contexts. I find a construction for producing small designs by making use of a fixed-point theorem from algebraic topology. The bound on size turns out to be particularly nice in the case of designs on a homogeneous space. I also use a variant of this technique to find designs of nearly optimal size for the Beta distribution and on the sphere.

Part III deals with three results on polynomial threshold functions. It begins with an introduction to the topic providing common motivational background for the topic and discussing a couple of common themes in this work. It is suggested that the reader look at this introductory material before reading any of the associated chapters.

Chapter F is based on [27], and proves that limited independence is enough to fool polynomial threshold functions of Gaussians. The proof is based on a technique that I have been developing along with Jelani Nelson and Ilias Diakonikolas called the FT-Mollification method, which deals with problems involving k-independence by combining techniques from approximation theory with anticoncentration results. The main new innovation in this work is a structure theorem for polynomials that allows us to write them in terms of other polynomials each with small higher moments.

Chapter G is based on [32], and contains proofs of asymptotically optimal bounds on the Gaussian surface are and noise sensitivity of polynomial threshold functions. The bound on noise sensitivity is very short and consists largely of taking advantage of a particular symmetry. The bound on surface area is then derived from the noise sensitivity bound using a few simple ideas which unfortunately need to be supported by some non-trivial technical results.

Chapter H is based on [31], and introduces a new pseudorandom generator for polynomial threshold functions. Many of the ideas behind this result had their origins in the proofs from the previous two Chapters.

Part I

Number Theory

Chapter A

On A Problem Relating to the ABC Conjecture

1 Introduction

For an integer n, let $r(n) = \prod_{p|n} p$ be the product of its prime divisors. The ABC-Conjecture states that for any $\epsilon > 0$ there are only finitely many solutions to the equation $A+B+C = 0$ in relatively prime integers A, B, C so that $\max(|A|, |B|, |C|) > r(ABC)^{1+\epsilon}$. Although this conjecture is not known for any value of ϵ, when the restrictions on A, B, C are loosened there are conjectural predictions of the number of solutions that are more tractable. In particular:

Conjecture 1. *Given constants $0 < a, b, c \leq 1$, and $\epsilon > 0$, with $a + b + c > 1$ then for sufficiently large N the number of solutions to $A+B+C = 0$ in relatively prime integers A, B, C with $|A|, |B|, |C| \leq N$ and $r(A) \leq |A|^a, r(B) \leq |B|^b, r(C) \leq |C|^c$ is between $N^{a+b+c-1-\epsilon}$ and $N^{a+b+c-1+\epsilon}$.*

We will henceforth refer to the number of such triples A, B, C as the number of solutions to the ABC problem with parameters (a, b, c) or with parameters (a, b, c, N).

Conjecture 1 was stated (with slightly different terminology) in [43]. Furthermore, [43] alludes to a proof of this Conjecture in the case when $5/6 \leq a, b, c \leq 1$. We extend this result to a wider range of values. In particular, we show:

(Note: from this point onward when using an N^ϵ in asymptotic notation, it will be taken to mean that the bound holds for any $\epsilon > 0$, though the asymptotic constants may depend on ϵ.)

Theorem 1. *For $0 < a, b, c \leq 1$, with $\min(a, b, c) + \max(a, b, c) > 1$, the number of solutions to the ABC problem with parameters (a, b, c, N) is $\Omega(N^{a+b+c-1})$.*

Theorem 2. *For $0 < a, b, c \leq 1$, the number of solutions to the ABC problem with parameters (a, b, c, N) is $O(N^{a+b+c-1+\epsilon} + N^{1+\epsilon})$.*

Together these Theorems provide a proof of Conjecture 1 whenever $a + b + c \geq 2$.

We use two main techniques to prove these Theorems. First we use lattice methods. The idea is to fix the non-squarefree parts of A, B, C and to then count the number of solutions. For example if A, B, C have non-squarefree parts α, β, γ, we need to count solutions to an equation of the form $\alpha X + \beta Y + \gamma Z = 0$.

Secondly we use a result of Heath-Brown on the number of integer points on a conic. The idea here is to note that if A, B, C are highly divisible they are likely divisible by large squares. Writing $A = \alpha X^2, B = \beta Y^2, C = \gamma Z^2$, then for fixed values of α, β, γ the solutions to the ABC problem correspond to integer points of small size on a particular conic.

In Section 2 we cover the lattice methods. In particular, in Section 2.1, we use these methods to prove Theorem 1. The proof of our upper bound will involve breaking our argument into cases based upon the approximate sizes of various parameters of A, B, and C. In Section 2.2, we introduce some ideas and notation that we will use when making these arguments. In Section 2.3, we use lattice methods to prove an upper bound on the number of solutions. These techniques will work best when A, B, and C have relatively few repeated factors. In Section 3, we prove another upper bound, this time using our methods involving conics. These results will turn out to be most effective when A,B and C have many repeated factors. Finally, we combine these results with our upper bound from the previous Section to prove Theorem 2.

2 Lattice Methods

2.1 Lattice Lower Bounds

We begin with the following Lemma:

Lemma 3. *Let L be a two dimensional lattice and P a convex polygon. Let m be the minimum separation between points of L. Then*

$$|L \cap P| = \frac{Volume(P)}{CoVolume(L)} + O\left(\frac{Perimeter(P)}{m} + 1\right).$$

Proof. Begin with a reduced basis of L. We apply a linear transformation to the problem so that L is a square lattice. We can do this since this operation has no effect on $|L \cap P|$ or $\frac{\text{Volume}(P)}{\text{CoVolume}(L)}$ and increases $\frac{\text{Perimeter}(P)}{m}$ by at most a constant factor. We now note that if we draw a fundamental domain around each point of $|L \cap P|$, their union is sandwiched between the set of points within distance $\sqrt{2}m$ of P, and the set of points where the disc of radius

$\sqrt{2}m$ around them is contained in P. Computing the areas of these two regions gives our result. □

We are now prepared to prove our lower bound.

proof of Theorem 1. Assume without loss of generality that $a \leq b \leq c$. We have by assumption that $a + c > 1$.

Pick x, y, z the smallest possible integers so that $2^{x-1} > N^{1-a}$, $3^{y-1} > N^{1-b}$, $5^{z-1} > N^{1-c}$. We will attempt to find solutions of the form $A = X2^x, B = Y3^y, C = Z5^z$. Let L be the lattice of triples of integers A, B, C so that $A + B + C = 0$, $2^x|A$, $3^y|B$, $5^z|C$. Let L_n be the sublattice of L so that $n|A, B, C$. Using an appropriate normalization of the volume of the plane $A + B + C = 0$, the covolume of L is $U := 2^x 3^y 5^z = \Theta(N^{3-a-b-c})$. The covolume of L_n is $n^2 U/\gcd(n, 30)$. Let v be the shortest vector in L. Let $m = |v|$. Let M be the length of the shortest vector which is not a multiple of v. Note that the vectors $(0, 3^y 5^z, -3^y 5^z), (2^x 5^z, 0, -2^x 5^z)$ are in L and are linearly independent. Therefore M is at most the larger of the lengths of these vectors so $M = O(N^{1-\delta})$ for some $\delta > 0$. Note also that $U = \Theta(mM)$.

Let P be the polygon in the plane $A + B + C = 0$ defined by $|A|, |B|, |C| \leq N$. The number of solutions to the ABC problem with parameters (a, b, c, N) is at least the number of points in $L \cap P$ with relatively prime coordinates. This is

$$\sum_n \mu(n)|L_n \cap P|.$$

Removing the points that are multiples of v we get

$$\sum_{n=1}^{O(N/M)} \mu(n)|L_n \cap P \backslash \langle v \rangle|.$$

(We can stop the sum at $O(N/M)$ since if $w \in L_n$ for n squarefree, then $w/(n/\gcd(n, 30)) \in L$ and all points of L not a multiple of v are of norm at least M) Letting V be the volume of P and noting that the shortest vector in L_n has length $\Theta(nm)$, the above equals

$$\sum_{n=1}^{O(N/M)} \left(\frac{\mu(n)\gcd(n, 30)}{n^2}\right)\left(\frac{V}{U}\right) + O(N/(nm) + 1).$$

The main term is

$$\begin{aligned}\left(\frac{V}{U}\right)\left(\sum_{n=1}^{\infty} \frac{\mu(n)\gcd(n, 30)}{n^2} + O(M/N)\right) &= \Theta\left(\frac{V}{U}\right) \\ &= \Theta\left(\frac{N^2}{N^{3-a-b-c}}\right) \\ &= \Theta(N^{a+b+c-1}).\end{aligned}$$

The error term is

$$\begin{aligned} O(\log(N)N/m + N/M) &= O(\log(N)NM/U + N/\sqrt{U}) \\ &= O(\log(N)N^{a+b+c-1-\delta} + N^{(a+b+c-1)/2}). \end{aligned}$$

This completes our proof. □

2.2 Diadic Intervals

Before beginning our work on upper bounds, we discuss some ideas involving diadic intervals that we will make use of. First a definition:

Definition. *A diadic interval is an interval of the form* $[2^n, 2^{n+1}]$ *for some integer* n.

In the process of proving upper bounds we will often wish to count the number of solutions to an ABC problem in which some functions of A, B, C lie in fixed diadic intervals. We may for example claim that the number of solutions to an ABC problem where $f(A), f(B), f(C)$ lie in fixed diadic intervals is $O(X)$. Here f will be some specified function and we are claiming that for any triple of diadic intervals I_A, I_B, I_C the number of A, B, C that are solutions to the appropriate ABC problem and so that additionally $f(A) \in I_A, f(B) \in I_B$ and $f(C) \in I_C$ is $O(X)$. When we do this, it will often be the case that X depends on $f(A), f(B), f(C)$ and not just the parameters of the original ABC problem we were trying to solve. By this we mean that our upper bound is valid if the $f(A), f(B), f(C)$ appearing in it are replaced by any numbers in the appropriate diadic intervals. Generally this freedom will not matter since fixing diadic intervals for $f(A), f(B), f(C)$ already fixes their values up to a multiplicative constant. It should also be noted that $[1, N]$ can be covered by $O(\log N)$ diadic intervals. Thus if we prove bounds on the number of solutions in which a finite number of parameters (each at most N) lie in fixed diadic intervals, we obtain an upper bound for the number of solutions with no such restrictions that is at most N^ϵ larger than the bound for the worst set of intervals.

2.3 Lattice Upper Bounds

In order to prove upper bounds, we will need a slightly different form of Lemma 3.

Lemma 4. *Let* L *be a 2 dimensional lattice and* P *a convex polygon, centrally symmetric about the origin. Then the number of vectors in* $L \cap P$ *which are not positive integer multiples of other vectors in* L *is*

$$O\left(\frac{Volume(P)}{CoVolume(L)} + 1\right).$$

Proof. If $L \cap P$ only contains the origin or multiples of a single vector, the result follows trivially. Otherwise P contains two linearly independent vectors of L. This means that P contains at least half of some fundamental domain. Therefore $2P$ contains some whole fundamental domain. Therefore $4P$ contains all of the fundamental domains centered at any of the points in $L \cap P$. Therefore $|L \cap P| = O\left(\frac{\text{Volume}(P)}{\text{CoVolume}(L)}\right)$. □

We will also make extensive use of the following proposition:

Proposition 5. *For any m, the number of k with $|k| \leq N$ and $r(k) = m$ is $O(N^\epsilon)$, where the implied constant depends on ϵ but not N or m.*

Proof. Let $m = p_1 p_2 \cdots p_n$, where $p_1 < p_2 < \cdots < p_n$ are primes (if m is not squarefree we have no solutions). Then all such k must be of the form $\prod_{i=1}^n p_i^{a_i}$ for some integers $a_i \geq 1$ with $\sum_i a_i \log(p_i) \leq \log(N)$. Note that if for each such k you consider the unit cube defined by $\prod_{i=1}^n [a_i - 1, a_i] \subset \mathbb{R}^n$, these cubes have disjoint interiors and are contained in a simplex of volume $\frac{1}{n!} \prod_i \frac{\log(N)}{\log(p_i)}$. Hence the number of such k is at most

$$\frac{1}{n!} \prod_i \frac{\log(N)}{\log(p_i)} = O\left(\frac{\log(N)}{n}\right)^n.$$

Now we must also have that $n! \leq \prod_i p_i \leq N$ or there will be no solutions, so $n = O\left(\frac{\log(N)}{\log\log(N)}\right)$. Now $n \log N - n \log n$ is increasing for $n < N/e$ so the number of solutions is at most

$$\begin{aligned} O\left(\frac{\log(N)}{\log(N)/\log\log(N)}\right)^{O\left(\frac{\log(N)}{\log\log(N)}\right)} &= \exp\left(O\left(\frac{\log(N)\log\log\log(N)}{\log\log(N)}\right)\right) \\ &= O(N^\epsilon). \end{aligned}$$

□

We now need some more definitions. For an integer n define

$$u(n) := \prod_{p||n} p$$

to be the product of primes that divide n exactly once. Let

$$e(n) := \prod_{p^\alpha||n, \alpha>1} p^\alpha = n/u(n)$$

be the product of primes dividing n more than once counted with their appropriate multiplicity. Finally, let

$$v(n) = \prod_{p^2|n} p = r(n)/u(n) = r(e(n))$$

be the product of primes dividing n more than once.

We can now prove the first part of our upper bound

Proposition 6. *Fix $0 < a, b, c \leq 1$. The number of solutions to the ABC problem with parameters (a, b, c, N) and with $v(A), v(B), v(C)$ lying in fixed diadic intervals is*

$$O(N^{a+b+c-1+\epsilon} + v(A)v(B)v(C)N^{\epsilon}).$$

Note that since $v(A), v(B), v(C) \leq \sqrt{N}$, this would already give a weaker but non-trivial version of Theorem 2.

Proof. We will begin by additionally fixing diadic intervals for $|A|, |B|, |C|$, $u(A)$, $u(B), u(C)$, $e(A), e(B), e(C)$. Since there are only $\log(N)$ possible intervals for each, our total number of solutions will be greater by a factor of at most a factor of $O(N^{\epsilon})$. Assume without loss of generality that $|A| \leq |B| \leq |C|$.

There are $O(v(A)v(B)v(C))$ ways to fix the values of $v(A), v(B), v(C)$ within their respective diadic intervals. Given these, by Proposition 5 there are $O(N^{\epsilon})$ possible values of $e(A), e(B), e(C)$. Pick a triple of values for $e(A), e(B), e(C)$. We assume these are relatively prime, for otherwise they could not correspond to any valid solutions to our ABC problem. Define the lattice L to consist of triples of integers which sum to 0, and are divisible by $e(A), e(B), e(C)$ respectively. We define the polygon P to be the set of (x_1, x_2, x_3) so that $x_1 + x_2 + x_3 = 0$ and $|x_1|$ is bounded by the upper end of the diadic interval for $|A|$, and $|x_2|, |x_3|$ are likewise bounded by the intervals for $|B|$ and $|C|$. The number of solutions to our ABC problem with the specified values of $e(A), e(B), e(C)$ and with $|A|, |B|, |C|$ in the appropriate diadic intervals is at most the number of vectors in $L \cap P$ that are not positive integer multiples of other vectors in L. By Lemma 4 this is

$$O\left(\frac{\text{Volume}(P)}{\text{CoVolume}(L)} + 1\right) = O\left(\frac{|AB|}{e(A)e(B)e(C)} + 1\right).$$

Multiplying this by the number of ways we had to choose values for $v(A)$, $v(B)$, $v(C)$, $e(A), e(B), e(C)$, we get that the total number of solutions to our original ABC problem with the specified diadic intervals for $|A|, v(A), e(A)$, etc. is at most

$$\begin{aligned} &O\left(\frac{N^{\epsilon}|ABC|v(A)v(B)v(C)}{e(A)e(B)e(C)|C|} + N^{\epsilon}v(A)v(B)v(C)\right) \\ &= O\left(\frac{N^{\epsilon}r(A)r(B)r(C)}{|C|} + N^{\epsilon}v(A)v(B)v(C)\right) \\ &\leq O(N^{\epsilon}|A|^{a}|B|^{b}|C|^{c-1} + N^{\epsilon}v(A)v(B)v(C)). \end{aligned}$$

Where the last step comes from noting that for any solution to our ABC problem, $r(A) \leq |A|^{a}$, etc. Since $a+b+c > 1$, $|A^{a}B^{b}C^{c-1}|$ is maximized with respect to $N \geq |C| \geq |B| \geq |A|$

when $|A| = |B| = |C| = N$. So we obtain the bound

$$O(N^{a+b+c-1+\epsilon} + N^{\epsilon} v(A)v(B)v(C)).$$

□

3 Points on Conics

The bound from Proposition 6 is useful so long as $v(A), v(B), v(C)$ are not too big. When they are large, we shall use different techniques. In particular, if $v(A), v(B), v(C)$ are large, then A, B, C are divisible by large squares. We will fix non-square parts of these numbers and bound the number of solutions using a Theorem from [23]

Theorem 7 ([23] Theorem 2)**.** *Let q be an integral ternary quadratic form with matrix M. Let $\Delta = |\det M|$, and assume that $\Delta \neq 0$. Write Δ_0 for the highest common factor of the 2×2 minors of M. Then the number of primitive integer solutions of $q(x) = 0$ in the box $|x_i| \leq Ri$ is*

$$\ll \{1 + (\frac{R_1 R_2 R_3 \Delta_0^2}{\Delta})^{1/2}\} d_3(\Delta).$$

Where $d_3(\Delta)$ is the number of ways of writing Δ as a product of three integers.

Putting this into a form that fits our needs slightly better:

Corollary 8. *For a, b, c relatively prime integers, the number of solutions to $aX^2 + bY^2 + cZ^2 = 0$ in relatively prime integers X, Y, Z with $|X| \leq R_1$, $|Y| \leq R_2$, $|Z| \leq R_3$ is*

$$O\left(\left(1 + \sqrt{\frac{R_1 R_2 R_3}{|abc|}}\right)(|abc|)^{\epsilon}\right).$$

Proof. This follows immediately from applying the above Theorem to the obvious quadratic form, noting that $\Delta = |abc|$, $\Delta_0 = 1$ and that $d_3(N) = O(N^{\epsilon})$. □

We now have all of the machinery ready to prove the upper bound. We make one final pair of definitions.

Let

$$S(n) := \prod_{p^{\alpha} || n} p^{[\alpha/2]} = \sup\{m : m^2 | n\}.$$

be the largest number whose square divides n. Let

$$T(n) = |n/S(n)^2|.$$

We use Corollary 8 to prove another upper bound.

Proposition 9. *Fix diadic intervals for $S(A), T(A), S(B), T(B), S(C)$, and $T(C)$. The number of solutions of the ABC problem with S and T of A, B, C lying in these intervals is*

$$O\left(\left(T(A)T(B)T(C) + \sqrt{S(A)S(B)S(C)T(A)T(B)T(C)}\right)N^{\epsilon}\right).$$

Proof. There are $O(T(A)T(B)T(C))$ choices for the values of $T(A), T(B)$, $T(C)$ lying in their appropriate intervals. Fixing these values, we count the number of solutions to

$$\pm T(A)S(A)^2 \pm T(B)S(B)^2 \pm T(C)S(C)^2 = 0$$

with $S(A), S(B), S(C)$ relatively prime and in the appropriate intervals. By Corollary 8 this is at most

$$O\left(\left(1 + \sqrt{\frac{S(A)S(B)S(C)}{T(A)T(B)T(C)}}\right)N^{\epsilon}\right).$$

Hence the total number of solutions to our ABC problem with $T(A), S(A)$, etc. lying in appropriate intervals is

$$O\left(\left(T(A)T(B)T(C) + \sqrt{S(A)S(B)S(C)T(A)T(B)T(C)}\right)N^{\epsilon}\right).$$

□

We are now prepared to prove our upper bound on the number of solutions to an ABC problem.

Proof of Theorem 2. It is enough to prove our Theorem after fixing $S(A)$, $T(A)$, $v(A)$, $S(B), T(B), v(B)$, $S(C), T(C), v(C)$ to all lie in fixed diadic intervals, since there are only $O(\log(N)^9) = O(N^{\epsilon})$ choices of these intervals. It should be noted that $S(n) \geq v(n)$ for all n. By Proposition 6 we have the number of solutions is at most

$$O\left(N^{a+b+c-1+\epsilon} + S(A)S(B)S(C)N^{\epsilon}\right).$$

By Proposition 9, noting that $N \geq T(A)S(A)^2, T(B)S(B)^2, T(C)S(C)^2$, we know that the number of solutions is at most

$$O\left(N^{3+\epsilon}(S(A)S(B)S(C))^{-2} + N^{3/2+\epsilon}(S(A)S(B)S(C))^{-1/2}\right).$$

If $S(A)S(B)S(C) \geq N$ this latter bound is $O(N^{1+\epsilon})$ and if $S(A)S(B)S(C) \leq N$, the former bound is $O(N^{a+b+c-1+\epsilon} + N^{1+\epsilon})$. So in either case we have our desired bound. □

4 Conclusion

We have proven several bounds on the number of solutions to ABC type problems. In particular we have proven the Conjecture 1 so long as $a + b + c \geq 2$.

Our lower bounds can not be extended by this technique because if a, b, c are much smaller we will not be guaranteed that the lattices we look at will have anything more than multiples of their shortest vector in the range we are concerned with.

Our upper bounds likely cannot be extended much either, because when $S(A)S(B)S(C) \sim N$, then Corollary 8 only says that we have $O(N^\epsilon)$ solutions for each choice of the T's. To do better than this, we would need to show that for some reasonable fraction of T's that there were no solutions.

Chapter B

The Goldbach Problem for Primes in Chebotarev Classes

1 Introduction and Statement of Results

In 1937 Vinogradov proved that any sufficiently large odd number could be written as the sum of three primes. In addition, he managed to provide an asymptotic for the number of ways to do so, proving (as stated in [24] Theorem 19.2)

Theorem 1 (Vinogradov, statement taken from [24] Theorem 19.2)**.** *For N a positive integer and A any real number then*

$$\sum_{n_1+n_2+n_3=N} \Lambda(n_1)\Lambda(n_2)\Lambda(n_3) = \mathfrak{G}_3(N)N^2 + O(N^2 \log^{-A}(N)). \tag{1}$$

Where

$$\mathfrak{G}_3(N) = \frac{1}{2}\prod_{p|N}(1-(p-1)^{-2})\prod_{p\nmid N}(1+(p-1)^{-3}),$$

$\Lambda(n)$ is the Von Mangoldt function, and the asymptotic constant in the O depends on A.

It is easy to see that the contribution to the left hand side of Equation 1 coming from one of the n_i a power of prime is negligible, and thus this side of the equation may be replaced by a sum over triples p_1, p_2, p_3 of primes that sum to N of $\log(p_1)\log(p_2)\log(p_3)$. This implies that any sufficiently large odd number can be written as a sum of three primes since $\mathfrak{G}_3(N)$ is bounded below by a constant for N odd.

It should also be noted that the main term, $N^2\mathfrak{G}_3(N)$ can be written as

$$C_\infty \prod_p C_p.$$

Where

$$C_\infty = \frac{N^2}{2}$$

and

$$C_p = \begin{cases} (1-(p-1)^{-2}) & \text{if } p|N \\ (1+(p-1)^{-3}) & \text{else} \end{cases}.$$

When written this way, there is a reasonable heuristic explanation for Theorem 1. To begin with, the Prime Number Theorem says that the Von Mangoldt function, is approximated by the distribution assigning 1 to each positive integer. C_∞ provides an approximation to the number of solutions based on this heuristic. The C_p can be thought of as corrections to this heuristic. They can be thought of as local contributions coming from congruential information about the primes p_i. C_p can easily be seen to be equal to

$$\frac{p\#\left\{(n_1,n_2,n_3) \in \left((\mathbb{Z}/p\mathbb{Z})^*\right)^3 : n_1+n_2+n_3 \equiv N \pmod p\right\}}{(p-1)^3}.$$

This can be thought of as a correction factor coming from the fact that no prime (except for p) is a multiple of p. This term is the ratio of the probability that three randomly chosen non-multiples of p add up to N modulo p divided by the probability that three randomly chosen (arbitrary) numbers add up to N modulo p.

There have been a number of generalizations to Theorem 1 over the years. $\mathfrak{G}(N)$ is viewed in the light described above, it is straightforward to generalize Theorem 1 to find asymptotics for the number of solutions to any linear equation in at least three prime numbers. In particular for fixed non-zero integers $a_1, \ldots, a_k$ ($k \geq 3$), and integers X and N, one can prove a formula similar to Equation 1 to approximate the sum over primes $p_1, \ldots, p_k \leq X$ so that $\sum_{i=1}^k a_i p_i = N$ of $\prod_{i=1}^k \log(p_i)$. Furthermore if $k = 2$, one can show that the analogous formula will hold for all but $O(X \log^{-B}(X))$ of the possible values for N.

In [60], this result was generalized to counting the number of representations of N as a sum of three primes all within approximately $N^{5/6}$ of $N/3$. In [54], bounds on the number of ways of writing N as sums of primes were made explicit enough to prove that any even natural number could be written as the sum of at most 18 primes. Other generalizations include [50] and [61], which generalize the Goldbach problem to number fields, and count the number of ways of writing elements of the ring of integers as a sum of elements generating prime ideals. Several papers, such as [19] and [46], deal with the problem of writing N as a sum of primes each taken from a specified arithmetic progression.

In this Chapter, we prove a new generalization of Theorem 1, counting solutions to similar equations where in addition the primes p_i are required to lie in specified Chebotarev classes. In particular, after fixing Galois extensions $K_i/\mathbb{Q}$ and conjugacy classes C_i of $\mathrm{Gal}(K_i/\mathbb{Q})$, we find an asymptotic for the sum of $\prod_{i=1}^k \log(p_i)$ over primes $p_1, \ldots, p_k \leq X$ so that $[K_i/\mathbb{Q}, p_i] = C_i$ for each i and $\sum_{i=1}^k a_i p_i = N$. Note that the results on writing N as a sum of primes

from arithmetic progressions, will follow as a special case of this when K_i is abelian over $\mathbb{Q}$ (although our bounds are probably worse). In particular we prove:

Theorem 2. *Let $k \geq 3$ be an integer. Let $K_i/\mathbb{Q}$ be finite Galois extensions $(1 \leq i \leq k)$ and $G_i = Gal(K_i/\mathbb{Q})$. Let $a_1, \ldots, a_k$ be non-zero integers with no common divisor. Let C_i be a conjugacy class of G_i for each i. Let K_i^a be the maximal abelian extension of $\mathbb{Q}$ contained in K_i, and let D_i be its discriminant. Let D be the least common multiple of the D_i. Let H_i^0 be the subgroup of $(\mathbb{Z}/D\mathbb{Z})^*$ corresponding to K_i^a via global class field theory. Let H_i be the coset of H_i^0 corresponding to the projection of an element of C_i to $Gal(K_i^a/\mathbb{Q})$. Additionally let N be an integer and let A and X be positive numbers, then*

$$\sum_{\substack{p_i \leq X \\ [K_i/\mathbb{Q}, p_i] = C_i \\ \sum_i a_i p_i = N}} \prod_{i=1}^{k} \log(p_i) = \left(\prod_{i=1}^{k} \frac{|C_i|}{|G_i|} \right) C_\infty C_D \left(\prod_{p \nmid D} C_p \right) + O\left(X^{k-1} \log^{-A}(X) \right). \tag{2}$$

Where the sum on the right hand side is over sets of prime numbers $p_1, \ldots, p_k$, each at most X, so that $\sum_{i=1}^{k} a_i p_i = N$, and so that the Artin symbol $[K_i/\mathbb{Q}, p_i]$ lands in the conjugacy class C_i of G_i for all $1 \leq i \leq k$. On the right hand side,

$$C_\infty = \int_{\substack{x_i \in [0,X] \\ \sum_i a_i x_i = N}} \left(\sum_{i=1}^{k} a_i \frac{\partial}{\partial x_i} \right) dx_1 \wedge dx_2 \wedge \ldots \wedge dx_k,$$

$$C_D = D \left(\frac{\#\{(x_i) \in ((\mathbb{Z}/D\mathbb{Z})^*)^k : x_i \in H_i, \sum_{i=1}^{k} a_i x_i \equiv N \pmod{D}\}}{\prod_{i=1}^{k} |H_i|} \right),$$

and the second product is over primes p not dividing D of

$$C_p = p \left(\frac{\#\{(x_i) \in ((\mathbb{Z}/p\mathbb{Z})^*)^k : \sum_{i=1}^{k} a_i x_i \equiv N \pmod{D}\}}{(p-1)^k} \right).$$

The implied constant in the O term may depend on k, K_i, C_i, a_i, and A, but not on X or N. Additionally, if $k = 2$ and K_i, C_i, a_i, A, X are fixed, then Equation (2) holds for all but $O(X \log^{-A}(X))$ values of N.

The introduction of Chebotarev classes leads to two main differences between our asymptotic and the classical one. For one, the Chebotarev Density Theorem tells us that there are fewer primes in these Chebotarev classes than out of them and causes us to introduce a factor of $\prod_{i=1}^{k} \left(\frac{|C_i|}{|G_i|} \right)$. Secondly, Global Class Field Theory tells us that the prime p_i will necessarily lie in the subset H_i of $(\mathbb{Z}/D\mathbb{Z})^*$, giving us the correction factor C_D rather than $\prod_{p|D} C_p$ to account for required congruence relations that these primes satisfy.

It should be noted that the error term is $o(X^{k-1}\log(X)^{-1})$, whereas if N is bounded away from both the largest and smallest possible values that can be taken by $\sum_i a_i x_i$ for $x_i \in [0, X]$, then C_∞ will be on the order of X^{k-1}. For K_i, C_i fixed, the first term on the right hand side is a constant. C_D is either 0 or is bounded away from both 0 and ∞. Lastly, for $p \nmid Dn\prod_i a_i$, inclusion-exclusion tells us that $C_p = 1 + O(p^{-2})$, and for $p|N$, $p \nmid D\prod_i a_i$, $C_p = 1 + O(p^{-1})$. This means that unless $C_p = 0$ for some p, $\prod_p C_p$ is within a bounded multiple of $\prod_{p|N}(1 + O(p^{-1})) = \exp(O(\log\log\log N))$. Therefore, unless $C_D = 0$, $C_p = 0$ for some p, or N is near the boundary of the available range, the main term on the right hand side of Equation 2 dominates the error.

2 Overview

Our proof will closely mimic the proof in [24] of Theorem 1. We provide a brief overview of the proof given in [24], discuss our generalization and provide an outline for the rest of the Chapter.

2.1 The Proof of Theorem 1

On a very generally level, the proof given in [24] depends on writing

$$\Lambda = \Lambda^\sharp + \Lambda^\flat.$$

Here $\Lambda^\sharp$ is a nice approximation to the Von Mangoldt function obtained essentially by Sieving out multiples of small primes and $\Lambda^\flat$ is an error term. It is relatively easy to deal with the sum

$$\sum_{n_1+n_2+n_3=N} \Lambda^\sharp(n_1)\Lambda^\sharp(n_2)\Lambda^\sharp(n_3),$$

yielding the main term in Equation 1. This leaves additional terms, each involving at least one $\Lambda^\flat$. These terms are dealt with by showing that $\Lambda^\flat$ is small in the sense that its generating function has small L^∞ norm.

To prove this bound on $\Lambda^\flat$, they make use of [24] Theorem 13.10, which states that for any A

$$\sum_{m\le x} \mu(m)e^{2\pi i\alpha m} \ll x\log^{-A}(x)$$

(μ is the Möbius function) with the implied constant depending only on A. This in turn is proved by splitting into cases based on whether or not α is near a rational number of small denominator.

If α is close to a rational number, the sum can be bounded through the use of Dirichlet L-functions. They prove bounds on $\sum_{n\le x}\chi(n)\mu(n)$ for χ a Dirichlet character ([24] (5.80)).

To prove this, their main ingredients are their Theorem 5.13, which gives bounds on the sums of coefficients of the logarithmic derivative of an L-function, along with some bounds on zero-free regions and Siegel zeroes.

If α is not well approximated by a rational number with small denominator, their bound is proved by rewriting the sum using some combinatorial identities ([24] (13.39)) and using the quadratic form trick. (Their actual bound is given in [24] Theorem 13.9.)

2.2 Outline of Our Proof

Our proof of Theorem 2 is similar in spirit to the proof of Theorem 1 given in [24]. We differ in a few ways, some just in the way we choose to organize our information and some from necessary complications due to the increased generality. We provide below and outline of our proof and a comparison of our techniques to those used in [24].

Instead of dealing directly with $\Lambda, \Lambda^\sharp$ and $\Lambda^\flat$ as is done in [24], we instead deal directly with their generating functions. In Section 3.1, we define G, which is our equivalent of the generating function for Λ. As it turns out, G is somewhat difficult to deal with directly, so we define a related function F, that is better suited for techniques involving Hecke L-functions. In Proposition 3 we prove that we can write G approximately as an appropriate sum of F's.

In Section 3.2 we define $G^\sharp$ and $G^\flat$, which are analogues of the generating functions for $\Lambda^\sharp$ and $\Lambda^\flat$. We also define analogous $F^\sharp$ and $F^\flat$. We do our sieving to write $G^\sharp$ slightly differently than the way [24] does to define $\Lambda^\sharp$, essentially writing our version of $\Lambda^\sharp$ as a product of local factors. This will later on produce some sums over smooth numbers, so in Lemma 6 we bound the number of smooth numbers, so that we may bound errors coming from sums over them.

We next work on proving that $F^\flat$ has small L^∞ norm (this is somewhat equivalent to [24] showing that generating functions of $\Lambda^\flat$ or μ are small). As in [24], we split into two cases based on whether or not we are near a rational number.

In Section 4.1 we deal with the approximation near rationals. First, in Section 4.1.1, we generalize some necessary results about L-functions and Siegel zeroes. In Section 4.1.2, we use these to produce an approximation of F, and in Section 4.1.3, we show that this also approximates $F^\sharp$.

In Section 4.2, we deal with showing that $F^\flat$ is small away from rationals. This is where we need to most diverge from the proof in [24]. The primary reason for this is that the classical case reduces the problem to bounds on trigonometric sums, which are relatively easy. Unfortunately, the analogue of these sums in our setting is a sum over ideals $\mathfrak{a}$ of $e^{2\pi i \alpha N(\mathfrak{a})}$, or something similar. In order to deal with these sums, we need some results about when multiples of a number with poor rational approximation have a good rational approximation (in Section 4.2.2). In Section 4.2.3, we use this result to prove bound on sums of the type described above, and in Section 4.2.4 use these bounds to prove bounds on F. In Section 4.2.1 we prove bounds for $F^\sharp$. Combining these bounds, we obtain bounds on $F^\flat$.

In Section, 5 we use our bounds on $F^\flat$ to prove bounds on $G^\flat$. Finally, in Section 6, we use this bound to prove Theorem 2. In Section 6.1, we introduce the appropriate product generating functions, and deal with the terms coming from $G^\flat$'s. In Section 6.2, we produce the main term of our Theorem.

Finally, in Section 7, we show an application of our Theorem to constructing elliptic curves whose discriminants split completely over specified number fields.

3 Preliminaries

In this Section, we introduce some of the basic terminology and results that will be used throughout the rest of the Chapter. In Section 3.1, we define the functions F and G along with some of the basic facts relating them. In Section 3.2, we define $F^\sharp$ and $G^\sharp$ along with some related terminology and again prove some basic facts. Finally, in Section 3.3, we prove a result on the distribution of smooth numbers that will prove useful to us later.

3.1 G and F

We begin with a standard definition:

Definition. *Let $e(x)$ denote the function $e(x) = e^{2\pi i x}$.*

We now define G as the generating function for the set primes $p \leq X$ with $[K/\mathbb{Q}, p] = C$ each weighted by $\log(p)$.

Definition. *Suppose that $K/\mathbb{Q}$ is a finite Galois extension, $G = Gal(K/\mathbb{Q})$, C a conjugacy class of G, and X a positive real number. We then define the generating function*

$$G_{K,C,X}(\alpha) = \sum_{\substack{p \leq X \\ [K/\mathbb{Q},p]=C}} \log(p) e(\alpha p).$$

Where the sum is over primes, p, with $p \leq X$ and $[K/\mathbb{Q}, p] = C$.

G is a little awkward to deal with and we would rather work with a related function defined in terms of characters. We first need one auxiliary definition:

Definition. *Let $L/\mathbb{Q}$ be a number field. Let Λ_L be the Von Mangoldt function on ideals of L, $\Lambda_L : \{\text{Ideals of } L\} \to \mathbb{R}$ defined by*

$$\Lambda_L(\mathfrak{a}) = \begin{cases} \log(N(\mathfrak{p})) & \text{if } \mathfrak{a} = \mathfrak{p}^n \\ 0 & \text{otherwise} \end{cases}$$

which assigns $\log(N(\mathfrak{p}))$ to a power of a prime ideal $\mathfrak{p}$, and 0 to ideals that are not powers of primes.

We now define

Definition. *If $L/\mathbb{Q}$ is a number field, ξ a Grossencharacter of L, and X a positive number, define the function*

$$F_{L,\xi,X}(\alpha) = \sum_{N(\mathfrak{a})\leq X} \Lambda_L(\mathfrak{a})\xi(\mathfrak{a})e(\alpha N(\mathfrak{a})).$$

Where the sum above is over ideals $\mathfrak{a}$ of L with norm at most X.

Notice that the sum in the definition of F is determined up to $O(\sqrt{X})$ by the terms coming from $\mathfrak{a}$ a prime splitting completely over $\mathbb{Q}$.

For both F and G, we will often suppress some of the subscripts when they are clear from context. We now demonstrate the relationship between F and G.

Proposition 3. *Let K and C be as above. Pick a $c \in C$. Let $L \subseteq K$ be the fixed field of c. Then we have that*

$$G_{K,C,X}(\alpha) = \frac{|C|}{|G|}\left(\sum_{\chi} \overline{\chi}(c)F_{L,\chi,X}(\alpha)\right) + O(\sqrt{X}). \tag{3}$$

Where the sum is over characters χ of the subgroup $\langle c\rangle \subset G$, which, by global class field theory, can be thought of as characters of L.

Proof. We begin by considering the sum on the right hand side of Equation (3). It is equal to

$$\begin{aligned}\sum_{\chi} \overline{\chi}(c)F_{L,\chi,X}(\alpha) &= \sum_{\chi}\sum_{N(\mathfrak{a})\leq X} \Lambda_L(\mathfrak{a})\overline{\chi}(c)\chi(\mathfrak{a})e(\alpha N(\mathfrak{a})) \\ &= \sum_{N(\mathfrak{a})\leq X} \Lambda_L(\mathfrak{a})e(\alpha N(\mathfrak{a}))\sum_{\chi}\overline{\chi}(c)\chi([K/L,\mathfrak{a}]) \\ &= \operatorname{ord}(c)\sum_{\substack{N(\mathfrak{a})\leq X \\ [K/L,\mathfrak{a}]=c}} \Lambda_L(\mathfrak{a})e(\alpha N(\mathfrak{a})).\end{aligned}$$

Up to an error of $O(\sqrt{X})$, we can ignore the contributions from elements whose norms are powers of primes, because there are $O(\sqrt{X}/\log(X))$ higher powers of primes with norm at most X. Therefore the above equals

$$\operatorname{ord}(c)\sum_{\substack{N(\mathfrak{p})\leq X \\ [K/L,\mathfrak{p}]=c \\ N(\mathfrak{p}) \text{ is prime}}} \log(N(\mathfrak{p}))e(\alpha N(\mathfrak{p})) + O(\sqrt{X}).$$

We need to determine now which primes $p \in \mathbb{Z}$ are the norm of an ideal $\mathfrak{p}$ of L with $[K/L, \mathfrak{p}] = c$, and for such p, how many such $\mathfrak{p}$ lie over it. Each such $\mathfrak{p}$ must have only one prime $\mathfrak{q}$ of K over it and it must be the case that $[K/\mathbb{Q}, \mathfrak{q}] = c$. Hence the p we wish to find are exactly those that have a prime $\mathfrak{q}$ lying over them with $[K/\mathbb{Q}, \mathfrak{q}] = c$. These are exactly the primes p so that $[K/\mathbb{Q}, p] = C$. Hence the term $e(\alpha n)$ appears in the above sum if and only if n is a prime p with $[K/\mathbb{Q}, p] = C$. We next need to compute the coefficient of this term. The coefficient will be $\mathrm{ord}(c)\log(p)$ times the number of primes $\mathfrak{p}$ of L over p with $[K/\mathbb{Q}, \mathfrak{p}] = c$. These primes are in 1-1 correspondence with primes $\mathfrak{q}$ of K over p with $[K/\mathbb{Q}, \mathfrak{q}] = c$. Now for such p, there will be $\frac{|G|}{\mathrm{ord}(c)}$ primes of K over it, and $\frac{|G|}{|C|\mathrm{ord}(c)}$ of them will have the correct Artin symbol. Hence the coefficient of $e(\alpha p)$ for such p will be exactly $\frac{|G|}{|C|}\log(p)$. Therefore the sum on the right hand side of Equation (3) is

$$\frac{|G|}{|C|} \sum_{\substack{p \leq X \\ [K/\mathbb{Q}, p] = C}} \log(p) e(\alpha p) + O(\sqrt{X}).$$

Multiplying by $\frac{|C|}{|G|}$ completes the proof of the Proposition. □

3.2 Local Approximations

Here we define some simpler functions meant to approximate F and G. In order to do so we will need a number of auxiliary definitions:

Definition. *For p a prime let*

$$\Lambda_p(n) = \begin{cases} 0 & \text{if } p|n \\ \frac{1}{1-p^{-1}} & \text{else} \end{cases}.$$

Λ_p can be thought of as a local approximation to the Von Mangoldt function, based only on the residue of n modulo p. Putting these functions together we get

Definition. *Let z be a positive real. Define a function Λ_z by*

$$\Lambda_z(n) = \prod_{p \leq z} \Lambda_p(n) = \begin{cases} 0 & \text{if } p|n \text{ for some prime } p < z \\ \prod_{p<z} \frac{1}{1-p^{-1}} & \text{otherwise} \end{cases}.$$

Note that by slight abuse of notation we have already defined several functions denoted by Λ with some subscript. We will disambiguate these by context and by consistently using subscripts either the same as or nearly identical to those used in the original definition (so Λ_z will always use z as its subscript, even though this represents a variable).

There are also some related definitions which will prove useful later.

Definition. *Let*

$$C(z) = \prod_{p \leq z} \frac{1}{1 - p^{-1}}.$$

$$P(z) = \prod_{p \leq z} p.$$

$$P(z, q) = \prod_{p \leq z, p \nmid q} p.$$

We note that

$$\Lambda_z(n) = C(z) \sum_{d | (n, P(z))} \mu(d),$$

that

$$\Lambda_z(n) = C(z) \sum_{d | (n, P(z,q))} \mu(d) \cdot \left(\begin{cases} 1 & \text{if } (n, q) = 1 \\ 0 & \text{otherwise} \end{cases} \right),$$

and that

$$C(z) = \Theta(\log(z)).$$

We will need some other local contributions to the Von Mangoldt function to take into account splitting information. In particular we define:

Definition. *Let $K/\mathbb{Q}$ be a Galois extension and $C \subset G = Gal(K/\mathbb{Q})$ a conjugacy class of the Galois group. Let the image of C in G^{ab} correspond via Global Class Field Theory to a coset H of some subgroup of $(\mathbb{Z}/D_K\mathbb{Z})^*$ for D_K the discriminant of K. We define $\Lambda_{K,C}$ to be the arithmetic function:*

$$\Lambda_{K,C}(n) = \begin{cases} \frac{\phi(D_K)}{|H|} & \text{if } n \in H \\ 0 & \text{otherwise} \end{cases}.$$

This accounts for the congruence conditions implied by n being a prime with Artin symbol C.

Definition. *Let L be a number field. Consider the image of $Gal(\bar{\mathbb{Q}}/L)$ in $Gal(\bar{Q}/\mathbb{Q})^{ab}$. By global class field theory, this corresponds to a subgroup H_L of $(\mathbb{Z}/D_L\mathbb{Z})^*$ for D_L the discriminant of L. Let*

$$\Lambda_{L/\mathbb{Q}}(n) = \begin{cases} \frac{\phi(D_L)}{|H_L|} & \text{if } n \in H_L \\ 0 & \text{otherwise} \end{cases}.$$

$\Lambda_{L/\mathbb{Q}}$ accounts for the congruence conditions that are implied by being a norm from L down to $\mathbb{Q}$.

We are now prepared to define our approximations $F^\sharp$ and $G^\sharp$ to F and G.

Definition. *For $K/\mathbb{Q}$ Galois, C a conjugacy class in $Gal(K/\mathbb{Q})$, and z and X positive numbers, we define the generating function*

$$G^{\sharp}_{K,C,X,z}(\alpha) = \frac{|C|}{|G|} \sum_{n \leq X} \Lambda_{K,C}(n)\Lambda_z(n)e(\alpha n).$$

We also let

$$G^{\flat}_{K,C,X,z}(\alpha) = G_{K,C,X}(\alpha) - G^{\sharp}_{K,C,X,z}(\alpha).$$

Definition. *For $L/\mathbb{Q}$ a number field, ξ a Grossencharacter of L, X and z positive numbers, we define the function*

$$F^{\sharp}_{L,\xi,X,z}(\alpha) = \begin{cases} \sum_{n \leq X} \Lambda_{L/\mathbb{Q}}(n)\Lambda_z(n)\chi(n)e(\alpha n) & \text{if } \xi = \chi \circ N_{L/\mathbb{Q}} \\ 0 & \text{otherwise} \end{cases}$$

where the first case is if ξ can be written as $\chi \circ N_{L/\mathbb{Q}}$ for some Dirichlet character χ. We note that although there may be several Dirichlet characters χ so that $\xi = N_{L/\mathbb{Q}} \circ \chi$, that the product of $\chi(n)$ with $\Lambda_{L/\mathbb{Q}}(n)$ is independent of the choice of such a χ. We also let

$$F^{\flat}_{L,\xi,X,z}(\alpha) = F_{L,\xi,X}(\alpha) - F^{\sharp}_{L,\xi,X,z}(\alpha).$$

Again for these functions we will often suppress some of the subscripts.

We claim that $F^{\sharp}$ and $G^{\sharp}$ are good approximations of F and G, and in particular we will prove that:

Theorem 4. *Let $K/\mathbb{Q}$ be a finite Galois extension, and let C be a conjugacy class of $Gal(K/\mathbb{Q})$. Let A be a positive number and B a sufficiently large multiple of A. Then if X is a positive number, $z = \log^B(X)$, and α any real number, then*

$$\left|G^{\flat}_{K,C,X,z}(\alpha)\right| = O\left(X \log^{-A}(X)\right), \tag{4}$$

where the implied constant depends on K, C, A, and B, but not on X or α.

Theorem 5. *Given $L/\mathbb{Q}$ a number field, and ξ a Grossencharacter of L, let A be a positive number and B a sufficiently large multiple of A. Then if X is a positive number, $z = \log^B(X)$, and α any real number, then*

$$\left|F^{\flat}_{L,\xi,X,z}(\alpha)\right| = O\left(X \log^{-A}(X)\right), \tag{5}$$

where the implied constant depends on L, ξ, A, and B, but not on X or α.

The proofs of these Theorems will be the bulk of Sections 4 and 5.

3.3 Smooth Numbers

We also need some results on the distribution of smooth numbers. We begin with a definition:

Definition. *Let $S(z,Y)$ be the number of $n \leq Y$ so that $n|P(z)$. In other words the number of $n \leq Y$ so that n is squarefree and has no prime factors bigger than z.*

We will need the following bound on $S(z,Y)$:

Lemma 6. *If $z = \log^B(X)$ and $Y \leq X$, then*

$$S(z,Y) \ll Y^{1-1/(2B)} \exp\left(O(\sqrt{\log(X)})\right)$$

Proof. Notice that

$$\int_{y=0}^{Y} S(z,y)dy = \frac{1}{2\pi}\int_{1-i\infty}^{1+i\infty} (s(s+1))^{-1} \prod_{p\leq z}(1+p^{-s})Y^{s+1}ds.$$

Note that for $\Re(s) > \frac{1}{2}$,

$$\left|\prod_{p\leq z}(1+p^{-s})\right| = \left|\exp\left(\sum_{p\leq z} p^{-s} + O(1)\right)\right| \ll \exp\left(\frac{z^{1-\Re(s)}}{1-\Re(s)}\right).$$

Changing the line of integration to $1-\Re(s) = \frac{1}{2B}$, we get that the integrand is at most $s^{-2}Y^{2-1/(2B)}\exp\left(O(\sqrt{\log(X)})\right)$. Integrating and evaluating at $2Y$, we get that

$$Y^{2-1/(2B)}\exp\left(O(\sqrt{\log(X)})\right) \gg \int_{y=0}^{2Y} S(z,y)dy \gg YS(z,Y),$$

proving our result. □

4 Approximation of F

In this Section, we will prove Theorem 5, restated here:

Theorem 5. *Given $L/\mathbb{Q}$ a number field, and ξ a Grossencharacter of L, let A be a positive number and B a sufficiently large multiple of A. Then if X is a positive number, $z = \log^B(X)$, and α any real number, then*

$$\left|F^{\flat}_{L,\xi,X,z}(\alpha)\right| = O\left(X\log^{-A}(X)\right),$$

where the implied constant depends on L, ξ, A, and B, but not on X or α.

In order to prove Theorem 5, we will split into cases based upon whether α is well approximated by a rational number of small denominator. If it is (the smooth case), we proceed to use the theory of L-functions to approximate F. If α is not well approximated (the rough case), we generalize results on exponential sums over primes to show that $|F|$ is small. In either case, $F^\sharp$ is not difficult to approximate.

We note that by Dirichlet's approximation Theorem, we can always find a pair (a, q) with a and q relatively prime and $q < M = \Theta(X \log^{-B}(X))$ with $\left|\alpha - \frac{a}{q}\right| \leq \frac{1}{qM}$. We consider the smooth case to be the one where $q \leq z$.

4.1 α Smooth

In this Section, we will prove the following Proposition:

Proposition 7. *Let L be a number field, and ξ a Grossencharacter. If $z = \log^B(X)$, $Y \leq X$ and $\alpha = \frac{a}{q}$ with a and q relatively prime and $q \leq z$, then for some constant $c > 0$ (depending only on L, ξ and B),*

$$|F^\flat_{L,\xi,Y,z}(\alpha)| = O\left(X \exp\left(-c\sqrt{\log(X)}\right)\right).$$

We note that this result can easily be extended to all smooth α. In particular we have:

Corollary 8. *Let L and ξ be as above. Let A be a constant, and B a sufficiently large multiple of A. Let $z = \log^B(X)$. Suppose that $\alpha = \frac{a}{q} + \theta$ with a and q relatively prime, $q \leq z$ and $|\theta| \leq \frac{1}{qM}$. Then*

$$|F^\flat_{L,\xi,X,z}(\alpha)| = O(X \log^{-A}(X)).$$

Proof (Given Proposition 7). Letting $F^\flat_{L,\xi,X}(\alpha) = \sum_{n \leq X} a_n e(\alpha n)$, we have by Abel summation and Proposition 7 that

$$\begin{aligned}
F^\flat_{L,\xi,X}(\alpha) &= \sum_{n \leq X} a_n e\left(\frac{na}{q}\right) e\left(n\theta\right) \\
&= (1 - e(\theta)) \left(\sum_{Y \leq X} F^\flat_{L,\xi,Y}\left(\frac{a}{q}\right) e(Y\theta)\right) + F^\flat_{L,\xi,X}\left(\frac{a}{q}\right) e((X+1)\theta) \\
&= X^{-1} \log^B(X) \sum_{Y \leq X} O\left(X \log^{-A-B}(X)\right) + O\left(X \log^{-A}(X)\right) \\
&= O\left(X \log^{-A}(X)\right).
\end{aligned}$$

□

In order to prove Proposition 7, we will need to separately approximate F and $F^\sharp$. For the former, we will also need to review some basic facts about Hecke L-functions.

4.1.1 Results on L-functions

Fix a number field L and a Grossencharacter ξ. We consider Hecke L-functions of the form $L(\xi\chi, s)$ where χ is a Dirichlet character of modulus $q \leq z = \log^B(X)$ thought of as a Grossencharacter via $\chi(\mathfrak{a}) = \chi(N_{L/\mathbb{Q}}(\mathfrak{a}))$. We let d be the degree of L over $\mathbb{Q}$, and let D_L be the discriminant. We let $\mathfrak{m}$ be the modulus of the character ξ, and q the modulus of χ. We note that $\xi\chi$ has modulus at most $q\mathfrak{m}$. Therefore by [24], in the paragraph above Theorem 5.35, $L(\xi\chi)$ has conductor $\mathfrak{q} \leq 4^d|d_K|N(\mathfrak{m})q^d$, and by Theorem 5.35 of [24], for some constant c depending only on L, $L(\xi\chi, s)$ has no zero in the region

$$\sigma > 1 - \frac{c}{d\log(|d_K|N(\mathfrak{m})q^d(|t|+3))}$$

except for possibly one Siegel zero. Note also that $L(\xi\chi, s)$ has a simple pole at $s = 1$ if $\xi = \bar{\chi}$, and otherwise is holomorphic. Noting that

$$\frac{-L'(\xi\chi, s)}{L(\xi\chi, s)} = \sum_{\mathfrak{a}} \Lambda_L(\mathfrak{a})\xi\chi(\mathfrak{a})N(\mathfrak{a})^{-s},$$

and that the n^{-s} coefficient of the above is at most $d\log(n)$, we may apply Theorem 5.13 of [24] and obtain for a suitable constant $c > 0$,

$$\sum_{N(\mathfrak{a})\leq Y} \Lambda_L(\mathfrak{a})\xi(\mathfrak{a})\chi(\mathfrak{a}) = \tag{6}$$
$$rY - \frac{Y^\beta}{\beta} + O\left(Y\exp\left(\frac{-c\log Y}{\sqrt{\log Y} + 3\log(q^d) + O(1)}\right)(\log(Yq^d) + O(1))^4\right),$$

where the term $\frac{Y^\beta}{\beta}$ should be taken with β the Siegel zero if it exists; $r = 0$ unless $\xi\chi = 1$, in which case, $r = 1$; and the implied constants may depend on L, ξ but not on χ or Y.

In order to make use of Equation 6, we will need to prove bounds on the size of Siegel zeroes. In particular we show that:

Lemma 9. *For all L and ξ, and all $\epsilon > 0$, there exists a $c(\epsilon) > 0$ so that for every Dirichlet character χ of modulus q and every Siegel zero β of $L(\xi\chi, s)$,*

$$\beta > 1 - \frac{c(\epsilon)}{q^\epsilon}.$$

Proof. We follow the proof of Theorem 5.28 part 2 from [24], and note the places where we differ. We note that Theorem 5.35 states that we only need by concerned when $\xi\chi$ is totally real. We then consider two such χ having Siegel zeros. We use, $L(s) = \zeta_L(s)L(\xi\chi_1, s)L(\xi\chi_2, s)L(\xi^2\chi_1\chi_2)$, which has conductor $O(q_1q_2)^{2d}$ instead of the analogous one from [24]. This gives us a convexity bound on the integral term of $O((q_1q_2)^d x^{1-\beta})$, instead of the one listed. Again assuming

that $\beta > 3/4$, we take $x > c(q_1q_2)^{4d}$. We notice that we still have (5.64) for $\sigma > 1 - 1/d$ sufficiently large by noting that $|\sum_{N(\mathfrak{a})\leq x} \xi\chi(\mathfrak{a})| = O(x^{1-1/d} + \max(x, q))$. Therefore, Equation (5.75) of [24] becomes

$$L(\xi\chi_2, 1) \gg (1 - \beta_1)(q_1q_2)^{-4d(1-\beta_1)}(\log(q_1q_2))^{-2}.$$

The rest of the argument from [24] carries over more or less directly. □

4.1.2 Approximation of F

We prove

Proposition 10. *With L, ξ, χ, Y, r as above, $X \geq Y$ and $z = \log^B(X)$,*

$$F_{L,\xi\chi,Y}(0) = rY + O\left(X\exp(-c\sqrt{X})\right). \tag{7}$$

Where again c depends on L, ξ but not χ, X, Y.

Proof. Applying Lemma 9 with $\epsilon = 1 - 1/(2B)$ to Equation 6, we get that

$$\begin{aligned}\sum_{N(\mathfrak{a})\leq Y} \Lambda_L(\mathfrak{a})\xi(\mathfrak{a})\chi(\mathfrak{a}) &= rY - \frac{Y^\beta}{\beta} + O\left(X\exp(-c\sqrt{\log(X)})\right)\\ &= rY + O\left(Y\exp(-c(\epsilon)\sqrt{\log(Y)}\right)\\ &\quad + O\left(X\exp(-c\sqrt{\log(X)})\right)\\ &= rY + O\left(X\exp(-c\sqrt{\log(X)})\right).\end{aligned}$$

□

4.1.3 Approximation of $F^\sharp$

Proposition 11. *With L, ξ, χ, X, Y, r as above, $z = \log^B(X)$,*

$$F^\sharp_{L,\xi\chi,Y}(0) = rY + O\left(X\exp(-c\sqrt{X})\right).$$

Proof. If $\xi\chi$ is not of the form $\chi' \circ N_{L/\mathbb{Q}}$, then $F^\sharp = 0$ and we are done. Otherwise let $\xi\chi$ be

as above with χ' a character of modulus q'. We have that

$$\begin{aligned} F^{\sharp}_{L,\chi',Y}(0) &= \sum_{n\leq Y} \Lambda_{L/\mathbb{Q}}(n)\Lambda_z(n)\chi'(n) \\ &= C(z) \sum_{n\leq Y} \sum_{d|(P(z,q'D_L),n)} \mu(d)\Lambda_{L/\mathbb{Q}}(n)\chi'(n) \\ &= C(z) \sum_{d|P(z,q'D_L)} \sum_{n=dm\leq Y} \mu(d)\Lambda_{L/\mathbb{Q}}(n)\chi'(n) \\ &= C(z) \sum_{d|P(z,q'D_L)} \mu(d)\chi'(d) \sum_{m\leq Y/d} \Lambda_{L/\mathbb{Q}}(dm)\chi'(m). \end{aligned}$$

Consider for a moment the inner sum over m. It is periodic with period dividing $q'D_L$. Note that the sum over a period is 0 unless χ' is trivial on H_L, in which case the average value is $\overline{\chi'}(d)\frac{\phi(q'D_L)}{q'D_L}$. Since $r = 1$ if χ' vanishes on H_L and $r = 0$ otherwise, we have that:

$$F^{\sharp}_{L,\chi',Y}(0) = C(z)\left(\frac{\phi(q'D_L)}{q'D_L}\right) \sum_{\substack{d|P(z,qD_L)\\ d\leq Y}} \left(\frac{r\mu(d)Y}{d} + O(q'D_L)\right).$$

The sum of error term here is at most $O\left(C(z)qS(z,Y)\right)$ which by Lemma 6 is $O\left(Y^{1-1/(2B)}\log^2(z)q\exp(O(\sqrt{\right.}$
The remaining term is

$$rYC(z)\frac{\phi(q'D_L)}{q'D_L} \sum_{\substack{d|P(z,q'D_L)\\ d\leq Y}} \frac{\mu(d)}{d}.$$

The error introduced by extending the sum to all $d|P(z,q'D_L)$ is at most

$$O\left(YC(z)\int_Y^{\infty} S(z,y)y^{-2}dy\right).$$

By Lemma 6 this is

$$O\left(Y^{1-1/(2B)}\log(z)\exp(O(\sqrt{\log(X)}))\right).$$

Once we have extended the sum we are left with

$$\begin{aligned} rYC(z)\frac{\phi(q'D_L)}{q'D_L} \sum_{d|P(z,q'D_L)} \frac{\mu(d)}{d} &= rYC(z)\left(\frac{\phi(q'D_L)}{q'D_L}\right)\left(\frac{\phi(P(z,q'D_L))}{P(z,q'D_L)}\right) \\ &= rYC(z)\left(\frac{\phi(P(z))}{P(z)}\right) \\ &= rY. \end{aligned}$$

Hence

$$F^{\sharp}_{L,\xi_\chi,Y,z}(0) = rY + O\left(Y^{1-1/(2B)}\log^2(z)q^2\exp(O(\sqrt{\log(X)}))\right)$$
$$= rY + O\left(X\exp(-c\sqrt{X})\right).$$

□

4.1.4 Proof of Proposition 7

Proof. Combining Propositions 10 and 11 we obtain that

$$F^{\flat}_{L,\xi_\chi,Y,z}(0) = O\left(X\exp(-c\sqrt{X})\right).$$

Our Proposition follows immediately after noting that

$$F^{\flat}_{L,\xi,X,z}\left(\frac{a}{q}\right) = \sum_{\chi \bmod q} e_\chi F^{\flat}_{L,\xi_\chi,X,z}(0).$$

Where e_χ is the appropriate Gauss sum. □

4.2 α Rough

In this Section, we will show that $|F^{\flat}(\alpha)|$ is small for α not well approximated by a rational of small denominator. We will do this by showing that both $|F(\alpha)|$ and $|F^{\sharp}(\alpha)|$ are small. The proof of the latter will resemble the proof of Proposition 11. The proof of the former will require some machinery including some Lemmas about rational approximations and exponential sums of polynomials.

4.2.1 Bounds on $F^{\sharp}$

Proposition 12. *Fix L a number field, and ξ a Grossencharacter. Fix B and let $z = \log^B(X)$. Let α be a real number. If there exist relatively prime integers a and q so that*

$$\left|\alpha - \frac{a}{q}\right| < \frac{1}{q^2},$$

then, $|F^{\sharp}_{L,\xi,z}(\alpha)|$ is

$$O\left(X\log(X)\log(z)q^{-1} + q\log(q)\log(z) + X^{1-1/(4B)}\exp(O(\sqrt{\log(X)})\right).$$

Proof. We note that the result is trivial unless $\xi = N_{L/\mathbb{Q}}(\chi)$ for some Dirichlet character χ of modulus Q. Hence we may assume that

$$F^{\sharp}_{L,\xi,z}(\alpha) = \sum_{n \leq X} \Lambda_{L/\mathbb{Q}}(n)\Lambda_z(n)\chi(n)e(\alpha n).$$

Let D_L be the discriminant of L. We note that

$$\begin{aligned} F^{\sharp}_{L,\xi,z}(\alpha) &= \sum_{n \leq X} \Lambda_{L/\mathbb{Q}}(n)\Lambda_z(n)\chi(n)e(\alpha n) \\ &= C(z) \sum_{n \leq X} \sum_{d|(n,P(z,QD_L))} \mu(d)\Lambda_{L/\mathbb{Q}}(n)\chi(n)e(\alpha n) \\ &= C(z) \sum_{d|P(z,QD_L)} \mu(d)\chi(d) \sum_{md=n\leq X} \Lambda_{L/\mathbb{Q}}(dm)\chi(m)e(\alpha dm) \\ &= O\left(C(z) \sum_{d|P(z,QD_L)} \left| \sum_{m \leq X/d} \Lambda_{L/\mathbb{Q}}(dm)\chi(m)e(\alpha dm) \right| \right). \end{aligned}$$

In order to analyze the last sum, we split it up based on the conjugacy class of m modulo QD_L. Each new sum is geometric series with ratio of terms $e(\alpha QD_L d)$. Hence we can bound this sum as $\min\left(\frac{X}{d}, \frac{QD_L}{2||dQD_L\alpha||}\right)$, where $||x||$ is the distance from x to the nearest integer. Therefore we have that $|F^{\sharp}_{\chi,z}(\alpha)|$ is

$$O\left(C(z) \sum_{d \leq X^{1-1/(4B)}} \min\left(\frac{X}{d}, \frac{QD_L}{2||dQD_L\alpha||}\right) + C(z)X^{1/4B}S(z,X) \right).$$

We bound the sum in the first term by looking at what happens as d ranges over an interval of length $\frac{q}{3QD_L}$. We get that $dQD_L\alpha = x_0 + kQ\alpha$ for x_0 the value at the beginning of the interval and k an integer at most $\frac{q}{3QD_L}$. Notice that $kQD_L\alpha$ is within $\frac{1}{3q}$ of $\frac{kQD_La}{q}$, which must be distinct for different values of k. Hence none of the fractional parts of $dQD_L\alpha$ can be within $\frac{1}{3q}$ of each other. Hence the sum over this range of d is at most $\frac{X}{d} + \frac{2QD_L}{2/(3q)} + \frac{2QD_L}{2(2/2q)} + \ldots = O\left(\frac{X}{d} + 3qQD_L\log(q)\right)$. Furthermore the $\frac{X}{d}$ term does not show up in the first such interval, since when $d = 0$, $dQD_L\alpha$ is an integer. We have $3QD_LX^{1-1/(4B)}/q+1$ of these intervals. Therefore, the first term is at most

$$\begin{aligned} &O\left(\sum_{\ell=1}^{3QD_LX^{1-1/(4B)}} \frac{X}{(\ell q/(3QD_L))} + 9Q^2D_L^2\log(q)X^{1-1/(4B)} + 3qQD_L\log(q) \right) \\ &= O\left(X\log(X)q^{-1} + \log(q)X^{1-1/(4B)} + q\log(q) \right). \end{aligned}$$

The other term is bounded by Lemma 6 as

$$O\left(\log(z)X^{1-1/(4B)}\exp\left(O(\sqrt{\log(X)})\right)\right).$$

Putting these bounds together, we get that $|F^{\sharp}_{L,\xi,z}(\alpha)|$ is at most

$$O\left(X\log(X)\log(z)q^{-1}+q\log(q)\log(z)+X^{1-1/(4B)}\exp(O(\sqrt{\log(X)})\right).$$

□

4.2.2 Lemmas on Rational Approximation

In the coming Sections, we will need some results on rational approximation of numbers. In particular, we will need to know how often multiples of a given α have a good rational approximation. In order to discuss these issues, we first make the following definition:

Definition. *We say that a real number α has a* rational approximation with denominator *q if there exist relatively prime integers a and q so that*

$$\left|\alpha-\frac{a}{q}\right|<\frac{1}{q^2}.$$

We now prove a couple of Lemmas about this definition.

Lemma 13. *Let X, Y, A be positive integers. Let α be a real number with rational approximation of denominator q. Suppose that for some B, that $XYB^{-1} > q > B$. Then for all but*

$$O\left(Y\left(A^{3/2}B^{-1/2}+A^2B^{-1}+\log(AY)A^3X^{-1}\right)\right)$$

of the integers n with $1 \le n \le Y$, $n\alpha$ has a rational approximation with denominator q' for some $XA^{-1} > q' > A$.

Proof. By Dirichlet's approximation theorem, $n\alpha$ always has a rational approximation $\frac{a}{q'}$ with $q' < XA^{-1}$ and

$$\left|n\alpha-\frac{a}{q'}\right|<\frac{1}{q'XA^{-1}}.$$

Therefore, $n\alpha$ lacks an appropriate rational approximation only when the above has a solution for some $q' \le A$. If such is the case, then, dividing by n, we find that α is within $(q')^{-1}n^{-1}X^{-1}A$ of some rational number of denominator d so that $d|nq'$. Note that this error is at most $d^{-1}X^{-1}A$.

Given such a rational approximation to α with denominator d, we claim that it contributes to at most YA^2d^{-1} bad n's. This is because there are at most A values of q', and for each

value of q', we still need that n is a multiple of $\frac{d}{(d,q')} \geq dA^{-1}$. Hence for each q', there are at most YA^2d^{-1} bad n.

Next, we pick an integer n_0. We will now consider only $Y \geq n \geq n_0$ so that αn has no suitable rational approximation. We do this by analyzing the denominators d for which some rational number of denominator d approximates α to within $X^{-1}A(\max(d, n_0))^{-1}$. Suppose that we have some $d \neq q$ which does this. α is within q^{-2} of a number with denominator q, and within $X^{-1}n_0^{-1}A$ of one with denominator d. These two rational numbers differ by at least $(dq)^{-1}$ and therefore

$$(dq)^{-1} \leq q^{-2} + X^{-1}An_0^{-1}.$$

Hence either dq^{-1} or $X^{-1}An_0^{-1}dq$ is at least $\frac{1}{2}$. Hence either $d \geq \frac{q}{2}$, or

$$d \geq \frac{Xn_0}{2Aq} \geq \frac{n_0B}{2AY}.$$

Therefore the smallest such d is at least the minimum of $\frac{q}{2}$ and $\frac{n_0B}{2AY}$.

Next suppose that we have two different such denominators, say d and d'. The fractions they represent are separated by at least $(dd')^{-1}$ and yet are both close to α. Therefore

$$(dd')^{-1} \leq X^{-1}A(d^{-1} + d'^{-1}).$$

Therefore we have that $\max(d, d') \geq \frac{X}{2A}$. Hence there is at most one such denominator less than $\frac{X}{2A}$.

Next we wish to bound the number of such denominators d in a dyadic interval $[K, 2K]$. We note that the corresponding fractions are all within $X^{-1}AK^{-1}$ of α, and that any two are separated from each other by at least $(2K)^{-2}$. Therefore the number of such d is at most $1 + 8KX^{-1}A$.

To summarize we potentially have the following d each giving at most YA^2d^{-1} bad n's.

- One d at least $\min\left(\frac{q}{2}, \frac{n_0B}{2AY}\right)$.
- For each diadic interval $[K, 2K]$ with $K \geq \frac{X}{2A}$ at most $10KX^{-1}A$ such d's

Notice that there are $\log(2AY)$ such diadic intervals, and that each contributes at most $10YA^3X^{-1}$ bad n's. We also potentially have n_0 bad n's from the numbers less than n_0. Hence the number of n for which there is no suitable rational approximation of $n\alpha$ is at most

$$O\left(n_0 + YA^2B^{-1} + Y^2A^3B^{-1}n_0^{-1} + \log(AY)YA^3X^{-1}\right).$$

Substituting $n_0 = YA^{3/2}B^{-1/2}$ yields our result. □

We will also need the following related Lemma:

Lemma 14. *Let X, A, C be positive integers. Let α be a real number with rational approximation of denominator q. Suppose that for some $B > 2A$, that $XB^{-1} > q > B$. Then there exists a set S of natural numbers so that*

- *elements of S are of size at least $\Omega(BA^{-1})$.*
- *The sum of the reciprocals of the elements of S is $O(A^2B^{-1} + X^{-1}A^4C)$.*
- *for all positive integers $n \leq C$, either n is a multiple of some element of S or $n\alpha$ has a rational approximation with some denominator q' with $XA^{-1}n^{-1} > q' > A$.*

Proof. We use the same basic techniques as the proof of Lemma 13. We note that $n\alpha$ always has a rational approximation $\frac{a}{q'}$ accurate to within $\frac{1}{q'XA^{-1}n^{-1}}$ with $q' < XA^{-1}n^{-1}$. This means that we have an appropriate rational approximation of $n\alpha$ unless this q' is less than A. If this happens, it is the case that

$$\left|\alpha - \frac{a}{nq'}\right| \leq \frac{1}{q'XA^{-1}}.$$

Hence to each such n we can assign a rational approximation $\frac{a}{nq'}$ of α. The n that are assigned to a rational approximation $\frac{a}{d}$ are those so that

$$\left|\alpha - \frac{a}{d}\right| \leq \frac{1}{XA^{-1}D}$$

where $D = \frac{d}{(n,d)} \leq A$. Hence it suffices to let S be the set of all $\frac{d}{D}$ where $A \geq D$, $D|d$ and

$$\left|\alpha - \frac{a}{d}\right| \leq \frac{1}{XA^{-1}D} \leq \frac{1}{XA^{-1}}.$$

We have to show that the elements of S are big enough and that the sum of their reciprocals satisfies the appropriate bound. Note that for each such d, it contributes at most $O\left(\frac{A^2}{d}\right)$ to the sum of reciprocals.

We note that if we have any such denominator d other than q, α is within q^{-2} of a rational number of denominator q and within $X^{-1}A$ of one of denominator d. Hence we have that

$$(dq)^{-1} \leq q^{-2} + X^{-1}A.$$

Hence

$$d \geq \min\left(\frac{q}{2}, \frac{X}{2A}\right) \geq \frac{q}{2}.$$

Note that in any case $\frac{d}{A} = \Omega(BA^{-1})$, and hence all of the terms in S are sufficiently large. Additionally, this s contributes at most $O(A^2B^{-1})$ to the sum of reciprocals.

Next note that if we have two of these approximations with denominators d and d' that

$$(dd')^{-1} \leq 2XA^{-1}.$$

Therefore the second largest such d is at least $\sqrt{2XA^{-1}}$.

Next we consider the contribution from all such approximations with d lying in a diadic interval $[K, 2K]$ all of these approximations are within $X^{-1}A$ of α and are separated from each other by at least $\frac{1}{4K^2}$. Therefore, there are at most $1+8X^{-1}AK^2$. If we ensure that K is at least $\sqrt{2XA^{-1}}$, this is $O(X^{-1}AK^2)$. Hence all of these d's contribute at most $O(X^{-1}A^3K)$ to the sum of reciprocals. Furthermore we can ignore terms with $K > AC$. Taking the sum of this over K a power of 2, we get at most $O(X^{-1}A^4C)$.

□

We will be using Lemma 14 to bound the number of ideals of L so that $N(\mathfrak{a})\alpha$ has a good rational approximation. In order to do this we will also need the following:

Lemma 15. *Fix L be a number field. Let n be a positive integer, and let X and ϵ be positive real numbers. Then we have that:*

$$\sum_{\substack{n|N(\mathfrak{a}) \\ N(\mathfrak{a})<X}} \frac{1}{N(\mathfrak{a})} = O\left(\frac{X\log(X)n^\epsilon}{n}\right),$$

$$\sum_{\substack{n|N(\mathfrak{ab}) \\ N(\mathfrak{ab})<X}} \frac{1}{N(\mathfrak{ab})} = O\left(\frac{X\log^2(X)n^\epsilon}{n}\right).$$

(The first sum is over ideals $\mathfrak{a}$ so that $n|N(\mathfrak{a})$ and $N(\mathfrak{a}) \leq X$, the second over pairs of ideals $\mathfrak{a}$ and $\mathfrak{b}$, so that $N(\mathfrak{a}\cdot\mathfrak{b})$ satisfies the same conditions).

Proof. We will prove the first of the two equations and note that the second follows from a similar argument. Let $d = [L : \mathbb{Q}]$. Let $p_1, \ldots, p_k$ be the distinct primes dividing n. We claim that for such an ideal $\mathfrak{a}$ must be a multiple of some ideal $\mathfrak{a}_0$ with $N(\mathfrak{a}_0) = nm$ with $m = \prod_{i=1}^k p_i^{a_i}$ for some $0 \leq a_i < d$. We obtain this by starting with the ideal $\mathfrak{a}_0 = (1)$ and repeatedly multiplying by primes of $\mathfrak{a}/\mathfrak{a}_0$ whose norm is a power of one of the p_i that do not yet divide $N(\mathfrak{a}_0)$ sufficiently many times. Since this prime has norm no bigger than p_i^d we cannot overshoot by more than $d-1$ factors of any p_i. We note that the number of possible values of m is k^d. Since $k = O(\log(n))$ this is $O(n^\epsilon)$. For each value of m there are $O(n^\epsilon)$ ideals of norm exactly nm, and hence there are $O(n^\epsilon)$ possible ideals $\mathfrak{a}_0$.

We now need to bound the sum over ideals $\mathfrak{b}$ so that the norm of $\mathfrak{a}_0\mathfrak{b}$ is at most X of $\frac{1}{N(\mathfrak{a}_0\mathfrak{b})}$. This is at most $\frac{1}{n}$ times the sum over ideals $\mathfrak{b}$ of norm at most X of $\frac{1}{N(\mathfrak{b})}$. This latter sum is $O(\log(X))$. This completes the proof. □

4.2.3 Lemmas on Exponential Sums

We will need a Lemma on the size of exponential sums of polynomials along the lines of Lemma 20.3 of [24]. Unfortunately, the X^ϵ term that shows up there will be unacceptable for our application. So instead we prove:

Lemma 16. *Pick a positive integer X. Let $[X] = \{1, 2, \ldots, X\}$. Let P be a polynomial with leading term cx^k for some integer $c \neq 0$. Let α be a real number with a rational approximation of denominator q. Then*

$$\left| \sum_{x \in [X]} e(\alpha P(x)) \right| \ll |c| X \left(\frac{1}{q} + \frac{1}{X} + \frac{q}{X^k} \right)^{10^{-k}},$$

where the implied constant depends on k, but not on the coefficients of P.

Note that the 10^{-k} in the exponent is not optimal and was picked for convenience.

Proof. We proceed by induction on k. We take as a base case $k = 1$. Then we have that P is a linear function with linear term c. α is within q^{-2} of a rational number of denominator q. Therefore $c\alpha$ is within cq^{-2} of a number of denominator between qc^{-1} and q. If $c \geq q/2$, there is nothing to prove. Otherwise, $c\alpha$ cannot be within $q^{-1} - cq^{-2} = O(q^{-1})$ of an integer. Therefore the sum is at most $O(\min(X, q))$, which clearly satisfies the desired inequality.

We now perform the induction step. We assume our inequality holds for polynomials of smaller degree. Squaring the left hand side of our inequality, we find that

$$\left| \sum_{x \in [X]} e(\alpha P(x)) \right| = \left(\sum_{a,b \in [X]} e(\alpha(P(a) - P(b))) \right)^{1/2}.$$

Breaking the inner sum up based on the value of $n = a - b$, we note that $P(n + b) - P(b)$ is a polynomial of degree $k - 1$ with leading term $nckx^{k-1}$. Letting $[X_n]$ be the interval of length $X - |n|$ that b could be in given that $b \in [X]$ and $b + n \in [X]$, we are left with at most

$$\left(\sum_{n \in [-X, X]} \left| \sum_{b \in [X_n]} e(\alpha(P(b + n) - P(b))) \right| \right)^{1/2}$$
$$= \left(\sum_{n \in [-X, X]} \left| \sum_{b \in [X_n]} e\left((n\alpha) \left(\frac{P(b + n) - P(b)}{n} \right) \right) \right| \right)^{1/2}.$$

Let $B = \min(q, X^k/q)$. We consider separately the terms in the above sum where $n\alpha$ has no rational approximation with denominator between $B^{1/5}$ and $X^{k-1}B^{-1/5}$. By Lemma 13 with parameters $A = B^{1/5}, B = B, Y = X, X = X^{k-1}$, the number of such n is at most

$$O(X(B^{-1/5} + \log(X)B^{3/5}X^{1-k})).$$

Each of those terms contributes $O(X)$ to the sum and hence together they contribute at most

$$O(X(B^{-1/10} + \log(X)B^{3/10}X^{(1-k)/2})).$$

Which is within the required bounds.

For the other terms, the inductive hypothesis tells us that the sum for fixed n is at most

$$O\left(|c|X\left(B^{-1/5} + \frac{1}{X-|n|} + \frac{X^{k-1}B^{-1/5}}{(X-|n|)^{k-1}}\right)^{-10^{k-1}}\right).$$

Summing over n and taking a square root gives an appropriate bound. □

We apply this Lemma to get a bound on exponential sums of norms of ideals of a number field. In particular we show that:

Lemma 17. *Fix L a number field of degree d, and ξ a Grossencharacter of modulus $\mathfrak{m}$. Then given a positive number X and a real number α which has a rational approximation of denominator q, we have that*

$$\left|\sum_{N(\mathfrak{a})\leq X} \xi(\mathfrak{a})e(\alpha N(\mathfrak{a}))\right| = O\left(X\left(\frac{1}{q} + \frac{1}{X^{1/d}} + \frac{q}{X}\right)^{10^{-d}/2}\right).$$

Where the implied constant depends only on L and ξ.

Proof. First we split the sum up into ideal classes modulo $\mathfrak{m}$. In order to represent an element of such a class we note that for $\mathfrak{a}_0$ a fixed element of such a class then other elements $\mathfrak{a}$ in the same class correspond to points $\frac{\mathfrak{a}}{\mathfrak{a}_0}$ in some lattice in $L \otimes \mathbb{R}$. Note that points in this lattice overcount these ideals since if two differ by an element of O_L^*, they correspond to the same ideal. On the other hand, if we take some fundamental domain of the elements of unit norm in $L \otimes \mathbb{R}$ modulo the elements of O_L^* that are 1 modulo $\mathfrak{m}$, and consider its positive multiples, we get a smoothly bounded region, R so that ideals in this class correspond to lattice points in R. Notice that the norm is a polynomial form of degree d on R. By taking the intersection of R with the set of points of norm at most X we get a region R_X whose lattice points correspond exactly to the ideals in this class of norm at most X. Let $Y = (qX)^{1/2d}$. We attempt to partition these lattice points into segments of length Y with a particular orientation. The number that we fail to include is proportional to Y times the surface area

of R_X which is $O(YX^{1-1/d})$. On each segment, we attempt to approximate ξ by a constant. We note that on this region ξ is a smoothly varying function of $x/(N(x))^{1/d}$. Therefore the error introduced at each point, x, is $O(YN(x)^{-1/d})$. The sum over points in R_X of $N(x)^{-1/d}$ is, by Abel Summation, $O(\int_{x=0}^{X} x^{-1/d}dx) = O(X^{1-1/d})$, thus producing another error of size at most $O(YX^{1-1/d})$. We are left with a sum over $O(XY^{-1})$ segments of length Y of the exponential sum of $e(\alpha N(x))$. Recalling that $N(x)$ is a polynomial of degree d with rational leading coefficient with bounded numerator and denominator, we may (perhaps after looking at only every k^{th} point to make the leading coefficient integral) apply Lemma 16 and get that the sum over each segment is

$$O\left(Y\left(\frac{1}{q}+\frac{1}{Y}+\frac{q}{Y^d}\right)^{10^{-d}}\right).$$

Noting that each of the error terms we introduced is less than the bound given, we are done. □

Abel summation yields the following Corollary.

Corollary 18. *Fix L a number field of degree d, and ξ a Grossencharacter of modulus $\mathfrak{m}$. Then given a positive number X and a real number α which has a rational approximation of denominator q, we have that*

$$\left|\sum_{N(\mathfrak{a})\leq X}\log(N(\mathfrak{a}))\xi(\mathfrak{a})e(\alpha N(\mathfrak{a}))\right| = O\left(X\log(X)\left(\frac{1}{q}+\frac{1}{X^{1/d}}+\frac{q}{X}\right)^{10^{-d}/2}\right).$$

4.2.4 Bounds on F

We are finally ready to prove our bound on F.

Proposition 19. *Fix a number field L of degree d and a Grossencharacter ξ. Let $X \geq 0$ be a real number. Let α be a real number with a rational approximation of denominator q where $XB^{-1} > q > B$ for some $B > 0$. Then $F_{L,\xi,X}(\alpha)$ is*

$$O\left(X\log^2(X)\left(B^{-10^{-d}/12}+X^{-10^{-d}/60}+X^{-10^{-d}/10d}+\log^{d^2/2}(X)B^{-1/12}\right)\right).$$

Where the asymptotic constant may depend on L and ξ, but not on X, q, B or α.

Note that a bound for $F_{L,\xi,X}(\alpha)$ is already known for the case when $L/\mathbb{Q}$ is abelian. In [2], they prove bounds on exponential sums over primes in an arithmetic progression. By Class Field Theory, this is clearly equivalent to proving bounds on F (or more precisely, G) when L is abelian over $\mathbb{Q}$. Proposition 19 can be thought of as a generalization of this result.

Proof. Our proof is along the same lines as Theorem 13.6 of [24]. We first note that the suitable generalization of Equation (13.39) of [24] still applies. Letting $y = z = X^{2/5}$ (y and z are variables used in (13.39) of [24]), we find that $F_{L,\xi,X}(\alpha)$ equals

$$\begin{aligned}
&\sum_{\substack{N(\mathfrak{ab})\leq X\\ N(\mathfrak{a})<X^{2/5}}} \mu(\mathfrak{a})\xi(\mathfrak{a})\log(N(\mathfrak{b}))\xi(b)e(\alpha N(\mathfrak{a})N(\mathfrak{b}))\\
&- \sum_{\substack{N(\mathfrak{abc})\leq X\\ N(\mathfrak{b}),N(\mathfrak{c})\leq X^{2/5}}} \mu(\mathfrak{b})\Lambda_L(\mathfrak{c})\xi(\mathfrak{bc})\xi(\mathfrak{a})e(\alpha N(\mathfrak{bc})N(\mathfrak{a}))\\
&+ \sum_{\substack{N(\mathfrak{abc})\leq X\\ N(\mathfrak{b}),N(\mathfrak{c})\geq X^{2/5}}} \mu(\mathfrak{b})\xi(\mathfrak{b})\Lambda_L(\mathfrak{c})\xi(\mathfrak{ac})e(\alpha N(\mathfrak{b})N(\mathfrak{ac})) + O(X^{2/5}).
\end{aligned}$$

The first term, we bound using Corollary 18 on the sum over $\mathfrak{b}$. Let $A = B^{1/4} \leq X^{1/8}$. By Lemmas 14 and 15, we can bound the sum over terms where $\alpha N(\mathfrak{a})$ has no rational approximation with denominator between A and $\frac{X}{AN(\mathfrak{a})}$ by

$$O\left(X\left(\log^2(X)\left(A^3B^{-1} + X^{-3/5}A^4\right)\left(\frac{B}{A}\right)^{\epsilon}\right)\right) = O\left(X\log^2(X)B^{-1/4+\epsilon}\right).$$

For other values of $\mathfrak{b}$, Corollary 18 bounds the sum as

$$O\left(X\log^2(X)\left(B^{-1/4} + X^{-3/5d}\right)^{10^{-d}/2}\right).$$

The second term is bounded using similar considerations. We let $A = \min(B^{1/4}, X^{1/41})$, and use Lemmas 14 and 15 to bound the sum over terms with $\mathfrak{b}$ and $\mathfrak{c}$ such that $N(\mathfrak{bc})\alpha$ has no rational approximation with norm between A and $\frac{X}{AN(\mathfrak{bc})}$ by

$$\begin{aligned}
&O\left(X\log^3(X)\left(\frac{B}{A}\right)^{\epsilon}\left(B^{-1/4} + X^{-1}A^4X^{4/5}\right)\right)\\
&\qquad = O\left(X\log^3(X)\left(B^{-1/4+\epsilon} + X^{-1/10}\right)\right).
\end{aligned}$$

Using Lemma 17, we bound the sum over other values of $\mathfrak{b}$ and $\mathfrak{c}$ as

$$O\left(X\log^2(X)\left(A^{-1} + X^{-1/5d}\right)^{10^{-d}/2}\right).$$

To bound the last sum, we first change to a sum over $\mathfrak{b}$ and $\mathfrak{d} = \mathfrak{a}\cdot\mathfrak{c}$. We have coefficients

$$x(\mathfrak{b}) = \mu(\mathfrak{b})\xi(\mathfrak{b}),$$

and

$$y(\mathfrak{d}) = \sum_{\substack{\mathfrak{a}\cdot\mathfrak{c}=\mathfrak{d} \\ N(\mathfrak{c})\geq X^{2/5}}} \Lambda_L(\mathfrak{c})\xi(\mathfrak{a}\mathfrak{c}).$$

We note that $|y(\mathfrak{d})| \leq \log(N(\mathfrak{d})) \leq \log(X)$. Our third term then becomes

$$\sum_{\substack{N(\mathfrak{b}\mathfrak{d})\leq X \\ N(\mathfrak{b}),N(\mathfrak{d})\geq X^{2/5}}} x(\mathfrak{b})y(\mathfrak{d})e(\alpha N(\mathfrak{b})N(\mathfrak{d})).$$

To this we apply the bilinear form trick. First, we split the sum over $\mathfrak{b}$ into parts based on which diadic interval (of the form $[K, 2K]$), the norm of $\mathfrak{b}$ lies in. Next, for each of these summands, we apply Cauchy-Schwartz to bound it by

$$\left(\left(\sum_{N(\mathfrak{b})\in[K,2K]} |x(\mathfrak{b})|^2\right)\left(\sum_{N(\mathfrak{b})\in[K,2K]}\left(\sum_{\substack{N(\mathfrak{d})\leq X/N(\mathfrak{b}) \\ N(mfd)\geq X^{2/5}}} y(\mathfrak{d})e(\alpha N(\mathfrak{b}\mathfrak{d}))\right)^2\right)\right)^{1/2}$$

$$= O(\sqrt{K})\left(\sum_{\substack{N(\mathfrak{b})\in[K,2K] \\ N(\mathfrak{d}),N(\mathfrak{d}')\leq X/N(\mathfrak{b}) \\ N(\mathfrak{d}),N(\mathfrak{d}')\geq X^{2/5}}} x(\mathfrak{d})\overline{x(\mathfrak{d}')}e(\alpha N(\mathfrak{b})(N(\mathfrak{d})-N(\mathfrak{d}')))\right)^{1/2}$$

$$\leq O(\sqrt{K}\log(X))\left(\sum_{\substack{X^{2/5}\leq N(\mathfrak{d}) \\ X^{2/5}\leq N(\mathfrak{d}')}}\left|\sum_{\substack{N(\mathfrak{b})\in[K,2K] \\ N(\mathfrak{b})\leq X/N(\mathfrak{d}) \\ N(\mathfrak{b})\leq X/N(\mathfrak{d}')}} e(\alpha(N(\mathfrak{d})-N(\mathfrak{d}'))N(\mathfrak{b}))\right|\right)^{1/2}.$$

We let $A = \min(B^{1/6}, X^{1/16})$. We bound terms separately based on whether or not $\alpha(N(\mathfrak{d}) - N(\mathfrak{d}'))$ has a rational approximation with denominator between A and KA^{-1}. Applying Lemma 13 with $X = K$, $Y = XK^{-1}$, $A = A$ and $B = B$, we get that the number of values of $N(\mathfrak{d}) - N(\mathfrak{d}')$ that cause this to happen is

$$O\left(XK^{-1}\left(B^{-1/6} + X^{3/32}B^{-1/2} + \log(X)A^4K^{-1}\right)\right) = O(XK^{-1}B^{-1/6}).$$

For each such difference, by the rearrangement inequality the number of $\mathfrak{d}, \mathfrak{d}'$ with norms at most $X/2K$ with $N(\mathfrak{d}) - N(\mathfrak{d}')$ equal to this difference is at most

$$\sum_{n=1}^{X/2K} (\text{Number of ideals with norm } n)^2.$$

Letting $W(n)$ be the number of ideals of L with norm n, we have that $W^2(n)$ is at most $\tau_d^2(n)$, where $\tau_d(n)$ is the number of ways of writing n as a product of d integers. This is because W is multiplicative and if we write a power of a prime p as the norm of an ideal, $\mathfrak{a}$, factoring $\mathfrak{a}$ into its primary parts gives a representation of n as a product of d integers. We therefore have that $W^2(n) \leq \tau_{d^2}(n)$ and hence the above sum is $O(XK^{-1}\log^{d^2}(X))$. Hence the total contribution from terms with such $\mathfrak{d}$ and $\mathfrak{d}'$ is at most

$$O\left(\left(K^{1/2}\log(X)\right)\left(K\left(XK^{-1}B^{-1/6}\right)\left(XK^{-1}\log^{d^2}(X)\right)\right)^{1/2}\right)$$
$$= O\left(X\log^{1+d^2/2}(X)B^{-1/12}\right).$$

The sum over the $O(\log(X))$ possible values for K of the above is

$$O\left(X\log^{2+d^2/2}(X)B^{-1/12}\right).$$

On the other hand, the sum over $\mathfrak{d}$ and $\mathfrak{d}'$ so that $\alpha(N(\mathfrak{d}) - N(\mathfrak{d}'))$ has a rational approximation with appropriate denominator is bounded by Lemma 17 by

$$O\left(\left(K^{1/2}\log(X)\right)\left(\left(XK^{-1}\right)^2 K\left(A^{-1}+K^{-1/d}\right)^{10^{-d}/2}\right)^{1/2}\right)$$
$$= O\left(X\log(X)\left(A^{-1}+K^{-1/d}\right)^{10^{-d}/4}\right).$$

Summing over all of the intervals we get

$$O\left(X\log^2(X)\left(A^{-1}+X^{-2/5d}\right)^{10^{-d}/4}\right).$$

Putting this all together, we get the desired bound for F. □

4.3 Putting it Together

We are finally prepared to prove Theorem 5.

Proof. We note that α can always be approximated by $\frac{a}{q}$ for some relatively prime integers a, q with $q \leq X\log^{-B}(X)$ so that

$$\left|\alpha - \frac{a}{q}\right| \leq \frac{1}{qX\log^{-B}(X)}.$$

We split into cases based upon weather $q \leq z$.

If $q \leq z$ our result follows from Corollary 8.

If $q \geq z$, our result follows from Propositions 12 and 19. □

5 Approximation of G

In this Section, we prove Theorem 4. We restate it here:

Theorem 4. *Let $K/\mathbb{Q}$ be a finite Galois extension, and let C be a conjugacy class of $Gal(K/\mathbb{Q})$. Let A be a positive integer and B a sufficiently large multiple of A. Then if X is a positive number, $z = \log^B(X)$, and α any real number, then*

$$\left|G^{\flat}_{K,C,X,z}(\alpha)\right| = O\left(X \log^{-A}(X)\right),$$

where the implied constant depends on K, C, A, B, but not on X or α.

Proof. Recall Proposition 3 which states that

$$G_{K,C,X}(\alpha) = \frac{|C|}{|G|}\left(\sum_{\chi} \overline{\chi}(c) F_{L,\chi,X}(\alpha)\right) + O(\sqrt{X}).$$

Where c is some element of C, and L is the fixed field of $\langle c\rangle \subset \mathrm{Gal}(K/\mathbb{Q})$. Therefore we know that $G_{K,C,X}(\alpha)$ is within $O(\sqrt{X})$ of

$$\frac{|C|}{|G|}\left(\sum_{\chi} \overline{\chi}(c) F_{L,\chi,X}(\alpha)\right).$$

Applying Theorem 5, this is within $O\left(X \log^{-A}(X)\right)$ of

$$\begin{aligned}
&\frac{|C|}{|G|}\left(\sum_{\chi} \overline{\chi}(c) F^{\sharp}_{L,\chi,X,z}(\alpha)\right) \\
&\quad = \frac{|C|}{|G|}\left(\sum_{\chi}\sum_{n\le X} \overline{\chi}(c)\Lambda_{L/\mathbb{Q}}(n)\Lambda_z(n)\chi(n)e(\alpha n)\right) \\
&\quad = \frac{|C|}{|G|}\left(\sum_{n\le X} \Lambda_z(n)e(\alpha n)\left(\Lambda_{L/\mathbb{Q}}(n)\sum_{\chi}\overline{\chi}(c)\chi(n)\right)\right).
\end{aligned}$$

Note that in the above, χ is summed over characters of $\mathrm{Gal}(K/L)$ and that $\chi(n)$ is taken to be 0 unless χ can be extended to a character of $\mathrm{Gal}(K/\mathbb{Q})^{ab}$. We wish the evaluate the inner sum over χ for some $n \in H_L$.

Let the kernel of the map $\langle c\rangle \to \mathrm{Gal}(K/\mathbb{Q})^{ab}$ be generated by c^k for some $k|\mathrm{ord}(c)$. Then $\chi(n)$ is 0 unless $\chi(c^k) = 1$. Therefore we can consider the sum as being over characters χ of $\langle c\rangle/c^k$. Taking K^a to be the maximal abelian subextension of K over $\mathbb{Q}$, this sum is then k if $[K^a/\mathbb{Q}, n] = c$ and 0 otherwise. Hence the sum over χ is non-zero if and only if

$n \in H_C$. The index of H_C in H_L is $[H_L : H_K]$, which is in turn the size of the image of $\langle c \rangle$ in $\mathrm{Gal}(K/\mathbb{Q})^{ab}$, or $|\langle c \rangle / \langle c^k \rangle| = k$. Hence $\Lambda_L(n) \sum_\chi \overline{\chi}(c)\chi(n) = \Lambda_{K,C}(n)$. Therefore $G_{K,C,X}(\alpha)$ is within $O\left(X \log^{-A}(X)\right)$ of

$$\frac{|C|}{|G|} \sum_{n \leq X} \Lambda_{K,C}(n)\Lambda_z(n)e(\alpha n) = G^\sharp_{K,C,X,z}(\alpha).$$

□

6 Proof of Theorem 2

We now have all the tools necessary to prove Theorem 2. Our basic strategy will be as follows. We first define a generating function H for the number of ways to write n as $\sum_i a_i p_i$ for p_i primes satisfying the appropriate conditions. It is easy to write H in terms of the function G. First, we will show that if H is replaced by $H^\sharp$ by replacing these G's by $G^\sharp$'s, this will introduce only a small change (in an appropriate norm). Dealing with $H^\sharp$ will prove noticeably simpler than dealing with H directly. We will essentially be able to approximate the coefficients of $H^\sharp$ using sieving techniques. Finally we combine these results to prove the Theorem.

6.1 Generating Functions

We begin with some basic definitions.

Definition. *Let K_i, C_i, a_i, X be as in the statement of Theorem 2. Then we define*

$$S_{K_i,C_i,a_i,X}(N) := \sum_{\substack{p_i \leq X \\ [K_i/\mathbb{Q},p_i]=C_i \\ \sum_i a_i p_i = N}} \prod_{i=1}^{k} \log(p_i).$$

(i.e. the left hand side of Equation 2). We define the generating function

$$H_{K_i,C_i,a_i,X}(\alpha) := \sum_N S_{K_i,C_i,a_i,X}(N)e(N\alpha).$$

Notice that this is everywhere convergent since there are only finitely many non-zero terms.

We know from basic facts about generating functions that

$$H_{K_i,C_i,a_i,X}(\alpha) = \prod_{i=1}^{k} G_{K_i,C_i,X}(a_i\alpha). \tag{8}$$

We would like to approximate the G's by corresponding $G^\sharp$'s. Hence we define

Definition.

$$H^\sharp_{K_i,C_i,a_i,z,X}(\alpha) := \prod_{i=1}^{k} G^\sharp_{K_i,C_i,z,X}(a_i\alpha).$$

$$H^\flat_{K_i,C_i,a_i,z,X}(\alpha) := H_{K_i,C_i,a_i,X}(\alpha) - H^\sharp_{K_i,C_i,a_i,z,X}(\alpha).$$

We now prove that this is a reasonable approximation.

Lemma 20. *Let A be a constant, and $z = \log^B(X)$ for B a sufficiently large multiple of A. If $k \geq 3$,*

$$\left|H^\flat_{K_i,C_i,a_i,z,X}\right|_1 = O(X^{k-1}\log^{-A}(X)).$$

If $k = 2$,

$$\left|H^\flat_{K_i,C_i,a_i,z,X}\right|_2 = O(X^{3/2}\log^{-A}(X)).$$

Where in the above we are taking the L^1 or L^2 norm respectively of $H^\flat_{K_1,C_i,a_i,X}$ as a function on $[0,1]$.

Proof. Our basic technique is to write each of the G's in Equation 8 as $G^\sharp + G^\flat$ and to expand out the resulting product. We are left with a copy of $H^\sharp$ and a number of terms which are each a product of k $G^\sharp$ or $G^\flat$'s, where each such term has at least one $G^\flat$. We need several facts about various norms of the $G^\sharp$ and $G^\flat$'s. We recall that the squared L^2 norm of a generating function is the sum of the squares of it's coefficients.

- By Theorem 4, the L^∞-norm of $G^\flat$ is $O\left(X\log^{-2A-k}(X)\right)$.
- The L^∞ norm of $G^\sharp$ is clearly $O\left(X\log\log(X)\right)$.
- $|G^\sharp|_2^2 = O(X\log\log^2(X))$.
- $|G|_2^2 = O(X\log(X))$.
- Combining the last two statements, we find that $|G^\flat|_2^2 = O(X\log(X))$.

For $k \geq 3$, we note that by Cauchy-Schwartz, the L^1 norm of a product of k functions is at most the products of the L^2 norms of two of them times the products of the L^∞ norms of the rest. Using this and ensuring that at least one of the terms we take the L^∞ norm of is a $G^\flat$, we obtain our bound on $|H^\flat|_1$.

For $k = 2$, we note that the L^2 norm of a product of two functions is at most the L^2 norm of one times the L^∞ norm of the other. Applying this to our product, ensuring that we take the L^∞ norm of a $G^\flat$ we get the desired bound on $|H^\flat|_2$. □

6.2 Dealing with $H^\sharp$

Now that we have shown that $H^\sharp$ approximates H, it will be enough to compute the coefficients of $H^\sharp$.

Proposition 21. *Let A be a constant, and $z = \log^B(X)$ for B a sufficiently large multiple of A. The $e(N\alpha)$ coefficient of $H^\sharp_{K_i,C_i,a_i,z,X}(\alpha)$ is given by the right hand side of Equation 2, or*

$$\left(\prod_{i=1}^{k} \frac{|C_i|}{|G_i|}\right) C_\infty C_D \left(\prod_{p \nmid D} C_p\right) + O\left(X^{k-1}\log^{-A}(X)\right).$$

Proof. We note that the quantity of interest is equal to

$$\left(\prod_{i=1}^{k} \frac{|C_i|}{|G_i|}\right) \sum_{\substack{n_1,\ldots,n_k \le X \\ \sum_{i=1}^k a_i n_i = N}} \left(\prod_{i=1}^{k} \Lambda_{K_i,C_i}(n_i)\right)\left(\prod_{i=1}^{k} \Lambda_z(n_i)\right). \tag{9}$$

First we consider the number of tuples of n_i that we are summing over. Making a linear change of variables with determinant 1 so that one of the coordinates is $x = \sum_i a_i n_i$, we notice that we are summing over the lattice points of some covolume 1 lattice in a convex region with volume C_∞ and surface area $O\left(X^{k-2}\right)$. Therefore if some affine sublattice L of the set of tuples of integers (n_i) so that $\sum_i a_i n_i = n$ of index I is picked, the number of tuples (n_i) in our sum in this class is $C_\infty/I + O(X^{k-2})$. We can write $\Lambda_{K_i,C_i}(n)$ as a sum of indicator functions for congruences classes of n modulo D. We can also write $\Lambda_z(n) = C(z)\sum_{d|(n,P(z))}\mu(d)$ another sum of indicator functions of congruence conditions. Hence we can write the expression in Equation 9 as a constant (which is $O(C(z))^k$) times the sum over certain affine sublattices of L of ± 1 times the number of points of the intersection of this sublattice with our region. These sublattices are of the following form:

$$\{n_i : \sum_i a_i n_i = n, n_i \equiv x_i \pmod{D}, d_i | n_i\},$$

where x_i are chosen elements of $H_i \subseteq (\mathbb{Z}/D\mathbb{Z})^*$, and $d_i | P(z)$ are integers.

We first claim that the contribution from terms with any d_i bigger than $X^{1/(2k)}$ is negligible. In fact, these terms cannot account for more than

$$\begin{aligned} O(C(z)^k)k \sum_{\substack{d|P(z) \\ d \ge X^{1/(2k)}}} \frac{X^{k-1}}{d} &= O(\log\log(X)^k)\int_{X^{1/(2k)}}^{\infty} X^{k-1}S(z,y)y^{-2}dy \\ &= O\left(X^{k-1-1/((2B)(2k))}\exp\left(O\left(\sqrt{\log(X)}\right)\right)\right). \end{aligned}$$

(Using Lemma 6 to bound the integral above). Throwing out these terms, we would like to approximate the number of tuples in each sublattice by C_∞ divided by the appropriate index. The error introduced by this approximation is $O\left(\log\log^k(X)X^{k-2}\right)$ per term times $O\left(\sqrt{X}\right)$ terms. Hence we can throw this error away. So we have a sum over sets of $d_i|P(z), d_i \leq X^{1/(2k)}$ and x_i of an appropriate constant times C_∞/I. We would like to remove the limitation that $d_i \leq X^{1/(2k)}$ in this sum. We note that once we get rid of the parts of the d_i that share common factors with $D\prod_i a_i$ (for which there are finitely many possible values), the value of I is at least $\frac{\prod_i d_i}{\gcd(d_i)}$. This is because we can compute the index separately for each prime p. If p divides some set of d_i other than all of them, we are forcing the corresponding x_i to have specified values modulo p, when otherwise these values could have been completely arbitrary. If we let $d = \gcd(d_i)$, then we can bound the sum of the reciprocals of the values of I that we are missing as

$$k\sum_d d^{-k+1}\left(\int_0^\infty S(z,y)y^{-2}dy\right)^{k-1}\left(\int_{X^{1/(2k)}/d}^\infty S(z,y)y^{-2}dy\right).$$

This is small by Lemma 6 and a basic computation.

We now wish to evaluate the sum over all x_i and d_i the sum of the appropriate constant times $\frac{1}{I}$. We note that if we were instead trying to evaluate

$$\frac{1}{M^{k-1}}\sum_{\substack{n_i \pmod M \\ \sum_i a_i n_i \equiv N \pmod M}}\left(\prod_{i=1}^k \Lambda_{K_i,C_i}(n_i)\right)\left(\prod_{i=1}^k \Lambda_z(n_i)\right),$$

for $M = DP(z)\prod a_i$, we would get exactly the above sum of $\frac{1}{I}$. This is because the same inclusion exclusion applies to each computation and the number of points in each sublattice here would be exactly $\frac{1}{I}$ of the total points. Hence our final answer up to acceptable errors is

$$C_\infty\left(\prod_{i=1}^k \frac{|C_i|}{|K_i|}\right)\left(\frac{\displaystyle\sum_{\substack{n_i \pmod M \\ \sum_i a_i n_i \equiv N \pmod M}}\left(\prod_{i=1}^k \Lambda_{K_i,C_i}(n_i)\right)\left(\prod_{i=1}^k \Lambda_z(n_i)\right)}{M^{k-1}}\right).$$

Note that $\Lambda_z(n) = \prod_{p\leq z}\Lambda_p(n)$ is a product of terms over the congruence class of n modulo p. Similarly $\Lambda_{K_i,C_i}(n)$ only depends on n modulo D. Therefore we may use the Chinese Remainder Theorem to write the fraction above as a produce of p-primary parts and a D-primary part.

For $p \nmid D, p|P(z)$, the p-primary factor is

$$\frac{\sum_{\substack{n_i \pmod{p^n} \\ \sum_i a_i n_i \equiv N \pmod{p^n}}} \left(\prod_{i=1}^{k} \Lambda_p(n_i)\right)}{(p^n)^{k-1}}$$

for some n. In fact we can use $n = 1$ since $\Lambda_p(n_i)$ only depends on n_i modulo p, and since p does not divide all the a_i, any solution to $\sum_i a_i n_i \equiv N \pmod{p}$ lifts to a solution modulo p^n in exactly $p^{(n-1)(k-1)}$ different ways. Hence the local factor is

$$\begin{aligned}
&\frac{\sum_{\substack{n_i \pmod{p} \\ \sum_i a_i n_i \equiv N \pmod{p}}} \left(\prod_{i=1}^{k} \Lambda_p(n_i)\right)}{p^{k-1}} \\
=&\frac{\left(\frac{p}{p-1}\right)^k \#\{(n_i) \pmod{p} : n_i \not\equiv 0 \pmod{p}, \sum_i a_i n_i \equiv N \pmod{p}\}}{p^{k-1}} \\
=&C_p.
\end{aligned}$$

Next we will compute the D-primary factor. Note that by reasoning similar to the above we can compute the factor modulo D rather than some power of D. Next we will consider the function $\Lambda_{K_i,C_i}(n) \prod_{p|D} \Lambda_p(n)$. This is 0 unless n is in H_i. Otherwise it is $\left(\frac{\phi(D)}{|H_i|}\right)\left(\prod_{p|D} \frac{p}{p-1}\right) = \frac{D}{|H_i|}$. Hence this factor is

$$\begin{aligned}
&\left(\prod_{i=1}^{k} \frac{D}{|H_i|}\right)\left(\frac{\#\{(n_i) \pmod{D} : n_i \in H_i, \sum_i a_i n_i \equiv N \pmod{D}\}}{D^{k-1}}\right) \\
&= C_D.
\end{aligned}$$

Putting these factors together we obtain our result. □

6.3 Putting it Together

We are finally able to prove Theorem 2

Proof. Let B be a sufficiently large multiple of A, and $z = \log^B(X)$.

For $k \geq 3$ we have that

$$S_{K_i,C_i,a_i,X}(N) = \int_0^1 H_{K_i,C_i,a_i,X}(\alpha) e(-N\alpha).$$

By Lemma 20 this is

$$\int_0^1 H^\sharp_{K_i,C_i,a_i,z,X}(\alpha)e(-N\alpha)$$

up to acceptable errors. This is the $e(N\alpha)$ coefficient of $H^\sharp_{K_i,C_i,a_i,z,X}(\alpha)$, which by Proposition 21 is as desired.

For $k = 2$, we let $T_{K_i,C_i,a_i,X}(N)$ be the corresponding right hand side of Equation 2. It will suffice to show that

$$\sum_{|n|\leq\sum_i |a_i|X} (S_{K_i,C_i,a_i,X}(N) - T_{K_i,C_i,a_i,X}(N))^2 = O(X^3 \log^{-2A}(X)).$$

If we define the generating function

$$J_{K_i,C_i,a_i,X}(\alpha) = \sum_{|N|\leq\sum_i |a_i|X} T_{K_i,C_i,a_i,X}(N)e(N\alpha)$$

we note that the above is equivalent to showing that

$$|H_{K_i,C_i,a_i,X} - J_{K_i,C_i,a_i,X}|_2 = O(X^{3/2} \log^{-A}(X)).$$

But by Lemma 20, we have that

$$|H_{K_i,C_i,a_i,X} - H^\sharp_{K_i,C_i,a_i,z,X}|_2 = O(X^{3/2} \log^{-A}(X)),$$

and by Proposition 21, we have

$$|H^\sharp_{K_i,C_i,a_i,z,X} - J_{K_i,C_i,a_i,X}|_2 = O(X^{3/2} \log^{-A}(X)).$$

This completes the proof. □

7 Application

We present an application of Theorem 2 to the construction of elliptic curves whose discriminants are divisible only by primes with certain splitting properties.

Theorem 22. *Let K be a number field. Then there exists an elliptic curve defined over $\mathbb{Q}$ so that all primes dividing its discriminant split completely over K.*

Proof. We begin by assuming that K is a normal extension of $\mathbb{Q}$. We will choose an elliptic curve of the form:

$$y^2 = X^3 + AX + B.$$

Here we will let $A = pq/4$, $B = npq^2$ where n is a small integer and p, q are primes that split over K. The discriminant is then

$$\begin{aligned} -16(4A^3 + 27B^3) &= -64p^3q^3/64 - 432n^2p^2q^4 \\ &= -p^2q^3(p + 432n^2q). \end{aligned}$$

Hence it suffices to find primes p, q, r that split completely over K with $p + 432n^2q - r = 0$. We do this by applying Theorem 2 with $k = 3$, $K_i = K$, $C_i = \{e\}$, and X large. As long as $C_D > 0$ and $C_p > 0$ for all p, the main term will dominate the error and we will be guaranteed solutions for sufficiently large X. If $n = D$, this will hold. This is because for C_D to be non-zero we need to have solutions $n_1 + 0n_2 - n_3 \equiv 0 \pmod{D}$ with n_i all in some particular subgroup of $(\mathbb{Z}/D\mathbb{Z})^*$. This can clearly be satisfied by $n_1 = n_3$. For $p = 2$, C_p is non-zero since there is a solution to $n_1 + 0n_2 - n_3 \equiv 0 \pmod{2}$ with none of the n_i divisible by 2 (take $(1, 1, 1)$). For $p > 2$, we need to show that there are solutions to $n_1 + 432D^2n_2 - n_3 \equiv 0 \pmod{p}$ with none of the n_i 0 modulo p. This can be done because after picking n_2, any number can be written as a difference of non-multiples of p. $\square$

Chapter C

Ranks of 2-Selmer Groups of Twists Elliptic Curves

1 Introduction

Let c_1, c_2, c_3 be distinct rational numbers. Let E be the elliptic curve defined by the equation

$$y^2 = (x - c_1)(x - c_2)(x - c_3).$$

We make the additional technical assumption that none of the $(c_i - c_j)(c_i - c_k)$ are squares. This is equivalent to saying that E is an elliptic curve over $\mathbb{Q}$ with complete 2-torsion and no cyclic subgroup of order 4 defined over $\mathbb{Q}$. For b a square-free number, let E_b be the twist defined by the equation

$$y^2 = (x - bc_1)(x - bc_2)(x - bc_3).$$

Let S be a finite set of places of $\mathbb{Q}$ including $2, \infty$ and all of the places at which E has bad reduction. Let D be a positive integer divisible by 8 and by the primes in S. Let $S_2(E_b)$ denote the 2-Selmer group of the curve E_b. We will be interested in how the rank varies with b and in particular in the asymptotic density of b's so that $S_2(E_b)$ has a given rank.

The parity of $\dim(S_2(E_b))$ depends only on the class of b as an element of $\prod_{\nu \in S} \mathbb{Q}_\nu^*/(\mathbb{Q}_\nu^*)^2$. We claim that for exactly half of these values this dimension is odd and exactly half of the time it is even.

Lemma 1. *For exactly half of the classes c in $(\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2$, if we pick b a positive representative of the class c then $\dim(S_2(E_b))$ is even.*

We put off the proof of this statement until Section 4.

Let $b = p_1 p_2 \dots p_n$ where p_i are distinct primes relatively prime to D. In [58] the rank of $S_2(E_b)$ is shown to depend only on the images of the p_i in $(\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2$ and upon

which p_i are quadratic residues modulo which p_j. There are $2^{n|S|+\binom{n}{2}}$ possible sets of values for these. Let $\pi_d(n)$ be the fraction of this set of possibilities that cause $S_2(E_b)$ to have rank exactly d. Then the main Theorem of [58] together with Lemma 1 implies that:

Theorem 2.

$$\lim_{n\to\infty} \pi_d(n) = \alpha_d.$$

where $\alpha_0 = \alpha_1 = 0$ *and* $\alpha_{n+2} = \frac{2^n}{\prod_{j=1}^{n}(2^j-1)\prod_{j=0}^{\infty}(1+2^{-j})}$.

The actual Theorem proved in [58] says that if, in addition, the class of b in $\prod_{\nu\in S} \mathbb{Q}_\nu^*/(\mathbb{Q}_\nu^*)^2$ is fixed, then the analogous $\pi_d(n)$ either converge to $2\alpha_d$ for d even and 0 for d odd, or to $2\alpha_d$ for d odd and 0 for d even.

This tells us information about the asymptotic density of twists of E whose 2-Selmer group has a particular rank. Unfortunately, this asymptotic density is taken in a somewhat awkward way by letting the number of primes dividing b go to infinity. In this Chapter, we prove the following more natural version of Theorem 2:

Theorem 3. *Let E be an elliptic curve over $\mathbb{Q}$ with full 2-torsion defined over $\mathbb{Q}$ so that in addition for E we have that*

$$\lim_{n\to\infty} \pi_d(n) = \alpha_d.$$

With α_d as given in Theorem 2. Then

$$\lim_{N\to\infty} \frac{\#\{b \le N : b \text{ square-free, } (b,D)=1 \text{ and } \dim(S_2(E_b)) = d\}}{\#\{b \le N : b \text{ square-free and } (b,D)=1\}} = \alpha_d.$$

Applying this to twists of E by divisors of D and noting that twists by squares do not affect the Selmer rank we have that

Corollary 4.

$$\lim_{N\to\infty} \frac{\#\{b \le N : \dim(S_2(E_b)) = d\}}{N} = \alpha_d.$$

and

Corollary 5.

$$\lim_{N\to\infty} \frac{\#\{-N \le b \le N : \dim(S_2(E_b)) = d\}}{2N} = \alpha_d.$$

Our technique is fairly straightforward. Our goal will be to prove that the average moments of the size of the Selmer groups will be as expected. As it turns out, this will be enough to determine the probability of seeing a given rank. In order to analyze the Selmer groups we follow the method described in [58]. Here the 2-Selmer group of E_b can be expressed as the intersection of two Lagrangian subspaces, U and W, of a particular

symplectic space, V, over $\mathbb{F}_2$. Although U, V and W all depend on b, once the number of primes dividing b has been fixed along with its congruence class modulo D, these spaces can all be written conveniently in terms of the primes, p_i, dividing b, which we think of as formal variables. Using the formula $|U \cap W| = \frac{1}{\sqrt{|V|}} \sum_{u \in U, w \in W} (-1)^{u \cdot w}$, we reduce our problem to bounding the size of the "characters" $(-1)^{u \cdot w}$ when averaged over b. These "characters" turn out to be products of Dirichlet characters of the p_i and Legendre symbols of pairs of the p_i. The bulk of our analytic work is in proving these bounds. These bounds will allow us to discount the contribution from most of the terms in our sum (in particular the ones in which Legendre symbols show up in a non-trivial way), and allow us to show that the average of the remaining terms is roughly what should be expected from Swinnerton-Dyer's result.

We should point out the connections between our work and that of Heath-Brown in [22] where he proves our main result for the particular curve

$$y^2 = x^3 - x.$$

We employ techniques similar to those of [22], but the algebra behind them is organized significantly differently. Heath-Brown's overall strategy is again to compute the average sizes of moments of $|S_2(E_b)|$ and use these to get at the ranks. He computes $|S_2(E_b)|$ using a different formula than ours. Essentially what he does is use some tricks specific to his curve to deal with the conditions relating to primes dividing D, and instead of considering each prime individually, he groups them based on how they occur in u and w. He lets D_i be the product of all primes dividing b that relate in a particular way (indexed by i). He then gets a formula for $|S_2(E_b)|$ that's a sum over ways of writing b as a product, $b = \prod D_i$, of some term again involving characters of the D_i and Legendre symbols. Using techniques similar to ours he shows that terms in this sum where the Legendre symbols have a non-negligible contribution (are not all trivial due to one of the D_i being 1) can be ignored. He then uses some algebra to show that the average of the remaining terms is the desired value. This step differs from our technique where we merely make use of Swinnerton-Dyer's result to compute our average. Essentially we show that the algebra and the analysis for this problem can be done separately and use [58] to take care of the algebra. Finally, Heath-Brown uses some techniques from linear algebra to show that the moment bounds imply the correct densities of ranks, while we use techniques from complex analysis.

In Section 2, we introduce some basic concepts that will be used throughout. In Section 3, we will prove the necessary character bounds. We use these bounds in Section 4 to establish the average moments of the size of the Selmer groups. Finally, in Section 5, we explain how these results can be used to prove our main Theorem.

2 Preliminaries

2.1 Asymptotic Notation

Throughout the rest of this paper we will make extensive use of O, and similar asymptotic notation. In our notation, $O(X)$ will denote a quantity that is at most $H \cdot X$ for some *absolute* constant H. If we need asymptotic notation that depends on some parameters, we will use $O_{a,b,c}(X)$ to denote a quantity that is at most $H(a,b,c) \cdot X$, where H is some function depending only on a, b and c.

2.2 Number of Prime Divisors

In order to make use of Swinnerton-Dyer's result, we will need to consider twists of E by integers $b \leq N$ with a specific number of prime divisors. For an integer m, we let $\omega(m)$ be the number of prime divisors of m. In our analysis, we will need to have estimates on the number of of such b with a particular number of prime divisors. We define

$$\Pi_n(N) = \#\{\text{primes } p \leq N \text{ so that } \omega(p) = n\}.$$

In order to deal with this we use Lemma A of [53] which states:

Lemma 6. *There exist absolute constants C and K so that for any ν and x*

$$\Pi_{\nu+1}(x) \leq \frac{Kx}{\log(x)} \frac{(\log\log x + C)^\nu}{\nu!}.$$

By maximizing the above in terms of ν it is easy to see that

Corollary 7.

$$\Pi_n(N) = O\left(\frac{N}{\sqrt{\log\log(N)}}\right).$$

It is also easy to see from the above that most integers of size roughly N have about $\log\log(N)$ prime factors. In particular:

Corollary 8. *There is a constant $c > 0$ so that for all N, the number of $b \leq N$ with $|\omega(b) - \log\log(N)| > \log\log(N)^{3/4}$ is at most*

$$2N \exp\left(-c\sqrt{\log\log(N)}\right).$$

In particular, the fraction of $b \leq N$ with $|\omega(b) - \log\log(N)| < \log\log(N)^{3/4}$ goes to 1 as N goes to infinity.

We will use Corollary 8 to restrict our attention only to twists by b with an appropriate number of prime divisors.

3 Character Bounds

Our main purpose in this section will be to prove the following Propositions:

Proposition 9. *Fix positive integers D, n, N with $4|D$, $\log\log N > 1$, and $(\log\log N)/2 < n < 2\log\log N$, and let $c > 0$ be a real number. Let $d_{i,j}, e_{i,j} \in \mathbb{Z}/2$ for $i, j = 1, \ldots, n$ with $e_{i,j} = e_{j,i}, d_{i,j} = d_{j,i}, e_{i,i} = d_{i,i} = 0$ for all i, j. Let χ_i be a quadratic character with modulus dividing D for $i = 1, \ldots, n$. Let m be the number of indices i so that at least one of the following hold:*

- *$e_{i,j} = 1$ for some j or*
- *χ_i has modulus not dividing 4 or*
- *χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.*

Let $\epsilon(p) = (p-1)/2$. Then if $m > 0$

$$\left| \frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} \prod_i \chi_i(p_i) \prod_{i<j} (-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}} \prod_{i<j} \left(\frac{p_i}{p_j}\right)^{e_{i,j}} \right| = O_{c,D}\left(Nc^m\right). \tag{1}$$

Note that m is the number of indices i so that no matter how we fix the values of p_j for the $j \neq i$ that the summand on the left hand side of Equation 1 still depends on p_i.

The sum can be thought of as a sum over all $b = \prod_i p_i$ with $\omega(b) = n$ and $b \leq N$, where the summand is a "character" defined by the $\chi_i, d_{i,j}$ and $e_{i,j}$. The $\frac{1}{n!}$ accounts for the different possible reorderings of the p_i. This Proposition will allow us to show that the "characters" in which the Legendre symbols make a non-trivial appearance add a negligible contribution to our moments.

Proposition 10. *Let n, N, D be positive integers with $\log\log N > 1$, and $(\log\log N)/2 < n < 2\log\log N$. Let $G = ((\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2)^n$. Let $f : G \to \mathbb{C}$ be a function with $|f|_\infty \leq 1$.*

Then

$$\frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} f(p_1,\ldots,p_n) \tag{2}$$

$$= \left(\frac{1}{|G|} \sum_{g \in G} f(g) \right) \left(\frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} 1 \right) + O_D \left(\frac{N(\log\log\log N)}{\log\log N} \right).$$

(Above $f(p_1,\ldots,p_n)$ is really f applied to the vector of their reductions modulo D).

Note that again, the sum can be thought of as a sum over all $b = \prod_i p_i$ with $b \leq N$ and $\omega(b) = n$. This Proposition says that the average of f over such $b = \prod_i p_i$ is roughly equal to the average of f over G. This will allow us to show that the average value of the remaining terms in our moment calculation equal what we would expect given Swinnerton-Dyer's result.

We begin with a Proposition that gives a more precise form of Proposition 9 in the case when the $e_{i,j}$ are all 0.

Proposition 11. *Let D, n, N be integers with $4|D$ with $\log\log N > 1$. Let $C > 0$ be a real number. Let $d_{i,j} \in \mathbb{Z}/2$ for $i, j = 1,\ldots,n$ with $d_{i,j} = d_{j,i}, d_{i,i} = 0$. Let χ_i be a quadratic character of modulus dividing D for $i = 1,\ldots,n$. Suppose that no Dirichlet character of modulus dividing D has an associated Siegel zero larger than $1 - \beta^{-1}$. Let*

$$B = \max(e^{(C+2)\beta \log\log N}, e^{K(C+2)^2 (\log D)^2 (\log\log(DN))^2}, n \log^{C+2}(N))$$

for K a sufficiently large absolute constant. Suppose that $B^n < \sqrt{N}$. Let m be the number of indices i so that either:

- *χ_i does not have modulus dividing 4 or*
- *χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.*

Then

$$\left| \frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} \prod_i \chi_i(p_i) \prod_{i<j} (-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}} \right| \tag{3}$$
$$= O\left(\frac{N}{\sqrt{\log\log(N)}}\right)\left(O\left(\frac{\log\log B}{n}\right)^m + O(\log N)^{-C}\right).$$

Note once again that m is the number of i so that if the values of p_j for $j \neq i$ are all fixed, the resulting summand will still depend on p_i.

The basic idea of the proof will be by induction on m. If $m = 0$, we can bound by the number of terms in our sum, giving a bound of $\Pi_n(N)$, which we bound using Corollary 7. If $m > 0$, there is some p_i so that no matter how we set the other p_j, our character still depends on p_i. We split into cases based on whether $p_i > B$. If $p_i > B$, we fix the values of the other p_j, and use bounds on character sums. For $p_i \leq B$, we note that this happens for only about a $\frac{\log\log B}{n}$ fraction of the terms in our sum, and for each possible value of p_i inductively bound the remaining sum. To deal with the first case we prove the following Lemma:

Lemma 12. *Let K be a sufficiently large constant. Take χ any non-trivial Dirichlet character of modulus at most D and with no Siegel zero more than $1 - \beta^{-1}$, $N, C > 0$ integers, and X any integer with*

$$X > \max(e^{(C+2)\beta\log\log N}, e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}).$$

Then

$$\left|\sum_{p\leq X} \chi(p)\right| \leq O(X\log^{-C-2}(N)).$$

Where the sum above is over primes p less than or equal to X.

Proof. [24] Theorem 5.27 implies that for any X that for some constant $c > 0$,

$$\sum_{n\leq Y} \chi(n)\Lambda(n) = YO\left(Y^{-\beta^{-1}} + \exp\left(\frac{-c\sqrt{\log(Y)}}{\log D}\right)(\log D)^4\right).$$

Note that the contribution to the above coming from n a power of a prime is $O(\sqrt{Y})$. Using Abel summation to reduce this to a sum over p of $\chi(p)$ rather than $\chi(p)\log(p)$, we find that

$$\sum_{p\leq X} \chi(p) \leq XO\left(X^{-\beta^{-1}} + \exp\left(\frac{-c\sqrt{\log(X)}}{\log D}\right)(\log D)^4\right) + O(\sqrt{X}).$$

The former term is sufficiently small since by assumption $X > e^{(C+2)\beta \log\log N}$. The latter term is small enough since $X > e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}$. The last term is small enough since clearly $X > \log^{2C+4}(N)$. □

For positive integers n, N, D, and S a set of prime numbers, denote by $Q(n, N, D, k, S)$ the maximum possible absolute value of a sum of the form given in Equation 3 with $m \geq k$, with the added restriction that none of the p_i lie in S. In particular a sum of the form

$$\frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ p_i \notin S \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} \prod_i \chi_i(p_i) \prod_{i<j} (-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}}$$

where χ_i are characters of modulus dividing D, and $d_{i,j} \in \{0, 1\}$.

We write the inductive step for our main bound as follows:

Lemma 13. *For integers, n, D, N, M, C and B with*

$$B > \max(e^{(C+2)\beta \log\log M}, e^{K(C+2)^2(\log D)^2(\log\log(DM))^2}, n\log^{C+2}(M)),$$

where $1 - \beta^{-1}$ is the largest Siegel zero of a Dirichlet character of modulus dividing D and K a sufficiently large constant, $1 \leq k \leq n$ and S a set of primes $\leq B$, then $Q(n, N, D, k, S)$ as described above is at most

$$O(N\log(N)\log^{-C-2}(M)) + \frac{1}{n} \sum_{\substack{p<B \\ p \notin S}} Q(n-1, N/p, D, k-1, S \cup \{p\}).$$

Proof. Since $k \geq 1$, there must be an i so that either χ_i has modulus bigger than 4 or has modulus exactly 4 and all of the $d_{i,j}$ are 0. Without loss of generality, n is such an index. We split our sum into cases depending on whether $p_n \geq B$. For $p_n \geq B$, we proceed by fixing all of the p_j for $j \neq n$ and summing over p_n. Letting $P = \prod_{i=1}^{n-1} p_i$, we have

$$\sum_{P=1}^{N/B} \frac{1}{n!} \sum_{\substack{P=p_1\ldots p_{n-1} \\ p_i \text{ distinct} \\ p_i \notin S \\ (D,P)=1}} a \sum_{\substack{B \leq p_n \leq N/P \\ p_n \neq p_j}} \chi(p_n)$$

where a is some constant of norm 1 depending on $p_1 \ldots p_{n-1}$, and χ is a non-trivial character of modulus dividing D, perhaps also depending on $p_1, \ldots, p_{n-1}$. The condition that $p_n \neq p_j$ alters the value of the inner sum by at most n. With this condition removed, we may bound the inner sum by applying Lemma 12 (taking the difference of the terms with $X = N/P$

and $X = B$). Hence the value of the inner sum is at most $O(N/P \log^{-C-2}(M) + n)$. Since $N/P \geq B \geq n \log^{C+2}(M)$, this is just $O(N/P \log^{-C-2}(M))$. Note that for each P, there are at most $(n-1)!$ ways of writing it as a product of $n-1$ primes (since the primes will be unique up to ordering). Hence, ignoring the extra $1/n$ factor, the sum above is at most

$$\sum_{P=1}^{N/B} O(N/P \log^{-C-2}(M)) = O(N \log(N) \log^{-C-2}(M)).$$

For $p_n < B$, we fix p_n and consider the sum over the remaining p_i. We note that for p a prime not in S and relatively prime to D, this sum is plus or minus one over n times a sum of the type bounded by $Q(n-1, N/p, D, k-1, S \cup \{p\})$. This completes our proof. □

We are now prepared to Prove Proposition 11

Proof. We prove by induction on k that for n, N, D, C, M, β, B as above with

$$B > \max(e^{(C+2)\beta \log\log M}, e^{K(C+2)^2 (\log D)^2 (\log\log(DM))^2}, n \log^{C+2}(M)),$$

and S a set of primes $\leq B$ that

$$\begin{aligned} Q(n, N, D, k, S) \leq & O\left(\frac{N}{\sqrt{\log\log N/B^n}}\right) O\left(\frac{\log\log B}{n}\right)^k \\ & + O(N \log(N) \log^{-C-2}(M)) \sum_{a=0}^{k-1} O\left(\frac{\log\log B}{n}\right)^a. \end{aligned} \tag{4}$$

Plugging in $M = N$, $k = m$, $S = \emptyset$, and

$$B = \max(e^{(C+2)\beta \log\log N}, e^{K(C+2)^2 (\log D)^2 (\log\log(DN))^2}, n \log^{C+2}(N)),$$

yields the necessary result.

We prove Equation 4 by induction on k. For $k = 0$, the sum is at most the sum over $b = p_1 \dots p_n$ with appropriate conditions of $\frac{1}{n!}$. Since each such b can be written as such a product on at most $n!$ ways, this is at most $\Pi_n(N)$, which by Corollary 7 is at most $O\left(\frac{N}{\sqrt{\log\log N}}\right)$ as desired.

For larger values of k, we use the inductive hypothesis and Lemma 13 to bound $Q(n, N, D, k, S)$

by

$$
\begin{aligned}
&O(N\log(N)\log^{-C-2}(M)) + \frac{1}{n}\sum_{p<B} Q(n-1, N/p, D, k-1, S') \\
\leq &O(N\log(N)\log^{-C-2}(M)) \\
&+ \frac{1}{n}\sum_{p<B}\frac{1}{p}O\left(\frac{N}{\sqrt{\log\log N/pB^{n-1}}}\right)O\left(\frac{\log\log B}{n-1}\right)^{k-1} \\
&+ \frac{1}{n}\sum_{p<B}\frac{1}{p}O(N\log(N)\log^{-C-2}(M))\sum_{a=0}^{k-2} O\left(\frac{\log\log B}{n-1}\right)^{a} \\
\leq &O(N\log(N)\log^{-C-2}(M)) \\
&+ O\left(\frac{N}{\sqrt{\log\log N/B^{n}}}\right)O\left(\frac{\log\log B}{n}\right)^{k} \\
&+ O(N\log(N)\log^{-C-2}(M))\sum_{a=0}^{k-2} O\left(\frac{\log\log B}{n}\right)^{a+1} \\
\leq &O\left(\frac{N}{\sqrt{\log\log N/B^{n}}}\right)O\left(\frac{\log\log B}{n}\right)^{k} \\
&+ O(N\log(N)\log^{-C-2}(M))\sum_{a=0}^{k-1} O\left(\frac{\log\log B}{n}\right)^{a}.
\end{aligned}
$$

Above we use that $\sum_{p<B}\frac{1}{p} = O(\log\log B)$ and that $\frac{1}{n}\left(\frac{1}{n-1}\right)^a = O\left(\left(\frac{1}{n}\right)^{a+1}\right)$ for $a \leq n$. This completes the inductive hypothesis, proving Equation 4, and completing the proof. □

We are now prepared to prove Proposition 10

Proof. First note that we can assume that $4|D$. This is because if that is not the case, we can split our sum up into two cases, one where none of the p_i are 2, and one where one of the p_i is 2. In either case we get a sum of the same form but now can assume that D is divisible by 4. We assume this so that we can use Proposition 11.

It is clear that the difference between the left hand side of Equation 2 and the main term

on the right hand side is

$$\frac{1}{|G|}\left(\sum_{\chi\in\widehat{G}\setminus\{1\}}\left(\frac{1}{n!}\sum_{\substack{p_1,\ldots,p_n\\ \text{distinct primes}\\ (D,p_i)=1\\ \prod_i p_i\leq N}}\chi(p_1,\ldots,p_n)\right)\left(\sum_{g\in G}f(g)\chi(g)\right)\right).$$

Using Cauchy-Schwarz we find that this is at most

$$\frac{1}{|G|}\sqrt{|G|}|f|_2\left(\sum_{\chi\in\widehat{G}\setminus\{1\}}\left|\frac{1}{n!}\sum_{\substack{p_1,\ldots,p_n\\ \text{distinct primes}\\ (D,p_i)=1\\ \prod_i p_i\leq N}}\chi(p_1,\ldots,p_n)\right|^2\right)^{1/2}.$$

We note that $|f|_2\leq\sqrt{|G|}$ and hence that $\frac{1}{|G|}\sqrt{|G|}|f|_2\leq 1$. Bounding the character sum using Proposition 11 (using the minimal possible value of B), we get $O\left(\frac{N^2}{\log\log N}\right)$ times

$$\sum_{\chi\in\widehat{G}\setminus\{1\}}O_D\left(\frac{\log\log\log N}{\log\log N}\right)^{2s}.$$

Where above s is the number of components on which χ (thought of as a product of characters of $(\mathbb{Z}/D\mathbb{Z})^*$) is non-trivial. Since each component of χ can either be trivial or have one of finitely many non-trivial values (each of which contributes $O_D((\log\log\log N)^2/(\log\log N)^2)$) and this can be chosen independently for each component, the inner sum is

$$\begin{aligned}\left(1+O_D\left(\frac{\log\log\log N}{\log\log N}\right)^2\right)^n-1&=\exp\left(O_D\left(\frac{(\log\log\log N)^2}{\log\log N}\right)\right)-1\\&=O_D\left(\frac{(\log\log\log N)^2}{\log\log N}\right).\end{aligned}$$

Hence the total error is at most

$$\frac{1}{|G|}\sqrt{|G|}\sqrt{|G|}O_D\left(\left(\frac{N^2\log\log\log^2(N)}{\log\log^2(N)}\right)^{1/2}\right)=O_D\left(\frac{N\log\log\log(N)}{\log\log(N)}\right).$$

□

The proof of Proposition 9 is along the same lines as the proof of Proposition 11. Again we induct on m. This time, we use Lemma 13 as our base case (when all of the $e_{i,j}$ are 0). If some $e_{i,j}$ is non-zero, we break into cases based on whether or not p_i and p_j are larger than some integer A (which will be some power of $\log(N)$). If both, p_i and p_j are large, then fixing the remaining primes and summing over p_i and p_j gives a relatively small result. Otherwise, fixing one of these primes at a small value, we are left with a sum of a similar form over the other primes. Unfortunately, doing this will increase our D by a factor of p_i, and may introduce characters with bad Siegel zeroes. To counteract this, we will begin by throwing away all terms in our sum where $D\prod_i p_i$ is divisible by the modulus of the worst Siegel zero in some range, and use standard results to bound the badness of other Siegel zeroes.

We begin with some Lemmas that will allow us to bound sums of Legendre symbols of p_i and p_j as they vary over primes.

Lemma 14. *Let Q and N be positive integers with $Q^2 \geq N$. Let a be a function $\{1, 2, \ldots, N\} \to \mathbb{C}$, supported on square-free numbers. Then we have that*

$$\sum_{\substack{\chi \text{ quadratic character} \\ \text{of modulus } p \text{ or } 4p,\ \leq Q}} \left|\sum_{n=1}^{N} a_n \chi(n)\right|^2 = O\left(Q\sqrt{N}|a|^2\right)$$

where the sum is over quadratic characters whose modulus is either a prime or four times a prime and is less than Q, and where $|a|^2 = \sum_{n=1}^{N} |a_n|^2$ is the squared L^2 norm.

Note the similarity between this and Lemma 3 of [22].

Proof. Let M be a positive integer so that $Q^2 \leq NM^2 \leq 4Q^2$. Let $b : \{1, 2, \ldots, M^2\} \to \mathbb{C}$ be the function $b_{n^2} = \frac{1}{M}$ and $b = 0$ on non-squares. Let $c = a * b$ be the multiplicative convolution of a and b. Note that since a is supported on square-free numbers and b supported on squares that $|c|^2 = |a|^2|b|^2 = |a|^2/M$. Applying the multiplicative large sieve inequality (see [24] Theorem 7.13) to c we have that

$$\sum_{q \leq Q} \frac{q}{\phi(q)} \sideset{}{^*}\sum_{\chi \bmod q} \left|\sum_n c_n \chi(n)\right|^2 \leq (Q^2 + NM^2 - 1)|c|^2. \tag{5}$$

The right hand side is easily seen to be

$$O(Q^2)|a|^2/M = O(Q^2|a|^2/(\sqrt{Q^2/N})) = O(Q\sqrt{N}|a|^2).$$

For the left hand side we may note that it only becomes smaller if we remove the $\frac{q}{\phi(q)}$ or ignore the characters that are not quadratic or do not have moduli either a prime or four times a prime. For such characters χ note that

$$\sum_n c_n \chi(n) = \left(\sum_n a_n \chi(n)\right)\left(\sum_n b_n \chi(n)\right) = \Omega(\sum_n a_n \chi(n)).$$

Where the last equality above follows from the fact that χ is 1 on squares not dividing its modulus, and noting that since its modulus divides four times a prime, the latter case only happens at even numbers of multiples of p. Hence the left hand side of Equation 5 is at least a constant multiple of

$$\sum_{\substack{\chi \text{ quadratic character} \\ \text{of modulus } p \text{ or } 4p,\ \leq Q}} \left| \sum_{n=1}^{N} a_n \chi(n) \right|^2.$$

This completes our proof. □

Lemma 15. *Let $A \leq X$ be positive numbers, and let $a, b : \mathbb{Z} \to \mathbb{C}$ be functions so that $|a(n)|, |b(n)| \leq 1$ for all n. We have that*

$$\left| \sum_{\substack{A \leq p_1, p_2 \\ p_1 p_2 \leq X}} a(p_1) b(p_2) \left(\frac{p_1}{p_2} \right) \right| = O(X \log(X) A^{-1/8}).$$

Where the above sum is over pairs of primes p_i bigger than A with $p_1 p_2 \leq X$, and where $\left(\frac{p_1}{p_2}\right)$ is the Legendre symbol.

Proof. We first bound the sum of the terms for which $p_1 \leq \sqrt{X}$.

We begin by partitioning $[A, \sqrt{X}]$ into $O(A^{1/4} \log(X))$ intervals of the form $[Y, Y(1 + A^{-1/4}))$. We break up our sum based on which of these intervals p_1 lies in. Once such an interval is fixed, we throw away the terms for which $p_2 \geq X/(Y(1 + A^{-1/4}))$. We note that for such terms $p_1 p_2 \geq X(1 + A^{-1/4})^{-1}$. Therefore the number of such terms in our original sum is at most $O(XA^{-1/4})$, and thus throwing these away introduces an error of at most $O(XA^{-1/4})$.

The sum of the remaining terms is at most

$$\sum_{A \leq p_2 \leq X/(Y(1+A^{-1/4}))} \left| \sum_{Y \leq p_1 \leq Y(1+A^{-1/4})} a(p_1) \left(\frac{p_1}{p_2} \right) \right|.$$

By Cauchy-Schwarz, this is at most

$$\sqrt{X/Y} \left(\sum_{A \leq p_2 \leq X/(Y(1+A^{-1/4}))} \left| \sum_{Y \leq p_1 \leq Y(1+A^{-1/4})} a(p_1) \left(\frac{p_1}{p_2} \right) \right|^2 \right)^{1/2}.$$

In the evaluation of the above, we may restrict the support of a to primes between Y and $Y(1 + A^{-1/4})$. Therefore, by Lemma 14, the above is at most

$$\sqrt{X/Y} O\left(\sqrt{(X/Y) Y^{1/2} (Y A^{-1/4})} \right) = O\left(X Y^{-1/4} A^{-1/8} \right) = O(X A^{-3/8}).$$

Hence summing over the $O(A^{1/4}\log(X))$ such intervals, we get a total contribution of $O(X\log(X)A^{-1/8})$.

We get a similar bound on the sum of terms for which $p_2 \leq \sqrt{X}$. Finally we need to subtract off the sum of terms where both p_1 and p_2 are at most $\sqrt{X}$. This is

$$\sum_{A\leq p_1\leq\sqrt{X}}\sum_{A\leq p_2\leq\sqrt{X}} a(p_1)b(p_2)\left(\frac{p_1}{p_2}\right).$$

This is at most

$$\sum_{A\leq p_2\leq\sqrt{X}}\left|\sum_{A\leq p_1\leq\sqrt{X}} a(p_1)\left(\frac{p_1}{p_2}\right)\right|.$$

By Cauchy-Schwarz and Lemma 14, this is at most

$$\sqrt{X^{1/2}}O\left(\sqrt{X^{1/2}X^{1/4}X^{1/2}}\right) = O(X^{7/8}) = O(XA^{-1/8}).$$

Hence all of our relevant factors are $O(X\log(X)A^{-1/8})$, thus proving our bound. □

As mentioned above, in proving Proposition 9, we are going to want to deal separately with the terms in which $D\prod_i p_i$ is divisible by a particular bad Siegel zero. In particular, for $X \leq Y$, let $q(X,Y)$ be the modulus of the Dirichlet character with the worst (closest to 1) Siegel zero of any Dirichlet character with modulus between X and Y. In analogy with the Q defined in the proof of Proposition 11, for integers n, N, D, k, X, Y and a set S of primes, we define $Q(n, N, D, k, X, Y, S)$ to be the largest possible value of

$$\left|\frac{1}{n!}\sum_{\substack{p_1,\ldots,p_n\\ \text{distinct primes}\\ p_i\notin S\\ q(X,Y)\nmid D\prod_i p_i\\ (D,p_i)=1\\ \prod_i p_i\leq N}}\prod_i \chi_i(p_i)\prod_{i<j}(-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}}\prod_{i<j}\left(\frac{p_i}{p_j}\right)^{e_{i,j}}\right| \tag{6}$$

where χ_i are Dirichlet characters of modulus dividing D, $e_{i,j}, d_{i,j} \in \{0,1\}$, and k is at most the number of indices i so that one of:

- $e_{i,j} = 1$ for some j or
- χ_i has modulus not dividing 4 or

- χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.

We wish to prove an inductive bound on Q. In particular we show:

Lemma 16. *Let n, N, D, k, X, Y be as above. Let β be a real number so that the worst Siegel zero of a Dirichlet series of modulus at most D other than $q(X, Y)$ is at most $1 - \beta^{-1}$. Let M, A, B, C be integers so that*

$$B > \max(e^{(C+2)\beta \log\log M}, e^{K(C+2)^2(\log D)^2(\log\log(DM))^2}, n\log^{C+2}(M), A)$$

for a sufficiently large constant K. Then for S a set of primes $\leq A$, we have that $Q(n, N, D, k, X, Y, S)$ is at most the maximum of

$$N\left(O\left(\frac{\log\log B}{n}\right)^k + O(\log(N)\log^{-C-2}(M))\sum_{a=0}^{k-1} O\left(\frac{\log\log B}{n}\right)^a\right)$$

and

$$O(N\log^2(N)A^{-1/8}) + \frac{2}{n}\sum_{p<A} Q(n-1, N/p, Dp, k-1, X, Y, S\cup\{p\})$$
$$+ \frac{1}{n(n-1)}\sum_{p_1,p_2<A} Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S\cup\{p_1,p_2\}).$$

Proof. We consider a sum of the form given in Equation 6. If all of the $e_{i,j}$ are 0, we have a form of the type handled in the proof of Proposition 11, and our sum is bounded by the first of our two expressions by Equation 4.

Otherwise, some $e_{i,j}$ is 1. Without loss of generality, this is $e_{n-1,n}$. We can also assume that $d_{n-1,n} = 0$ since adding or removing the appropriate term is equivalent the reversing the Legendre symbol. We split our sum into parts based on which of p_{n-1}, p_n are at least A. In particular we take the sum of terms with both at least A, plus the sum of terms where $p_{n-1} < A$ plus the sum of terms with $p_n < A$ minus the sum of terms with both less than A.

First consider the case where $p_{n-1}, p_n \geq A$. First fixing the values of $p_1, \ldots, p_{n-2}$, and letting $P = \prod_{i=1}^{n-2} p_i$, we consider the remaining sum over p_{n-1} and p_n. We have

$$\frac{\pm 1}{n!}\sum_{\substack{A\leq p_{n-1},p_n\\ p_{n-1}\neq p_n\\ (p_i,DP)=1\\ Q\not| DPp_{n-1}p_n\\ p_{n-1}p_n\leq N/P}} a(p_{n-1})b(p_n)\left(\frac{p_{n-1}}{p_n}\right).$$

Where a, b are some functions $\mathbb{Z} \to \mathbb{C}$ so that $|a(x)|, |b(x)| \leq 1$ for all x. We note that the condition that $(p_i, DP) = 1$ can be expressed by setting a and b equal to 0 for some

appropriate set of primes. We note that the condition that $q(X,Y)$ not divide $DPp_{n-1}p_n$ is only relevant if DP is missing only one or two primes of $q(X,Y)$. In the former case, it is equivalent to making one more value illegal for the p_i. In the latter case it eliminates at most two terms. The condition that the p_i are distinct removes at most $\sqrt{N/P}$ terms from our sum. Therefore, perhaps after setting a and b to 0 on some set of primes, the above is

$$\frac{\pm 1}{n!}\left(O(\sqrt{N/P}) + \sum_{\substack{A\le p_{n-1},p_n \\ p_{n-1}p_n\le N/P}} a(p_{n-1})b(p_n)\left(\frac{p_{n-1}}{p_n}\right)\right).$$

By Lemma 15, this is at most

$$\frac{1}{n!}O(N/P\log(N)A^{-1/8}).$$

Now for each $P \le N$, it can be written in at most $(n-2)!$ ways, hence the sum over all $p_{n-1}, p_n \ge A$ is at most

$$\sum_{P=1}^{N} O(N/P\log(N)A^{-1/8}) = O(N\log^2(N)A^{-1/8}).$$

Next for the case where $p_n < A$. We deal with this case by setting p_n to each possible value of size at most A individually. It is easy to check that after setting p_n to such a value p, the sum over the remaining p_i is $1/n$ times a sum of the form bounded by $Q(n-1, N, Dp, k-1, X, Y, S\cup\{p\})$. Hence the sum over all terms with $p_n < A$ is at most

$$\frac{1}{n}\sum_{p<A} Q(n-1, N/p, Dp, k-1, X, Y, S\cup\{p\}).$$

The sum of the terms with $p_{n-1} < A$ has the same bound, and the sum of terms with both less than A is similarly seen to be at most

$$\frac{1}{n(n-1)}\sum_{p_1,p_2<A} Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S\cup\{p_1,p_2\}).$$

□

We use this to prove an inductive bound on Q

Lemma 17. *Let $n, N, D, k, X, Y, S, M, A, B, C, \beta$ be as above. Assume furthermore that $Y \ge DA^n$,*

$$B > \max(e^{(C+2)\beta\log\log M}, e^{K(C+2)^2(\log Y)^2(\log\log(YM))^2}, n\log^{C+2}(M), A),$$

and that S contains only elements of size at most A. Let $L = n-k$, then $Q(n, N, D, k, X, Y, S)$ is at most

$$N\Bigg(O\left(\frac{\log\log B}{L}\right)^k + O\left(\log^2(N)A^{-1/8} + \log(N)\log^{-C-2}(M)\right)\sum_{a=0}^{k-1} O\left(\frac{\log\log B}{L}\right)^a\Bigg).$$

Note that we will wish to apply this Lemma with n about $\log\log N$, D a constant, A polylog N, X polylog N, $M = N$, $Y = DA^n$, and B its minimum possible value.

Proof. We proceed by induction on k, using Lemma 16 to bound Q. Our base case is when $Q(n, N, D, k, X, Y, S)$ is bounded by

$$N\left(O\left(\frac{\log\log B}{n}\right)^k + O(\log(N)\log^{-C-2}(M))\sum_{a=0}^{k-1} O\left(\frac{\log\log B}{n}\right)^a\right)$$

(which must happen if $k = 0$). In this case, our bound clearly holds.

Otherwise, $Q(n, N, D, k, X, Y, S)$ is bounded by

$$O(N\log^2(N)A^{-1/8}) + \frac{2}{n}\sum_{p<A} Q(n-1, N/p, Dp, k-1, X, Y, S\cup\{p\}) + \frac{1}{n(n-1)}\sum_{p_1,p_2<A} Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S\cup\{p_1,p_2\}).$$

Notice that the parameters of Q in the above also satisfy our hypothesis, so we may bound them inductively. Note also that for the above values of Q that the value of L is the same. Letting U and E be sufficiently large multiples of $\frac{\log\log B}{L}$ and $\left(\log^2(N)A^{-1/8} + \log(N)\log^{-C-2}(M)\right)$, the above is easily seen to be at most

$$N\left(E + U\left(O(U)^{k-1} + E\sum_{a=0}^{k-2} O(U)^a\right) + U^2\left(O(U)^{k-2} + E\sum_{a=0}^{k-3} O(U)^a\right)\right).$$

Suppose that we wish to prove our final statement where the $O(U)$ terms hide a constant of Z. By the above, we could complete the inductive step if we could show that

$$E + U\left((ZU)^{k-1} + E\sum_{a=0}^{k-2}(ZU)^a\right) + U^2\left((ZU)^{k-2} + E\sum_{a=0}^{k-3}(ZU)^a\right)$$

was at most

$$(ZU)^k + E\sum_{a=0}^{k-1}(ZU)^a.$$

On the other hand, by comparing terms, this follows trivially as long as $Z > 2$. This completes our inductive step and proves our Lemma. □

We are finally prepared to prove Proposition 9.

Proof. The basic idea will be to compare the sum in question to the quantity $Q(n, N, D, k, X, Y, \emptyset)$ for appropriate settings of the parameters. We begin by fixing the constant c in the Proposition statement. We let C be a constant large enough that $c^n > \log^{-C}(N)$ (recall that n was $O(\log\log N)$). We set A to $\log^{8C+16}(N)$, X to $\log^C(N)$ and Y to $DA^n = \exp(O_D(C(\log\log N)^2))$. We let $M = N$.

We note that β comes from either the worst Siegel zero of modulus less that X, or the second worst Siegel zero of modulus less than Y. By Theorem 5.28 of [24], β is at most $O_\epsilon(X^\epsilon)$ in the former case, and at most $O(\log(Y))$ in the latter case. Hence (changing ϵ by a factor of C), we have unconditionally, that $\beta = O_\epsilon(\log^\epsilon(N))$ for any $\epsilon > 0$. We next let

$$B = \max(e^{(C+2)\beta\log\log M}, e^{K(C+2)^2(\log Y)^2(\log\log(YM))^2}, n\log^{C+2}(M), A).$$

Hence for sufficiently large N (in terms of epsilon and D),

$$\log\log B < \epsilon\log\log N.$$

Finally we pick k so that $n/2 \geq k \geq m/2$. Thus $L = n - k > n/2 = \Omega(\log\log N)$. Noting that we satisfy the hypothesis of Lemma 16, we have that for N sufficiently large relative to ϵ and D that $Q(n, N, D, k, X, Y, \emptyset)$ is at most

$$N\left(O(\epsilon)^{m/2} + O\left(\log^2(N)\log^{-C-2}(N) + \log(N)\log^{-C-1}(N)\right)\sum_{a=0}^{k} O(\epsilon)^a\right).$$

If ϵ is small enough that the term $O(\epsilon)$ is at most $1/2$, this is at most

$$N\left(O(\epsilon)^{m/2} + \log^{-C}(N)\right).$$

If additionally the $O(\epsilon)$ term is less than c^2, this is

$$O(Nc^m).$$

Hence for N sufficiently large relative to c and D,

$$Q(n, N, D, k, X, Y, \emptyset) = O(Nc^m).$$

Therefore unequivocally,

$$Q(n, N, D, k, X, Y, \emptyset) = O_{c,D}(Nc^m).$$

Finally, we note that the difference between $Q(n, N, D, k, X, Y, \emptyset)$ and the term that we are trying to bound is exactly the sum over such terms where $p_1 \cdots p_n$ is divisible by $\frac{q(X,Y)}{gcd(q(X,Y),D)}$. Since $q(X, Y) \geq X$, there are only $O_D(N \log^{-C}(N))$ such products. Since each product can be obtained in at most $n!$ ways, each contributing at most $\frac{1}{n!}$, this difference is at most $O_D(N \log^{-C}(N)) = O(Nc^m)$. Therefore the thing we wish to bound is $O_{c,D}(Nc^m)$. □

4 Average Sizes of Selmer Groups

Here we use the results from the previous section to prove the following Proposition:

Proposition 18. *Let E be an elliptic curve as described above (full two torsion defined over $\mathbb{Q}$, no cyclic 4-isogeny defined over $\mathbb{Q}$). Let S be a finite set of places containing $2, \infty$ and all of the places where E has bad reduction. Let x be either -1 or a power of 2. Let $\omega(m)$ denote the number of prime factors of m. Say that $(m, S) = 1$ if m is an integer not divisible by any of the finite places in S. For positive integers N, let $\mathcal{S}_N$ denote the set of integers $b \leq N$ squarefree with $|\omega(b) - \log\log N| \leq (\log\log N)^{3/4}$, and $(b, S) = 1$. Then*

$$\lim_{N\to\infty} \frac{\sum_{\mathcal{S}_N} x^{\dim(S_2(E_b))}}{|\mathcal{S}_N|} = \sum_n x^n \alpha_n.$$

This says that the k^{th} moment of $|S_2(E_b)|$ averaged over $b \leq N$ with $|\omega(b) - \log\log N| \leq (\log\log N)^{3/4}$ is what you would expect it to be by Theorem 2, and that averaged over the same b's that the rank of the Selmer group is odd half of the time. The latter part of the Proposition follows from Lemma 1, which we prove now:

Proof of Lemma 1. First we replace E by a twist so that $c_i - c_j$ are pairwise relatively prime integers. It is now the case that E has everywhere good or multiplicative reduction, and we are now concerned with $\dim(S_2(E_{db}))$ for some constant $d|D$. By [44] Theorem 2.3, and [38] Corollary 1 we have that $\dim(S_2(E_{bd})) \equiv \dim(S_2(E))$ (mod 2) if and only if $(-1)^x \chi_{bd}(-N) = 1$ where $x = \omega(d)$, N is the product of the primes not dividing d at which E has bad reduction, and χ_{bd} is the quadratic character corresponding to the extension $\mathbb{Q}(\sqrt{bd})$. From this, the Lemma follows immediately. □

In order to prove the rest of Proposition 18, we will need a concrete description of the Selmer groups of twists of E. We follow the treatment given in [58]. Let $b = p_1 \cdots p_n$ where p_i are distinct primes relatively prime to S (we leave which primes unspecified for now). Let

$B = S \cup \{p_1, \ldots, p_n\}$. For $\nu \in B$ let V_ν be the subspace of $(u_1, u_2, u_3) \in (\mathbb{Q}_\nu^*/(\mathbb{Q}_\nu^*)^2)^3$ so that $u_1u_2u_3 = 1$. Note that V_ν has a symplectic form given by $(u_1, u_2, u_3) \cdot (v_1, v_2, v_3) = \prod_{i=1}^3 (u_i, v_i)_\nu$, where $(u_i, v_i)_\nu$ is the Hilbert Symbol. Let $V = \prod_{\nu \in B} V_\nu$ be a symplectic $\mathbb{F}_2$-vector space of dimension $2M$.

There are two important Lagrangian subspaces of V. The first, which we call U, is the image in V of $(\mathbb{Z}_B^*/(\mathbb{Z}_B^*)^2)_1^3$. The other, which we call W, is given as the product of W_ν over $\nu \in B$, where W_ν consists of points of the form $(x - bc_1, x - bc_2, x - bc_3)$ for $(x, y) \in E_b$. Note that we can write $W = W_S \times W_b$ where $W_S = \prod_{\nu \in S} W_\nu$ and $W_b = \prod_{\nu | b} W_\nu$. The Selmer group is given by

$$S_2(E_b) = U \cap W.$$

Let U' be the $\mathbb{F}_2$-vector space generated by the symbols ν, ν' for $\nu \in S$ and p_i, p_i' for $1 \leq i \leq n$. There is an isomorphism $f : U' \to U$ given by $f(\infty) = (-1, -1, 1), f(\infty') = (1, -1, -1), f(p) = (p, p, 1), f(p') = (1, p, p)$.

Note also that W_{p_i} is generated by $((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3))$ and $(b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2))$. If we define W' to be the $\mathbb{F}_2$-vector space generated by the symbols p_i, p_i' for $1 \leq i \leq n$, then there is an isomorphism $g : W' \to W_b$ given by $g(p_i) = ((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3)) \in W_{p_i}$ and $g(p_i') = (b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2)) \in W_{p_i}$.

Let $G = \prod_{\nu \in S \setminus \infty} \mathfrak{o}_\nu^*/(\mathfrak{o}_\nu^*)^2$ (here $\mathfrak{o}_\nu^*$ are the units in the ring of integers of k_ν). Note that if b is positive, W_S is determined by the restriction of b to G. So for $c \in G$ let $W_{S,c}$ be W_S for such b. Let $W_c' = W_{S,c} \times W'$. Then we have a natural map $g_c : W_c' \to V$ that is an isomorphism between W_c' and W if b restricts to c.

We are now ready to prove Proposition 18.

Proof. For $x = -1$ this Proposition just says that the parity is odd half of the time, which follows from Lemma 1. For $x = 2^k$ this says something about the expected value of $|S_2(E_b)|^k$. For $x = 2^k$ we will show that for each $n \in (\log\log N - (\log\log N)^{3/4}, \log\log N + (\log\log N)^{3/4})$ that

$$\sum_{\substack{b \leq N \\ b \text{ square-free} \\ \omega(b) = n \\ (b,S)=1}} |S_2(E_b)|^k$$

$$= \left(\sum_{\substack{b \leq N \\ b \text{ square-free} \\ \omega(b) = n \\ (b,S)=1}} 1 \right) \left(\sum_m \alpha_m (2^k)^m + \delta(n, N) \right) + O_{E,k}\left(\frac{N(\log\log\log N)^2}{\log\log N} \right).$$

Where $\delta(n,N)$ is some function so that $\lim_{N\to\infty}\delta(n,N)=0$. Summing over n and noting that there are $\Omega(N)$ values of $b\le N$ square-free with $|\omega(b)-\log\log N|<(\log\log N)^{3/4}$, and $(b,S)=1$ gives us our desired result.

In order to do this we need to better understand $|S_2(E_b)|=|U\cap W|$. For $v\in V$ we have since U is Lagrangian of size 2^M,

$$\frac{1}{2^M}\sum_{u\in U}(-1)^{u\cdot v}=\begin{cases}1 \text{ if } v\in U^\perp\\ 0 \text{ else}\end{cases}$$
$$=\begin{cases}1 \text{ if } v\in U\\ 0 \text{ else}\end{cases}.$$

Hence

$$\begin{aligned}|S_2(E_b)|&=|U\cap W|\\ &=\#\{w\in W: w\in U\}\\ &=\sum_{w\in W}\frac{1}{2^M}\sum_{u\in U}(-1)^{u\cdot w}\\ &=\frac{1}{2^M}\sum_{u\in U,w\in W}(-1)^{u\cdot w}\\ &=\frac{1}{2^M}\sum_{u\in U',w\in W_b'}(-1)^{f(u)\cdot g_b(w)}\\ &=\frac{1}{2^M}\sum_{c\in G}\frac{1}{|G|}\sum_{\chi\in\widehat{G}}\chi(bc)\sum_{u\in U',w\in W_c'}(-1)^{f(u)\cdot g_c(w)}\\ &=\frac{1}{2^M|G|}\sum_{\substack{c\in G,\chi\in\widehat{G}\\ u\in U',w\in W_c'}}\chi(bc)(-1)^{f(u)\cdot g_c(w)}.\end{aligned}$$

If we extend f and g_c to $f^k:(U')^k\to U^k$, $g_c^k:(W_c')^k\to V^k$, and extend the inner product on V to an inner product on V^k, we have that

$$|S_2(E_b)|^k=\frac{1}{2^{kM}|G|^k}\sum_{\substack{c\in G,\chi\in\widehat{G}\\ u\in(U')^k,w\in(W_c')^k}}\chi(bc)(-1)^{f^k(u)\cdot g_c^k(w)}. \tag{7}$$

Notice that once we fix values of c,χ,u,w in Equation 7 the summand (when treated as a function of $p_1,\dots,p_n$) is of the same form as the "characters" studied in Section 3.

We want to take the sum over all $b\le N$ square-free, $\omega(b)=n$, $(b,S)=1$, of $|S_2(E_b)|^k$. If we let D be 8 times the product of the finite odd primes in S, we note that each such b can

be expressed exactly $n!$ ways as a product $p_1, \ldots, p_n$ with p_i distinct, $(p_i, D) = 1$. Therefore this sum equals

$$\frac{1}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} \frac{1}{2^{kM}|G|^k} \sum_{\substack{c \in G, \chi \in \widehat{G} \\ u \in (U')^k, w \in (W'_c)^k}} \prod_i \chi(p_i)\chi(c)(-1)^{f^k(u)\cdot g_c^k(w)}.$$

Interchanging the order of summation gives us

$$\frac{1}{2^{kM}|G|^k} \sum_{\substack{c \in G, \chi \in \widehat{G} \\ u \in (U')^k, w \in (W'_c)^k}} \frac{\chi(c)}{n!} \sum_{\substack{p_1,\ldots,p_n \\ \text{distinct primes} \\ (D,p_i)=1 \\ \prod_i p_i \leq N}} \left(\prod_i \chi(p_i) \right) (-1)^{f^k(u)\cdot g_c^k(w)}.$$

Now the inner sum is exactly of the form studied in Proposition 9.

We first wish to bound the contribution from terms where this inner sum has terms of the form $\left(\frac{p_i}{p_j}\right)$, or in the terminology of Proposition 9 for which not all of the $e_{i,j}$ are 0. In order to do this, we will need to determine how many of these terms there are and how large their values of m are. Notice that terms of the form $\left(\frac{p_i}{p_j}\right)$ show up here when we are evaluating the Hilbert symbols of the form $(p, b(c_a - c_b))_p, (p, b(c_a - c_b))_q, (q, b(c_a - c_b))_p, (q, b(c_a - c_b))_q$ and in no other places.

Let $U_i \subset U'$ be the subspace generated by $p_i = (p_i, p_i, 1)$ and $p'_i = (1, p_i, p_1)$. For $u \in U'$ let u_i be its component in U_i in the obvious way. Let $W_i \subset W'$ be W_{p_i}. For $w \in W'_c$ let w_i be its component in W_i. Let U_0 be the $\mathbb{F}_2$-vector space with formal generators p and p'. We have a natural isomorphism between U_0 and U_i sending p to p_i and p' to p'_i. We will hence often think of u_i as an element of U_0. Similarly let W_0 be the $\mathbb{F}_2$-vector space with formal generators $((c_1-c_2)(c_1-c_3), b(c_1-c_2), b(c_1-c_3))$ and $(b(c_3-c_1), b(c_3-c_2), (c_3-c_1)(c_3-c_2))$. We similarly have natural isomorphisms between W_i and W_0 and will often consider w_i as an element of W_0 instead of W_i.

Additionally, we have a bilinear form $U_0 \times W_0 \to \mathbb{F}_2$ defined by:

$$\begin{aligned}
&p\cdot((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3)) \\
&\quad = p' \cdot (b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2)) \\
&\quad = 1 \\
&p'\cdot((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3)) \\
&\quad = p \cdot (b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2)) \\
&\quad = 0.
\end{aligned}$$

We notice that if $u \in U'$ and $w \in W_c'$, then the exponent of $\left(\frac{p_i}{p_j}\right)$ that appears in $(-1)^{f(u)\cdot g_c(w)}$ is $(u_i + u_j) \cdot (w_i + w_j)$. Similarly if $u \in (U')^k, w \in (W_c')^k$, the exponent of $\left(\frac{p_i}{p_j}\right)$ that appears in $(-1)^{f^k(u)\cdot g_c^k(w)}$ is $(u_i + u_j) \cdot (w_i + w_j)$, where u_*, w_* are thought of as elements of U_0^k and W_0^k.

We define a symplectic form on $T = U_0^k \times W_0^k$ by $(u, w) \cdot (u', w') = u \cdot w' + u' \cdot w$. Also define a quadratic form q on T by $q(u, w) = u \cdot w$. We claim that given some set of $t_i = (u_i, w_i)_{i\in I} \in T$ that $(u_i + u_j) \cdot (w_i + w_j) = 0$ for all pairs $i, j \in I$ only if all of the (u_i, w_i) lie in a translate of a Lagrangian subspace of T. First note that for $t = (u, w), t' = (u', w')$ that $(u + u') \cdot (w + w') = t \cdot t' + q(t) + q(t')$. We need to show that for all $i, j, k \in I$ that $(t_i + t_j) \cdot (t_i + t_k) = 0$. This is true because

$$\begin{aligned}
&(t_i + t_j) \cdot (t_i + t_k) \\
&\quad = t_i \cdot t_i + t_i \cdot t_k + t_j \cdot t_i + t_j \cdot t_k \\
&\quad = t_i \cdot t_k + t_j \cdot t_i + t_j \cdot t_k \\
&\quad = t_i \cdot t_k + t_j \cdot t_i + t_j \cdot t_k + 2q(t_i) + 2q(t_j) + 2q(t_k) \\
&\quad = (t_i \cdot t_j + q(t_i) + q(t_j)) + (t_i \cdot t_k + q(t_i) + q(t_k)) + (t_k \cdot t_j + q(t_k) + q(t_j)) \\
&\quad = 0.
\end{aligned}$$

So suppose that we have some $u \in \prod_{i=1}^n U_i^k$ and $w \in \prod_{i=1}^n W_i^k$, and suppose that we have a set of ℓ indices in $\{1, 2, \ldots, n\}$, which we call *active* indices, so that $(-1)^{f^k(u)\cdot g^k(w)}$ has terms of the form $\left(\frac{p_i}{p_j}\right)$ only if i, j are both active, and suppose furthermore that each active index shows up as either i or j in at least one such term. Let $t_i = (u_i, w_i) \in T$. We claim that t_i takes fewer than 2^k different values on non-active indices, i. We note that our notion of active indices is similar to the notion in [22] of linked indices.

Since $t_i \cdot t_j + q(t_i) + q(t_j) = 0$ for any two non-active indices t_i and t_j, all of these must lie in a translate of some Lagrangian subspace of T. Therefore t_i can take at most 2^k values on non-active indices. Suppose for sake of contradiction that all of these values are actually assumed by some non-active index. Then consider t_j for j an active index. The t_i for i either non-active or equal to j must similarly lie in a translate of a Lagrangian subspace. Since such a space is already determined by the non-active indices and since all elements of this affine subspace are already occupied, t_j must equal t_i for some non-active i. But this means that every t_j is assumed by some non-active index which implies that no terms of the form $\left(\frac{p_i}{p_j}\right)$ survive, yielding a contradiction.

Now consider the number of such u, w so that there are at most ℓ active indices and so that at least one of these terms survive. Once we fix the values t_i that are allowed to be taken by the non-active indices (which can only be done in finitely many ways), there are $\binom{n}{\ell}$ ways to choose the active indices, at most $2^k - 1$ ways to pick t_i for each non-active

index, and at most 2^{2k} ways for each active index. Hence the total number of such u,w with exactly ℓ active indices is

$$O\left(\binom{n}{\ell}\left(2^k-1\right)^{n-\ell}\left(2^{2k}\right)^{\ell}\right).$$

By Proposition 9, the value of the inner sum for such a (u,w) is at most $O_{E,k}\left(N(2^{-2k-1})^{\ell}\right)$. Hence summing over all $\ell > 0$ and recalling the 2^{-Mk} out front we get a contribution of at most

$$\begin{aligned} N2^{-nk}O_{E,k}\left(\sum_{\ell}\binom{n}{\ell}\left(2^k-1\right)^{n-\ell}\left(\frac{1}{2}\right)^{\ell}\right) &= N2^{-nk}O_{E,k}\left((2^k-1/2)^n\right) \\ &= NO_{E,k}\left((1-2^{-k-1})^n\right) \\ &= NO_{E,k}\left((\log N)^{-2^{-k-2}}\right). \end{aligned}$$

Therefore we may safely ignore all of the terms in which a $\left(\frac{p_i}{p_j}\right)$ shows up. This is our analogue of Lemma 6 in [22].

Notice that by the above analysis, that the number of remaining terms must be $O(2^{Mk})$. Additionally, for these terms we may apply Proposition 10. Therefore because of the 2^{-Mk} factor out front we have that up to an error of $O_E\left(\frac{(\log\log\log N)^2}{\log\log N}\right)$ that the sum over $b = p_1\cdots p_n$, square free, at most N, relatively prime to D, of $|S_2(E_b)|^k$ is the number of such b times the average over all possible values of $p_i \in G$, $\left(\frac{p_i}{p_j}\right)$, of $|S_2(E_b)|^k$. Furthermore, our work shows that this average is bounded in terms of k and E independently of n.

On the other hand, recalling the notation from Theorem 2, this average is just

$$\sum_d \pi_d(n)2^{kd}.$$

Using the fact that this is bounded for $k+1$ independently of n, we find that $\pi_d(n) = O_{k,E}(2^{-(k+1)d})$. In order to complete the proof of our Proposition, we need to show that

$$\lim_{n\to\infty}\sum_d(\pi_d(n)-\alpha_d)2^{kd} = 0.$$

But this follows from the fact that

$$\sum_{d>X}(\pi_d(n)-\alpha_d)2^{kd} = O_{E,k}\left(\sum_{d>X}2^{-d}\right) = O_{E,k}(2^{-X})$$

and that $\pi_d(n)\to\alpha_d$ for all d by Theorem 2.

□

5 From Sizes to Ranks

In this Section, we turn Proposition 18 into a proof of Theorem 3. This Section is analogous to Section 8 of [22], although our techniques are significantly different. We begin by doing some computations with the α_i.

Note that

$$\alpha_{n+2} = \left(\frac{1}{\prod_{j=0}^{\infty}(1+2^{-j})} \right) 2^{-\binom{n}{2}} \prod_{j=1}^{n} (1-2^{-j})^{-1}.$$

Now $\prod_{j=1}^{n}(1-2^{-j})^{-1}$ is the sum over partitions, P, into parts of size at most n of $2^{-|P|}$. Equivalently, taking the transpose, it is the sum over partitions P with at most n parts of $2^{-|P|}$. Multiplying by $2^{-\binom{n}{2}}$, we get the sum over partitions P with n distinct parts (possibly a part of size 0) of $2^{-|P|}$. Therefore, we have that

$$F(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \frac{x^2 \prod_{j=0}^{\infty}(1+2^{-j}x)}{\prod_{j=0}^{\infty}(1+2^{-j})}.$$

Since the x^{d+2} coefficient of $F(x)$ is also the sum over partitions, P into exactly d distinct parts (perhaps one of which is 0) of $2-|P|$ divided by $\prod_{j=0}^{\infty}(1+2^{-j})$. This implies in particular that $\sum_{n=0}^{\infty} \alpha_n$ equals 1 as it should.

Let T_N be the set of square-free $b \leq N$ with $(b, D) = 1$, and $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$. Let $C_d(N)$ be

$$\frac{\#\{b \in T_N : \dim(S_2(E_b)) = d\}}{|T_N|}.$$

Let $C(N) = (C_0(N), C_1(N), \ldots) \in [0,1]^{\omega}$. Theorem 3 is equivalent to showing that

$$\lim_{N\to\infty} C(N) = (\alpha_0, \alpha_1, \ldots).$$

Lemma 19. *Suppose that some subsequence of the $C(N)$ converges to some sequence $(\beta_0, \beta_1, \ldots) \in [0,1]^{\omega}$ in the product topology. Let $G(x) = \sum_n \beta_n x^n$. Then $G(x)$ has infinite radius of convergence and $F(x) = G(x)$ for $x = -1$ or x equals a power of 2. Also $\beta_0 = \beta_1 = 0$.*

This Lemma says that if the $C(N)$ have some limit that the naive attempt to compute moments of the Selmer groups from this limit would succeed.

Proof. The last claim follows from the fact that since E_b has full 2-torsion, its 2-Selmer group always has rank at least 2. Notice that $\sum_d C_d(N) x^d$ is equal to the average size of $x^{\dim(S_2(E_b))}$ over $b \leq N$ square-free, relatively prime to D with $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$. This

has limit $F(x)$ as $N \to \infty$ by Proposition 18 if x is -1 or a power of 2. In particular it is bounded. Therefore there exists an R_k so that

$$\sum_d C_d(N) 2^{kd} \leq R_k$$

for all N. Therefore $C_d(N) \leq R_k 2^{-kd}$ for all d, N. Therefore $\beta_d \leq R_k 2^{-kd}$. Therefore G has infinite radius of convergence.

Furthermore if we pick a subsequence, $N_i \to \infty$ so that $C_d(N_i) \to \beta_d$ for all d, we have that

$$\begin{aligned} F(2^k) &= \lim_{i\to\infty} \sum_d C_d(N_i) 2^{dk} \\ &= \lim_{i\to\infty} \sum_{d\leq X} C_d(N_i) 2^{dk} + O\left(\sum_{d>X} R_{k+1} 2^{-d}\right) \\ &= \lim_{i\to\infty} \sum_{d\leq X} C_d(N_i) 2^{dk} + O(R_{k+1} 2^{-X}) \\ &= \sum_{d\leq X} \beta_d 2^{dk} + O(R_{k+1} 2^{-X}). \end{aligned}$$

So

$$\lim_{X\to\infty} \sum_{d\leq X} \beta_d 2^{dk} = F(2^k).$$

Thus $G(2^k) = F(2^k)$. For $x = -1$ the argument is similar but comes from the equidistribution of parity rather than expectation of size. □

Lemma 20. *Suppose that $G(x) = \sum_n \beta_n x^n$ is a Taylor series with infinite radius of convergence. Suppose also that $\beta_n \in [0,1]$ for all n and that $G(x) = F(x)$ for x equal to -1 or a power of 2. Suppose also that $\beta_0 = \beta_1 = 0$. Then $\beta_n = \alpha_n$ for all n.*

Proof. First we wish to prove a bound on the size of the coefficients of G. Note that

$$F(2^k) = \frac{2^{2k}(1+2^k)(1+2^{k-1})\cdots}{(1+2^0)(1+2^{-1})\cdots} = 2^{2k} \prod_{j=1}^{k} (1+2^k) = O\left(2^{2k+k(k+1)/2}\right).$$

Now

$$2^{nk}\beta_n \leq G(2^k) = F(2^k) = O\left(2^{2k+k(k+1)/2}\right).$$

Therefore

$$\beta_n = O\left(2^{2k+k(k+1)/2-kn}\right).$$

Setting $k = n$ we find that

$$\beta_n = O\left(2^{-n^2/2+5n/2}\right) = O\left(2^{-\binom{n-2}{2}}\right).$$

The same can be said for F. Now consider $F - G$. This is an entire function whose x^n coefficient is bounded by $O\left(2^{-\binom{n-2}{2}}\right)$. Furthermore $F - G$ vanishes to order at least 2 at 0, and order at least 1 at -1 and at powers of 2. The bounds on coefficients imply that

$$|F(x) - G(x)| \leq O\left(\sum_n 2^{-\binom{n-2}{2}}|x|^n\right).$$

The terms in the above sum clearly decay rapidly for n on either side of $\log_2(|x|)$. Hence

$$\begin{aligned}|F(x) - G(x)| &= O\left(2^{(-\log_2(|x|)^2+5\log_2(|x|))/2+\log_2(|x|)^2}\right)\\ &= O\left(2^{(\log_2(|x|)^2+5\log_2(|x|))/2}\right).\end{aligned}$$

In particular $F - G$ is a function of order less than 1. Hence it must equal

$$Cx^{2+t}\prod_\rho(1 - x/\rho),$$

where the product is over non-zero roots ρ of $F - G$, and t is some non-negative integer. On the other hand, Jensen's Theorem tells us that if $C \neq 0$ the average value of $\log_2(|F - G|)$ on a circle of radius R is

$$\log_2|C| + (2 + t)\log_2 R + \sum_{|\rho|<R}\log_2(R/|\rho|).$$

Setting $R = 2^k$ and noting the contributions from $\rho = -1$ and $\rho = 2^j$ for $j < k$ we have

$$O(1) + 3k + \sum_{j<k}(k - j) = O(1) + 3k + \binom{k+1}{2} = O(1) + \frac{k^2 + 7k}{2} > \frac{k^2 + 5k}{2}$$

which is larger than $\log_2(|F - G|)$ can be at this radius. This provides a contradiction. □

We now prove Theorem 3.

Proof. Suppose that $C(N)$ does not have limit $(\alpha_0, \alpha_1, \ldots)$. Then there is some subsequence N_i so that $C(N_i)$ avoid some neighborhood of $(\alpha_0, \alpha_1, \ldots)$. By compactness, $C(N_i)$ must have some subsequence with a limit $(\beta_0, \beta_1, \ldots)$. By Lemmas 19 and 20, $(\alpha_0, \alpha_1, \ldots) = (\beta_0, \beta_1, \ldots)$. This is a contradiction.

Therefore $\lim_{N\to\infty} C(N) = (\alpha_0, \alpha_1, \ldots)$. Hence $\lim_{N\to\infty} C_d(N) = \alpha_d$ for all d. The Theorem follows immediately from this and the fact the fraction of $b \leq N$ square-free with $(b, D) = 1$ that have $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$ approaches 1 as $N \to \infty$. □

It should be noted that our bounds on the rate of convergence in Theorem 3 are non-effective in two places. One is our treatment in this last Section. We assume that we do not have an appropriate limit and proceed to find a contradiction. This is not a serious obstacle and if techniques similar to those of [22] were used instead, it could be overcome. The more serious problem comes in our proof of Proposition 9, where we make use of non-effective bounds on the size of Siegel zeroes. This latter problem may well be fundamental to our approach.

Chapter D

Projective Embeddings of the Deligne-Lusztig Curves

1 Introduction

There are four (twisted) Chevalley groups of rank 1. The associated Deligne-Lusztig varieties for the Coxeter classes of these groups all give affine algebraic curves. The completions of these curves have several applications including the representation theory of the associated group ([9, 42]), coding theory ([20]) and the construction of potentially interesting covers of $\mathbb{P}^1$ ([17]).

In this Chapter, we consider these curves associated to the groups $G = {}^2A_2, {}^2B_2$ and ${}^2G_2$. The remaining curve is associated to $G = A_1$ and is $\mathbb{P}^1$, but we do not cover this case as it is easy and doesn't follow many of the patterns found in the analysis of the other three cases. For each of these curves, we explicitly construct an embedding $C \hookrightarrow \mathbb{P}(W)$ where W is a representation of G of dimension 3,5, or 14 respectively, and provide an explicit system of equations cutting out C. The curve associated to ${}^2A_2$ is the Fermat curve. The curve associated to ${}^2B_2$ is also well-known though not immediately isomorphic to our embedding. As far as I know there is no standard embedding of the curve associated to ${}^2G_2$, although in [52] they construct an explicit curve with the correct genus, symmetry group and number points, which is likely the correct curve.

For each case, let d be the Coxeter number of the associated group, namely $d = 3, 4, 6$ for ${}^2A_2, {}^2B_2, {}^2G_2$, respectively. In [17], Gross proved that for each of these curves C, that $C/G^\sigma \cong \mathbb{P}^1$, with the corresponding map $C \to \mathbb{P}^1$ ramified over only three points corresponding to the images of the $\mathbb{F}_q$-points, the $\mathbb{F}_{q^d}$-points and the $\mathbb{F}_{q^{d+1}}$-points. In all cases, we write this map explicitly by finding the homogenous polynomials on C that correspond to the pullbacks of the degree-1 functions on $\mathbb{P}^1$ vanishing at each of these points, and demonstrating a linear relation between these functions.

In all cases we attempt to make our constructions canonical. We define the algebraic group G as the group of automorphisms of some vector space V preserving some additional structure. We define the Frobenius map on G by picking an isomorphism between V and some other space V' constructed functorially from V (for example, for ${}^2A_2$, we get the Frobenius map defining $SU(3)$ by picking a Frobenius-linear map between V and its dual). From V we construct W, another representation of G, given as a quotient of $\Lambda^2 V$. In each case we define our map $C \to \mathbb{P}(W)$ by sending a Borel, B, of G to the 2-form corresponding to the line in V that B fixes. The construction of the functions giving our map from C to $\mathbb{P}^1$ are all given by linear algebraic constructions.

There are a number of similarities in our techniques for the three different cases, suggesting that there may be a more general way to deal with all three at once, although we were unable to find such a technique. In addition to the similarity of overall approach, much of the feel of these constructions should be the same although they differ in the details. Additionally in all three cases we compute the degree of the embedding and find that it is given by $\frac{|G^\sigma|}{|B^\sigma||T^\sigma|}$ (here T is a Coxeter torus of G).

In Section 2, we describe some of the basic theory along with an outline of our general approach. In Section 3, we deal with the case of ${}^2A_2$; in Section 4, the case of ${}^2B_2$; and in Section 5, we deal with ${}^2G_2$. Much of the exposition in Section 2 is somewhat abstract and corresponds to relatively simple computations in Section 3, so if you are having trouble following in Section 2, it is suggested that you look at Section 3 in parallel to get a concrete example of what is going on.

2 Preliminaries

Here we provide an overview of the techniques and notation that will be common to our treatment of all three cases. In Section 2.1, we review the definition and basic theory of Deligne-Lusztig curves. In Section 2.2, we discuss some representations of G which will prove useful in our later constructions. In Section 2.3, we give a more complete overview of the common techniques to our different cases and describe a general category-theoretic construction that will provide us with the necessary Frobenius maps in each case. Finally in Section 2.4, we fix a couple of points of notation for the rest of the Chapter.

2.1 Basic Theory of the Deligne-Lusztig Curve

Let G be an algebraic group of type A_2, B_2 or G_2, defined over a finite field $\mathbb{F}_q$ with q equal to $q_0^2, 2q_0^2$ or $3q_0^2$ respectively. Let FR be the $\mathbb{F}_q$-Frobenius of G, and let σ be a Frobenius map so that FR $= \sigma^2$. We pick a an element w of the Weyl group of G of height 1. The Deligne-Lusztig variety is then defined to be the subvariety of the flag variety of G consisting of the Borel subgroups B so that B and $\sigma(B)$ are in relative position either w or 1 (actually

it is usually defined to be just the B where B and $\sigma(B)$ are in position w, but we use this definition, which constructs the completed curve). The resulting variety has an obvious G^σ action and in all of these cases is a smooth, complete algebraic curve.

In each of these cases, let B a Borel subgroup of G. Let T be a twisted Coxeter torus of G, that is a σ-invariant maximal torus such that the action of σ on T is conjugate to the action of $w\sigma$ on a split torus. Let d be $3, 4, 6$ for A_2, B_2, G_2 respectively. We will later make use of several facts about the points on the Deligne-Lusztig curve, C, defined over various fields and their behavior under the action of G^σ.

Here and throughout the rest of the Chapter we will use the phrase $\mathbb{F}_{q^n}$-point to mean a point defined over $\mathbb{F}_{q^n}$ but not defined over any smaller extension of $\mathbb{F}_q$. We make use of the following theorem of Lusztig:

Theorem 1. *1. G^σ acts transitively on the $\mathbb{F}_q$-points of C with stabilizer B^σ.*

2. C has no $\mathbb{F}_{q^n}$-points for $1 < n < d$.

3. G^σ acts transitively on the $\mathbb{F}_{q^d}$-points of C with stabilizer T^σ.

4. G^σ acts simply transitively on the $\mathbb{F}_{q^{d+1}}$-points of C.

5. G^σ acts freely on the $\mathbb{F}_{q^n}$-points of C for $n > d + 1$.

Recall that $C/G^\sigma \cong \mathbb{P}^1$ with the points of ramification given by the three points corresponding to the orbit of $\mathbb{F}_q$-points, the orbit of $\mathbb{F}_{q^d}$-points, and the orbit of $\mathbb{F}_{q^{d+1}}$-points.

We also make use of some basic counts (given in [17]). In particular, for A_2:

$$\begin{aligned} d &= 3 \\ |G^\sigma| &= q_0^3(q_0^3+1)(q-1) \\ |B^\sigma| &= q_0^3(q-1) \\ |T^\sigma| &= q - q_0 + 1 \\ \#\{\mathbb{F}_q - \text{points of } C\} &= q_0^3 + 1 \\ \#\{\mathbb{F}_{q^3} - \text{points of } C\} &= q_0^3(q_0+1)(q-1) \\ \#\{\mathbb{F}_{q^4} - \text{points of } C\} &= q_0^3(q_0^3+1)(q-1). \end{aligned}$$

For B_2:

$$\begin{aligned} d &= 4 \\ |G^\sigma| &= q^2(q^2+1)(q-1) \\ |B^\sigma| &= q^2(q-1) \\ |T^\sigma| &= q - 2q_0 + 1 \\ \#\{\mathbb{F}_q - \text{points of } C\} &= q^2+1 \\ \#\{\mathbb{F}_{q^4} - \text{points of } C\} &= q^2(q+2q_0+1)(q-1) \\ \#\{\mathbb{F}_{q^5} - \text{points of } C\} &= q^2(q^2+1)(q-1). \end{aligned}$$

For G_2:

$$\begin{aligned} d &= 6 \\ |G^\sigma| &= q^3(q^3+1)(q-1) \\ |B^\sigma| &= q^3(q-1) \\ |T^\sigma| &= q - 3q_0 + 1 \\ \#\{\mathbb{F}_q - \text{points of } C\} &= q^3+1 \\ \#\{\mathbb{F}_{q^6} - \text{points of } C\} &= q^3(q+3q_0+1)(q^2-1) \\ \#\{\mathbb{F}_{q^7} - \text{points of } C\} &= q^3(q^3+1)(q-1). \end{aligned}$$

2.2 Representation Theory

Let G be as above. Fix a Borel subgroup B. B is contained in two maximal parabolic subgroups, P_1 and P_2, corresponding to the short root and the long root respectively. There exists a representation V of G so that B fixes the complete flag $0 \subset L \subset M \subset S \subset \ldots \subset V$ and so that P_1 is the subgroup of G fixing L, and P_2 the subgroup fixing M. If G is A_2, B_2 or G_2, the dimension of V is $3, 4$ or 7 respectively.

Inside of $\Lambda^2 V$ is the representation W of G given by $W \subset \Lambda^2 V$ is the subrepresentation containing $\Lambda^2 M$. In our three cases, $\Lambda^2 V/W$ equals $0, 1$ or V respectively. If we are in any characteristic for A_2, characteristic 2 for B_2, or characteristic 3 for G_2, W has a quotient representation, V' of the same dimension as V so that the image of $G \to \mathrm{End}(V')$ is isomorphic to G. Picking an isomorphism between G and its image provides an endomorphism of G. If $G = A_2$, this is its outer automorphism. For G equal to B_2 or G_2 in characteristic 2 or 3 respectively, this endomorphism squares to the Frobenius endomorphism over the relevant prime field, giving a definition of σ.

2.3 Basic Techniques

Our basic techniques will be similar in all three cases and are as follows. For p a prime, let $q_0 = p^m$, and $q = p^{2m}$ or $q = p^{2m+1}$ as appropriate. Let $\mathcal{C}$ be the groupoid where an object of $\mathcal{C}$ is a representation of an algebraic group abstractly isomorphic to the representation V of G, and a morphism of $\mathcal{C}$ is an isomorphism of representations. In each case, we will reinterpret $\mathcal{C}$ as a groupoid whose objects are merely vector spaces with some additional structure (for example in the case of ${}^2A_2$ an object of $\mathcal{C}$ will be a three dimensional vector space with a volume form). This reinterpretation will allow us to work more concretely with $\mathcal{C}$ in each individual case, but as these structures are specific to which case we are in, we will ignore them for now.

In each case there will be two functors of interest from $\mathcal{C}$ to itself. The first is the Frobenius functor, $\mathrm{Fr} : \mathcal{C} \to \mathcal{C}$. This functor should be thought of as abstractly applying $\mathbb{F}_p$-Frobenius to each element. In particular there should be a natural Frobenius-linear transformation from V to $\mathrm{Fr}(V)$. In particular each of our $\mathcal{C}$ will have objects that are vector spaces, perhaps with some extra structure. As a vector space we will define $\mathrm{Fr}(V)$ as $\{[v] : v \in V\}$ where addition and multiplication are defined by $[v] + [w] = [v + w]$ and $k[v] = [\mathrm{Frobenius}^{-1}(k)v]$. The morphisms are unchanged by Fr. This should all be compatible with the extra structure accorded to an object of $\mathcal{C}$, except for inner products which must also be twisted by Frobenius. Note that giving a morphism $T : V \to \mathrm{Fr}^n(W)$ is the same as giving a $\mathrm{Frobenius}^{-n}$-linear map $S : V \to W$ that respects the additional $\mathcal{C}$-structure. Note therefore that a morphism $T : V \to \mathrm{Fr}^n(V)$ gives V an $\mathbb{F}_{p^n}$-structure. By abuse of notation, we will also use Fr to denote that natural Frobenius-linear map $V \to \mathrm{Fr}(V)$ defined by $v \to [v]$.

The other functor of importance is $' : \mathcal{C} \to \mathcal{C}$. This is the functor that takes V and gives V' as a subquotient of $\Lambda^2 V$. The exact construction of $'$ will vary from case to case. For ${}^2A_2$, V' will be the dual of V. It will be clear in all cases that there is a natural equivalence between $\mathrm{Fr}\circ' $ and $' \circ \mathrm{Fr}$.

Define a to be 0 or 1 so that $q = p^{2m+a}$. In each case, we will also find a natural transformation ρ from Fr^a to $''$ (again the details vary by case and we will not go into them here, though for ${}^2A_2$ it is the obvious isomorphism between a vector space and the dual of its dual). Lastly we pick a $V \in \mathcal{C}$ and a morphism $F : V' \to \mathrm{Fr}^{-m}(V)$.

This is enough to give a more explicit definition of σ. We construct a Frobenius map for $G = \mathrm{Aut}(V)$. We note the morphisms

$$\begin{aligned}
\mathrm{Fr}^{-a}(\rho_V) &: V \to \mathrm{Fr}^{-a}(V''), \\
\mathrm{Fr}^{-a}(F') &: \mathrm{Fr}^{-a}(V'') \to \mathrm{Fr}^{-m-a}(V'), \\
\mathrm{Fr}^{-m-a}(F) &: \mathrm{Fr}^{-m-a}(V') \to \mathrm{Fr}^{-2m-a}(V).
\end{aligned}$$

The composition

$$\mathfrak{F} := \mathrm{Fr}^{-m-a}(F) \circ (\mathrm{Fr}^{-a}(F')) \circ \mathrm{Fr}^{-a}(\rho_V) : V \to \mathrm{Fr}^{-2m-a}(V)$$

defines an $\mathbb{F}_q$-structure on V.

This gives us a Frobenius endomorphism $\mathrm{FR} : G \to G$ defined by the equation $\mathrm{Fr}^{-2m-a}(\mathrm{FR}(g))\circ \mathfrak{F} = \mathfrak{F} \circ g$. We define $\sigma : G \to G$ by the equation $\mathrm{Fr}^{-m}(\sigma(g)) \circ F = F \circ g'$. We have that:

$$\begin{aligned}
\sigma(\sigma(g)) &= \mathrm{Fr}^m(F) \circ \mathrm{Fr}^m(\sigma(g')) \circ \mathrm{Fr}^m(F^{-1}) \\
&= \mathrm{Fr}^m(F) \circ \mathrm{Fr}^{2m}(F') \circ \mathrm{Fr}^{2m}(g'') \circ \mathrm{Fr}^{2m}(F'^{-1}) \circ \mathrm{Fr}^m(F^{-1}) \\
&= \mathrm{Fr}^{2m+a}(\mathfrak{F} \circ \mathrm{Fr}^{-a}(\rho_V^{-1}) \circ \mathrm{Fr}^{-a}(g'') \circ \mathrm{Fr}^{-a}(\rho_V) \circ \mathfrak{F}^{-1}) \\
&= \mathrm{Fr}^{2m+a}(\mathfrak{F} \circ \mathrm{Fr}^{-a}(\rho_V^{-1} \circ g'' \circ \rho_V) \circ \mathfrak{F}^{-1}) \\
&= \mathrm{Fr}^{2m+a}(\mathfrak{F} \circ \mathrm{Fr}^{-a}(\mathrm{Fr}^a(g)) \circ \mathfrak{F}^{-1}) \\
&= \mathrm{Fr}^{2m+a}(\mathfrak{F} \circ g \circ \mathfrak{F}^{-1}) \\
&= \mathrm{FR}(g).
\end{aligned}$$

The third to last step above comes from the fact that ρ is a natural transformation. This gives us an endomorphism $\sigma : G \to G$ so that $\sigma^2 = \mathrm{FR}$.

We note that this technique for defining σ works most conveniently when by "algebraic group" we mean "group object in the category of varieties over $\overline{\mathbb{F}_q}$", since then G can be associated with $\mathrm{Aut}_{\mathcal{C}}(V)$, and we have an action of σ on G. On the other hand if you want "algebraic group" to mean "group object in the category of schemes", then the same technique should still work as long as we consider $\mathcal{C}$ as a category enriched in schemes.

We may pick our element w so that two Borels of G are in relative position w if they fix the same line in V. We then define a projective embedding $C \hookrightarrow \mathbb{P}(W)$ sending a Borel B to the two-form defined by the plane it fixes in V. In each case we will provide explicit polynomials that cut out the image of C. We will compute the degree of this embedding by finding a polynomial that vanishes exactly at the $\mathbb{F}_q$-points of C. In each case this degree will be $\frac{|G^\sigma|}{|B^\sigma||T^\sigma|}$. In each case for each $\mathbb{F}_q$-point, there is a hyperplane that intersects C only at that point but with large multiplicity. We also compute polynomials that cut out the $\mathbb{F}_{q^d}$-points and the $\mathbb{F}_{q^{d+1}}$-points. Lastly we find a linear relation between appropriate powers of these polynomials.

2.4 Notes

Throughout this Chapter, by an $\mathbb{F}_{q^n}$-point of a curve defined over $\mathbb{F}_q$, we will mean a point defined over $\mathbb{F}_{q^n}$ but not over any smaller extension of $\mathbb{F}_q$. Also throughout this Chapter, a Frobenius^n-linear map will always refer to the n^{th} power of the Frobenius map over the corresponding prime field.

3 The Curve Associated to ${}^2A_2$

3.1 The Group A_2 and its Representations

One of the groups associated to the Lie Algebra A_2 is the group $G = \mathrm{SL}_3$. This group acts naturally on a three dimensional vector space V. The Borels of G are defined by picking an arbitrary flag $0 \subset L \subset M \subset V$. As we range over Borel subgroups $\Lambda^2 M$ will span all of $\Lambda^2 V$, so our representation W will be given by $W = \Lambda^2 V$.

3.2 A Canonical Definition of ${}^2A_2$

The group A_2 is just SL_3. The group ${}^2A_2$ will turn out to be simply the special unitary group. Although this would be simple to derive directly, we will attempt to use the same basic technique as we will for the more complicated groups. Let $\mathcal{V}^3$ be the groupoid consisting of all three dimensional vector spaces over $\mathbb{F} := \overline{F_q}$ with a volume form, Ω. We define $\mathrm{Fr} : \mathcal{V}^3 \to \mathcal{V}^3$ as above. We define $' : \mathcal{V}^3 \to \mathcal{V}^3$ by letting V' equal $\mathrm{Hom}(V, \mathbb{F})$, the dual of V. Note that V' is naturally $\Lambda^2 V$ with the pairing $(v, \omega) = \frac{v \wedge \omega}{\Omega}$. We use these definitions interchangeably.

We have the obvious natural transformation $\rho : \mathrm{Id} \Rightarrow''$ so that $\rho_V(v)$ is the functional $\phi \in V' \to \phi(v)$. Given $V \in \mathcal{V}^3$ and $F : V' \to \mathrm{Fr}^{-m}(V)$, we can define σ by $\mathrm{Fr}^{-m}(\sigma(g)) \circ F = F \circ g'$. It is not hard to see that F defines a hermitian inner product on V and that $\sigma(g)$ is simply the adjoint of g with respect to this hermitian form.

3.3 The Deligne-Lusztig Curve

Let B be the Borel fixing the line $L = \langle v \rangle$ and the plane $M = \langle v, w \rangle = \ker(\phi)$, where $v \in V, \phi \in V'$. Note that we may think of ϕ as $\omega := v \wedge w \in \Lambda^2 V$. Now for $g \in B$, since g fixes M, g' must fix the line in V' containing ω. Similarly, since g fixes L, g' must fix the plane $L \wedge V$ in V'. Hence if B is defined by $L = \langle v \rangle$ and $M = \langle v, w \rangle$ and if u is some other linearly independent vector, then $\sigma(B)$ is defined by $\langle F(v \wedge w) \rangle$ and $\langle F(v \wedge w), F(v \wedge u) \rangle$.

Now for B to correspond to a point on the Deligne-Lusztig curve, it must therefore be the case that $L = \langle F(v \wedge w) \rangle$. We define the embedding of the Deligne-Lusztig curve C to $\mathbb{P}(\Lambda^2 V)$ by sending B to the line containing $\omega = v \wedge w$. We note that this is an embedding since given ω, we know that $\langle v \rangle = \langle F(\omega) \rangle$. This is a smooth embedding since the coordinates of C can be written as polynomials on $\mathbb{P}(\Lambda^2 V)$. For simplicity of notation we denote the pairing between V and V' as $(-, -)$. We note that the image of C is cut out by the equation:

$$(\omega, F(\omega)) = 0.$$

Thinking of F as defining a hermitian form on V', this just says that the norm of ω with respect to this hermitian form is 0. Hence C is just the Fermat curve of degree $q_0 + 1$.

3.4 Divisors of Note

We first compute a divisor that vanishes exactly on the $\mathbb{F}_q$-points. We note that a Borel B corresponds to a point over $\mathbb{F}_q$ if and only if $\sigma(\sigma(B)) = B$. We claim that this holds if and only if $\sigma(B) = B$. One direction holds trivially. For the other direction, if $\sigma(\sigma(B)) = B$, then $\sigma(\sigma(B))$ and fixes the same line as B and $\sigma(B)$. But the line fixed by $\sigma(B)$ is determined by the plane fixed by B. Hence B and $\sigma(B)$ fix both the same line and the same plane and hence are equal. Hence B corresponds to an $\mathbb{F}_q$-point if and only if $F(\langle v \wedge w, v \wedge u\rangle) = \langle v, w\rangle$. Since $F(v \wedge w) \propto v$, this happens exactly when $F(v \wedge u) \in \langle v, w\rangle$.

Letting u be some vector not in M we consider

$$\frac{(F(F(\omega) \wedge u) \wedge \omega) \otimes \Omega^{\otimes(q_0-1)}}{(\omega \wedge u)^{\otimes q_0}}. \tag{1}$$

Note that both numerator and denominator are multiples of $\Omega^{\otimes q_0}$ so the fraction makes sense. Note also that numerator and denominator are homogeneous of degree q_0 in u as an element of V/M. Hence the resulting expression is independent of u and hence a polynomial of degree $q - q_0 + 1$ in ω. It should be noted that this polynomial vanishes exactly when $(F(F(\omega) \wedge u) \wedge \omega) \propto F(v \wedge u) \wedge v \wedge w$ does, or in other words exactly at the $\mathbb{F}_q$-points. We could show that it vanishes simply at these points merely be a computation of degrees (knowing that we have the Fermat curve), but instead we will show it directly as that will be useful in our later examples, thus giving us another way of computing the degree of C.

Considering an analytic neighborhood of an $\mathbb{F}_q$-point in C, we may compute the polynomial in Equation 1 with u held constant. We would like to show that the derivative is non-zero. This is clearly equivalent to showing that the derivative of $(F(F(\omega) \wedge u) \wedge \omega)$ is non-zero. Suppose for sake of contradiction that it were zero. It is clear that $F(\omega) \wedge \omega$ is identically 0. Note also that the derivatives of $F(F(\omega) \wedge u)$ and $F(\omega)$ are both 0. Hence this would mean that $(F(F(\omega) \wedge u) \wedge d\omega = F(\omega) \wedge d\omega = 0$. But since $F(\omega)$ and $F(F(\omega) \wedge u)$ span M, this can only happen if $d\omega$ is proportional to ω, which means $d\omega = 0$ since we are in a projective space. Hence the polynomial in 1 vanishes exactly at the $\mathbb{F}_q$-points of C and with multiplicity 1. This proves that C is of degree $q_0 + 1$.

Recall that the polynomial $(\omega, F(\omega))$ is identically 0 on C. Consider the polynomial $(\omega, F(\mathfrak{F}(\omega)))$. This polynomial is clearly G^σ invariant. Since the divisor it defines has degree $(q_0 + 1)(q_0^3 + 1)$, it cannot vanish on any G^σ orbit other than the orbit of $\mathbb{F}_q$-points. Furthermore near an $\mathbb{F}_q$-point, ω_0, this divisor agrees with $(\omega, F(\mathfrak{F}(\omega_0)))$ to order q_0^3. Since the former vanishes to order at most $q_0 + 1$, $(\omega, F(\mathfrak{F}(\omega)))$ cannot vanish identically, and hence defines the divisor that vanishes to degree $q_0 + 1$ on each of the $\mathbb{F}_q$-points. Note also that since this divisor agrees with $(\omega, F(\mathfrak{F}(\omega_0)))$ to order q_0^3, the linear divisor $(\omega, F(\mathfrak{F}(\omega_0)))$ vanishes only at $\omega = \omega_0$ and to degree $q_0 + 1$ and nowhere else.

Next, consider the divisor $f = (\omega, F(\mathfrak{F}^2(\omega)))$. Note that if ω is defined over $\mathbb{F}_{q^3}$ that $\mathfrak{F}(f) = (\mathfrak{F}(\omega), F(\omega)) = (F(\omega), \omega) = 0$. Hence the divisor defined by f vanishes on the $\mathbb{F}_{q^3}$-

points of C. Additionally f agrees with $(\omega, F(\mathfrak{F}(\omega)))$ to order q_0^3 on any $\mathbb{F}_q$-point. Therefore f vanishes to order 1 on the $\mathbb{F}_{q^3}$-points, and order q_0+1 on the $\mathbb{F}_q$-points. This accounts for the entirety of the degree of the divisor so it therefore vanishes nowhere else.

Finally, consider the divisor defined by $(\omega, F(\mathfrak{F}^3(\omega)))$. Similarly to above this must vanish on the $\mathbb{F}_{q^4}$-points of C and to order exactly q_0+1 on the $\mathbb{F}_q$-points. The remaining degree unaccounted for is

$$(q_0^7+1)(q_0+1)-(q_0^3+1)(q_0+1)-q_0^3(q_0^3+1)(q_0-1)(q_0+1)=q_0^3(q_0+1)(q_0^3-q_0).$$

Since the remainder of the divisor is G^σ-invariant and cannot vanish on $\mathbb{F}_q$-points, the only orbit small enough is that of the $\mathbb{F}_{q^3}$-points. Hence the remainder of the divisor must be q_0 times the sum of the $\mathbb{F}_{q^3}$-points. Hence $(\omega, F(\mathfrak{F}^3(\omega)))$ vanishes to degree q_0+1 on the $\mathbb{F}_q$-points, to order q_0 on the $\mathbb{F}_{q^3}$-points, to order 1 on the $\mathbb{F}_{q^4}$-points, and nowhere else.

We now introduce several important polynomials. Let $F_3 = (\omega, F(\mathfrak{F}(\omega)))$, $F_5 = (\omega, F(\mathfrak{F}^2(\omega)))$, $F_7 = (\omega, F(\mathfrak{F}^3(\omega)))$. Let $P_1 = F_3^{1/(q_0+1)}$. We know that such a root exists since the $(q_0+1)^{st}$ root of the divisor of F_3 is the sum of the $\mathbb{F}_q$-points of C, and we know of polynomials with that divisor. Hence P_1 is a polynomial whose divisor is the sum of the $\mathbb{F}_q$-points of C. Let $P_3 = \frac{F_5}{F_3}$ and let $P_4 = \frac{F_7}{F_3 P_3^{q_0}}$. We know that these are polynomials (instead of just rational functions) because their associated divisors are the sum of the $\mathbb{F}_{q^3}$-points and the sum of the $\mathbb{F}_{q^4}$-points respectively.

We claim that the following relation holds:

$$P_4 - P_3^{q-q_0+1} + P_1^{q_0^3(q_0-1)} = 0. \tag{2}$$

Although the proof of Equation 2 is somewhat complicated we can much more easily prove that

$$P_4 - P_3^{q-q_0+1} + aP_1^{q_0^3(q-1)} = 0 \tag{3}$$

for some $(q_0+1)^{st}$ root of unity a. We do this by showing that for a properly chosen the above vanishes on the $\mathbb{F}_{q^4}$-points of C and the $\mathbb{F}_q$-points of C. Since this is more points than allowed by the degree of the polynomial, it implies that $P_4 - P_3^{q-q_0+1} + aP_1^{q_0^3(q_0-1)}$ must vanish identically. To show that it vanishes on the $\mathbb{F}_{q^4}$-points we show that for $Q \in C$ an $\mathbb{F}_{q^4}$-point that

$$P_3^{q_0^3+1}(Q) = P_1^{q_0^3(q-1)(q_0+1)}(Q) = F_3^{q_0^5-q_0^3}(Q).$$

This is obviously equivalent to showing that

$$F_5^{q_0^3+1}(Q) = F_3^{q_0^5+1}(Q).$$

If Q corresponds to a vector $\omega \in V$ defined over $\mathbb{F}_{q^4}$, we have that

$$\begin{aligned} F_5^{q_0^3+1}(Q) &= \big(\omega, F(\mathfrak{F}^2(\omega))\big) \cdot \big(F(\mathfrak{F}(\omega)), \mathfrak{F}^4(\omega)\big) \\ &= \big(\omega, F(\mathfrak{F}^2(\omega))\big) \cdot (\omega, F(\mathfrak{F}(\omega))) \\ &= (\omega, F(\mathfrak{F}(\omega))) \cdot \big(F(\mathfrak{F}^2(\omega)), \mathfrak{F}^4(\omega)\big) \\ &= F_3^{q_0^5+1}(Q). \end{aligned}$$

Next, we consider the $\mathbb{F}_q$-point. We need to show that if $Q \in C(\mathbb{F}_q)$ that

$$P_4(Q) = P_3^{q-q_0+1}(Q).$$

Equivalently, we will show that

$$F_7 F_3^q - F_5^{q+1}$$

vanishes to degree more than $(q+1)(q_0+1)$ at Q. This is easy since if we use our parameter ω with $\omega = \omega_0$ at Q (where $\omega_0 = \mathfrak{F}(\omega_0)$) than up to order $q_0^3 + q(q_0+1)$ the above is

$$(\omega, F(\omega_0)) \cdot (\omega, F(\omega_0))^q - (\omega, F(\omega_0))^{q+1} = 0.$$

This completes our proof of Equation 3.

4 The Curve Associated to ${}^2B_2$

4.1 The Group B_2 and its Representations

A form of B_2 is Sp_4. It has a natural representation on a four-dimensional symplectic vector space V. The Borels of G correspond to complete flags $0 \subset L \subset M \subset L^\perp \subset V$ where M is a lagrangian plane. We have another representation $W \subset \Lambda^2 V$ given as the kernel of the map $\Lambda^2 V \to \mathbb{F}$ defined by the alternating form on V. Equivalently, W is the subset consisting the 2-forms α so that $\alpha \wedge \omega = 0$, where ω is the 2-form corresponding to the alternating form on V. In characteristic 2, ω is contained in W, so we can take the quotient $V' = W/\langle\omega\rangle$.

4.2 A Canonical Definition of ${}^2B_2$

We define the groupoid SymSp_4 whose objects are symplectic spaces of dimension 4 over $\mathbb{F} = \overline{\mathbb{F}_2}$, and whose morphisms are symplectic linear transformations. We will write objects of SymSp_4 either as V or as (V, ω), where V is underlying vector space and ω is the 2-form associated to the symplectic form.

To each object (V, ω) in SymSp_4, we can pick a symplectic basis e_0, e_1, f_0, f_1 of V so that $(e_i, e_j) = (f_i, f_j) = 0, (e_i, f_j) = \delta_{i,j}$. We next note that there is a canonical volume form $\mu \in \Lambda^4 V$. In any characteristic other than 2, we could use $\mu = \omega \wedge \omega$, but here $\omega \wedge \omega = 0$. We instead use $\mu = e_0 \wedge e_1 \wedge f_0 \wedge f_1$. We have to prove:

Lemma 2. *$\mu = e_0 \wedge e_1 \wedge f_0 \wedge f_1$ is independent of the choice of symplectic basis.*

Proof. Any two symplectic bases can be interchanged by some symplectic transformation $A : V \to V$. We need only show that $\det(A) = 1$. Writing A in the basis of our first symplectic basis and letting

$$J = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

the statement that A is symplectic becomes $J = AJA^T$. Therefore taking determinants we find $\det(A)^2 = 1$ so $\det(A) = 1$ (since we are in characteristic 2). □

Next we note that for $(V, \omega) \in \mathrm{SymSp}_4$, that we have a bilinear form on $\Lambda^2 V$ given by $(\alpha, \beta) = \frac{\alpha \wedge \beta}{\mu}$. We note that this pairing is alternating since if e_I is a wedge of two 1-forms then clearly $(e_I, e_I) = 0$ and if $\alpha = \sum_I c_I e_I$ then

$$(\alpha, \alpha) = \sum_{I,J} c_I c_J \, (e_I, e_J) = \sum_I c_I^2 \, (e_I, e_I) + 2 \sum_{I<J} c_I c_J \, (e_I, e_J) = 0.$$

This alternating form is inherited by the subquotient $V' = \omega^\perp / \langle \omega \rangle$ of $\Lambda^2 V$. Lastly we note that the pairing on V' is non-degenerate. This can be seen by picking a symplectic basis, e_i, f_i, and noting that $e_0 \wedge e_1, e_0 \wedge f_1, f_0 \wedge f_1, e_1 \wedge f_1$ is a symplectic basis for V'. This allows us to define a functor $' : \mathrm{SymSp}_4 \to \mathrm{SymSp}_4$ by $V \to V'$ as above.

We also have the functor $\mathrm{Fr} : \mathrm{SymSp}_4 \to \mathrm{SymSp}_4$ as in Section 2.

We next define a natural transformation:

$$\rho :'' \Rightarrow \mathrm{Fr}.$$

To do this we must produce a Frobenius^{-1}-linear map $\rho_V : V'' \to V$ for $V \in \mathrm{SymSp}_4$. We provide this by producing a (linear,Frobenius-linear), pairing $V'' \times V \to \mathbb{F}$. We define this pairing on $(\alpha \wedge \beta, v)$ (for $\alpha, \beta \in \Lambda^2 V, v \in V$) by

$$(\alpha \wedge \beta, v) = (i_v(\alpha), i_v(\beta)) \, . \tag{4}$$

We note that the function in Equation 4 thought of as a map from $\Lambda^2 V \times \Lambda^2 V \times V \to \mathbb{F}$ is clearly linear and anti-symmetric in α and β, and clearly homogeneous of degree 2 in v. In order to show that it is well defined on $V'' \times V$ we need to demonstrate that it vanishes if

either α or β is a multiple of ω, that the map on $\Lambda^2 V' \times V$ vanishes on the symplectic form of V', and that it is additive in v. The first of these is true since

$$(i_v(\omega), i_v(\alpha)) = (v, i_v(\alpha)) = \alpha(v*, v*) = 0.$$

Where $v*$ is the dual of v, and the last equation holds since α is alternating. Let e_i, f_i be a symplectic basis of V.

Note that we have a basis

$$\{(e_0 \wedge e_1) \wedge (e_0 \wedge f_1), (e_0 \wedge e_1) \wedge (e_1 \wedge f_0), (e_1 \wedge f_0) \wedge (f_0 \wedge f_1), (e_0 \wedge f_1) \wedge (f_0 \wedge f_1)\}$$

of V''. To prove that our function is additive in v, notice that it suffices to check that

$$(i_u(\alpha), i_v(\beta)) = (i_v(\alpha), i_u(\beta))$$

for all $u, v \in V$. Hence it suffices to check the above for $\alpha = a \wedge b, \beta = a \wedge c$ for $a, b, c \in V$ and $(a, b) = (a, c) = 0$. In this case

$$\begin{aligned}(i_u(\alpha), i_v(\beta)) &= (i_u(a \wedge b), i_v(a \wedge c)) \\ &= ((u, a)\, b + (u, b)\, a, (v, a)\, c + (v, c)\, a) \\ &= (a, u)\,(a, v)\,(b, c)\,.\end{aligned}$$

Since this is symmetric in u and v, the operator is additive.

Hence, we have a $(\text{linear}, \text{Frobenius} - \text{linear})$ function $\Lambda^2 V' \times V \to \mathbb{F}$. To show that it descends to a function $V'' \times V \to \mathbb{F}$, it suffices to check that our form vanishes on

$$(e_0 \wedge e_1) \wedge (f_0 \wedge f_1) + (e_0 \wedge f_1) \wedge (f_0 \wedge e_1),$$

which is clear from checking on basis vectors (in fact our form vanishes on each term in the above sum) and by additivity.

Finally, we need to show that ρ_V is symplectic. This can be done by picking a symplectic basis and noting that by the above:

$$\begin{aligned}\rho_V((e_0 \wedge e_1) \wedge (e_0 \wedge f_1)) &= e_0 \\ \rho_V((e_0 \wedge e_1) \wedge (e_1 \wedge f_0)) &= e_1 \\ \rho_V((e_1 \wedge f_0) \wedge (f_0 \wedge f_1)) &= f_0 \\ \rho_V((e_0 \wedge f_1) \wedge (f_0 \wedge f_1)) &= f_1.\end{aligned}$$

It is clear that ρ defines a natural transformation.

Now given an $F : V' \to \mathrm{Fr}^{-m}(V)$, we get an $\mathbb{F}_q$-Frobenius $\mathfrak{F}$ on V, and we can now define $\sigma : G \to G$ as in Section 2 so that $\sigma^2 = \mathrm{FR}$.

4.3 The Deligne-Lusztig Curve

For simplicity, from here on out we will think of F as a Frobenius^m-linear map from $V' \to V$ rather than a linear one from $V' \to \mathrm{Fr}^{-m}(V)$.

We recall that for $V \in \mathrm{SymSp}_4$ that $G = \mathrm{Sp}_4(V)$ has Borel subgroups that correspond to the data of a line $L \subset V$ and a Lagrangian subspace $M \subset V$ so that $L \subset M$. The corresponding Borel, B is the set of elements g that fix both L and M. The Deligne-Lusztig curve can be thought of as the set of Borel subgroups, B, in the flag variety so that B and $\sigma(B)$ fix the same line.

Consider the Borel subgroup B fixing the line $\langle v \rangle$ and the plane $\langle v, w \rangle$ with $(v, w) = 0$. Notice that since $(v, w) = 0$ that $\omega \wedge v \wedge w = 0$, and thus $v \wedge w$ can be thought of as a (necessarily non-zero) element of V'. If g fixes $\langle v, w \rangle$, then g' clearly fixes the line of $v \wedge w$. Hence $F(v \wedge w)$ must be the line fixed by $\sigma(B)$.

Given V and F as above, we define the Deligne-Lusztig curve C and produce an embedding $C \hookrightarrow \mathbb{P}(W)$ by sending B to the line containing $\alpha = v \wedge w$ (recall $W \subset \Lambda^2 V$ was the orthogonal compliment of ω). This is an embedding since if we pick a point in the image we have fixed both the plane fixed by B and the line fixed by $\sigma(B)$ (and hence also the line fixed by B since they are the same). This embedding is smooth since all of the other coordinates of $B, \sigma(B)$ can be written as polynomials in α. Furthermore this embedding is given by a few simple equations. Namely:

- The quadratic relation equivalent to α being a pure wedge of two vectors (i.e. that it lies on the Grassmannian).
- The span of $\alpha, F(\alpha) \wedge V$ must have dimension at most 3.

The first condition guarantees that $F(\alpha) \neq 0$. The second condition implies that if we set $v = F(\alpha)$ that $\alpha = [v \wedge w]$ for some w. The fact that $\alpha \in W$ implies that $(v, w) = 0$. Together these imply that we have a point of the image.

These equations also cut out the curve scheme-theoretically. We show this for $m > 0$ by showing that the tangent space to the scheme defined by these equations is one dimensional at every $\bar{\mathbb{F}}$-point. If we are at some point α, we note that the derivative of $F(\alpha)$ is 0, so along any tangent line, $d\alpha$ must lie in $F(\alpha) \wedge V$. This restricts $d\alpha$ to a three-dimensional subspace of $\Lambda^2 V$. Additionally, $(\omega, d\alpha)$ must be zero, restricting to two dimensions. Finally, we are reduced to one dimension when we project onto $\Lambda^2 V / \langle \alpha \rangle$, which is the tangent space to projective space at α.

4.4 Divisors of Note

Let $q = 2^{2m+1}$. Let $q_0 = 2^m$. We wish to compute the degree of the above embedding of C. We do this by producing a polynomial that exactly cuts out the F_q-points of C and thus

must be a divisor of degree q^2+1. The $\mathbb{F}_q$-points are the points corresponding to Borel subgroups B so that $B = \sigma(\sigma(B))$. First, we claim that these are exactly the Borels for which $B = \sigma(B)$. This is because B and $\sigma(B)$ automatically fix the same line $\langle v \rangle$. If $\sigma(B)$ and $\sigma(\sigma(B))$ fix the same line then the 2-forms corresponding to the planes fixed by B and $\sigma(B)$ must by multiples of each-other modulo ω. But since ω cannot be written as a pure wedge of v with any vector, this implies that B and $\sigma(B)$ must fix the same plane, and hence be equal.

For any Borel B in C, we can pick $v, w, u \in V$ so that B fixes the line $L = \langle v \rangle$, the Lagrangian plane $M = \langle v, w \rangle$ and the space $S = \langle v \rangle^{\perp} = \langle v, w, u \rangle$. Notice that for $g \in B$ that g' fixes the Lagrangian plane $\langle v \wedge w, v \wedge u \rangle$. Hence $\sigma(B)$ fixes the plane $\langle F(v \wedge w), F(v \wedge u) \rangle$. Since $F(v \wedge w)$ is parallel to v, $B = \sigma(B)$ if and only if $F(v \wedge u)$ lies in M. Note that since $v \wedge w \perp v \wedge u$ in V' that $F(v \wedge u)$ must lie in S. We can then consider:

$$\left(\frac{\alpha}{v \wedge w}\right)^{q-2q_0+1} \left(\frac{F(v \wedge u)/M}{u/M}\right) (u, w)^{1-q_0} \left(\frac{F(v \wedge w)}{v}\right)^{2q_0-1}. \tag{5}$$

Note that the value in Expression 5 is independent of our choice of v, w, u, is never infinite, is zero exactly when $B = \sigma(B)$, and is homogenous of degree $q - 2q_0 + 1$ in α. Therefore it defines a polynomial of degree $q - 2q_0 + 1$ in our embedding that vanishes exactly on the F_q-points. We have left to show that it vanishes to order 1 at these points. Consider picking a local coordinate around an F_q-point of C and computing the derivative of Expression 5 with respect to this local coordinate. Clearly this derivative is some non-zero multiple of the derivative of

$$\left(\frac{F(v \wedge u)/M}{u/M}\right).$$

This can be rewritten as

$$\frac{F(v \wedge u) \wedge \alpha}{u \wedge \alpha}.$$

So in addition to picking a local coordinate, z, for α we can pick a local coordinate for v, u as well. We can let $v = F(\alpha)$, and let u be a fixed vector perpendicular to both v and dv/dz. Then for $m > 0$ we have that $dF(v \wedge u)/dz = 0$. Note that α must always be a pure wedge of $F(\alpha)$ with some perpendicular vector. Since $dF(\alpha) = 0$, $d\alpha$ must lie in $\langle v \wedge w, v \wedge u \rangle$. It cannot be parallel to $v \wedge w$ though since this is parallel to α. But, since $F(v \wedge u)$ is in $\langle v, w \rangle$ and not parallel to v, this means that $F(v \wedge u) \wedge d\alpha$ is non-zero. Hence the derivative of Expression 5 with respect to z at an $\mathbb{F}_q$-point is non-zero, and hence Expression 5 defines a divisor that is exactly the sum of the $\mathbb{F}_q$-points of C. Therefore the degree of our embedding must be:

$$\frac{q^2+1}{q - 2q_0 + 1} = q + 2q_0 + 1.$$

Note that the divisor (α, α) is identically 0, where the pairing is as elements of V'.

Next consider the polynomial $(\alpha, \mathfrak{F}(\alpha))$. We claim that this function is identically zero on C. We prove this by contradiction. If it were non-zero it would define a divisor of degree $(q+1)(q+2q_0+1)$ that would be clearly invariant under the action of G^σ on C. On the other hand, the only orbit small enough to be covered by this is the orbit of the $\mathbb{F}_q$-points, whose number does not divide the necessary degree.

Next consider the polynomial $(\alpha, \mathfrak{F}^2(\alpha))$. We claim that the divisor of this polynomial is simply $q+2q_0+1$ times the sum of the $\mathbb{F}_q$-points of C. Again the divisor must be G^σ invariant, and again the only orbit of small enough order is the orbit of $\mathbb{F}_q$-points. Hence if the polynomial is not uniquely zero it must be the divisor specified. On the other hand, consider a local coordinate $\alpha(z)$ around an $\mathbb{F}_q$-point. Then this divisor is equal to $(\alpha, \alpha_0) + O(z^{q^2})$ ($\alpha_0 = \alpha(0)$). (α, α_0) cannot vanish to degree more than $q+2q_0+1$ since it is a degree 1 polynomial. Hence $(\alpha, \mathfrak{F}^2(\alpha))$ vanishes to degree exactly $q+2q_0+1$ on each $\mathbb{F}_q$-point and nowhere else. Furthermore if β is the coordinate of an $\mathbb{F}_q$-point, then the degree 1 polynomial (α, β) vanishes at this point to degree $q+2q_0+1$ and nowhere else.

Next consider the polynomial $(\alpha, \mathfrak{F}^3(\alpha))$. This corresponds to a divisor of degree $(q^3+1)(q+2q_0+1)$. Note that if β is the coordinate vector for an $\mathbb{F}_{q^4}$-point of C that $(\beta, \mathfrak{F}^{-1}(\beta)) = 0$. Since $\beta = \mathfrak{F}^4(\beta)$, this implies that $(\beta, \mathfrak{F}^3(\beta)) = 0$. Therefore this divisor contains each of the $\mathbb{F}_{q^4}$ points. Also by the above this divisor vanishes to degree exactly $q+2q_0+1$ on the $\mathbb{F}_q$-points (and is hence non-zero). We have just accounted for a total degree of

$$q^2(q+2q_0+1)(q-1) + (q^2+1)(q+2q_0+1) = (q+2q_0+1)(q^3+1),$$

thus accounting for all of the vanishing of the polynomial. Hence this polynomial defines the divisor given by the sum of the $\mathbb{F}_{q^4}$-points plus $(q+2q_0+1)$ times the sum of the $\mathbb{F}_q$-points.

Lastly, consider the divisor $(\alpha, \mathfrak{F}^4(\alpha))$. By the arguments above it vanishes on the $\mathbb{F}_{q^5}$-points and to degree exactly $q+2q_0+1$ on the $\mathbb{F}_q$-points. We have so far accounted for a divisor of total degree

$$q^2(q^2+1)(q-1) + (q+2q_0+1)(q^2+1) = (q+2q_0+1)(q^4 - 2q_0q^3 + 2q_0q^2 + 1).$$

We are missing a divisor of degree $(q+2q_0+1)q^2(q-1)2q_0$. This divisor must be G^σ-invariant and cannot contain the orbit of $\mathbb{F}_q$-points. Hence the only orbit small enough is that of the $\mathbb{F}_{q^4}$-points, which can be taken $2q_0$ times. Hence this polynomial defines the divisor equal to the sum of the $\mathbb{F}_{q^5}$-points plus $2q_0$ times the sum of the $\mathbb{F}_{q^4}$-points, plus $q+2q_0+1$ times the sum of the $\mathbb{F}_q$-points.

Note that above we demonstrated that there was a polynomial that vanished to degree 1 exactly on the $\mathbb{F}_q$-points. Call this polynomial P_1. Note that we can choose P_1 so that $P_1^{q+2q_0+1} = (\alpha, \mathfrak{F}^2(\alpha))$. We also have polynomials $P_4 = \frac{(\alpha, \mathfrak{F}^3(\alpha))}{(\alpha, \mathfrak{F}^2(\alpha))}$ that vanishes to degree 1 on the $\mathbb{F}_{q^4}$ points and nowhere else. Finally we have $P_5 = \frac{(\alpha, \mathfrak{F}^4(\alpha))}{P_4^{2q_0}(\alpha, \mathfrak{F}^2(\alpha))}$ that vanishes exactly on

the $\mathbb{F}_{q^5}$-points. We claim that if the correct root P_1 is chosen that:

$$P_1^{q^2(q-1)} + P_4^{q-2q_0+1} + P_5 = 0.$$

We do this by demonstrating that this polynomial vanishes on all of the $\mathbb{F}_{q^5}$-points and all of the $\mathbb{F}_q$-points, which is more than a polynomial of its degree should be able to. In fact since everything can actually be defined over $\mathbb{F}_2$, the correct choice of P_1 will have to be the one given in Expression 5.

We begin by showing vanishing on the $\mathbb{F}_{q^5}$-points. We first show that $(\alpha, \mathfrak{F}^2(\alpha))^{q^2(q-1)} = P_4^{q^2+1}$ on the $\mathbb{F}_{q^5}$-points. We then note that by taking P_1 to be an appropriate root of $(\alpha, \mathfrak{F}^2(\alpha))$ we can cause $P_1^{q^2(q-1)} + P_4^{q-2q_0+1}$ to vanish on some $\mathbb{F}_{q^5}$-point. But then we note that P_1 is a multiple of the quantity in Expression 5, which is clearly G^σ invariant, and hence $P_1^{q^2(q-1)} + P_4^{q-2q_0+1}$ must vanish on all $\mathbb{F}_{q^5}$-points.

Let α be an $\mathbb{F}_{q^5}$-point. We wish to show that $(\alpha, \mathfrak{F}^2(\alpha))^{q^2(q-1)} = P_4^{q^2+1}(\alpha)$. This is equivalent to asking that

$$\left(\alpha, \mathfrak{F}^2(\alpha)\right)^{q^3+1} = \left(\alpha, \mathfrak{F}^3(\alpha)\right)^{q^2+1}.$$

But this is clear since

$$\begin{aligned}
\left(\alpha, \mathfrak{F}^2(\alpha)\right)^{q^3+1} &= \left(\alpha, \mathfrak{F}^2(\alpha)\right)^{q^3} \left(\alpha, \mathfrak{F}^2(\alpha)\right) \\
&= \left(\mathfrak{F}^3(\alpha), \mathfrak{F}^5(\alpha)\right) \left(\alpha, \mathfrak{F}^2(\alpha)\right) \\
&= \left(\alpha, \mathfrak{F}^3(\alpha)\right) \left(\mathfrak{F}^5(\alpha), \mathfrak{F}^2(\alpha)\right) \\
&= \left(\alpha, \mathfrak{F}^3(\alpha)\right) \left(\alpha, \mathfrak{F}^3(\alpha)\right)^{q^2} \\
&= \left(\alpha, \mathfrak{F}^3(\alpha)\right)^{q^2+1}.
\end{aligned}$$

Next, we need to show that at an $\mathbb{F}_q$-point that $P_4(\alpha)^{q-2q_0+1} = P_5(\alpha)$. This is equivalent to showing that

$$\left(\frac{(\alpha, \mathfrak{F}^3(\alpha))}{(\alpha, \mathfrak{F}^2(\alpha))}\right)^{q+1} = \frac{(\alpha, \mathfrak{F}^4(\alpha))}{(\alpha, \mathfrak{F}^2(\alpha))}.$$

Equivalently, we will show that

$$\left(\alpha, \mathfrak{F}^3(\alpha)\right)^{q+1} + \left(\alpha, \mathfrak{F}^4(\alpha)\right) \left(\alpha, \mathfrak{F}^2(\alpha)\right)^q$$

vanishes to degree more than $(q+1)(q+2q_0+1)$ on an $\mathbb{F}_q$-point. The above is equal to

$$\left(\mathfrak{F}(\alpha), \mathfrak{F}^4(\alpha)\right) \left(\alpha, \mathfrak{F}^3(\alpha)\right) + \left(\alpha, \mathfrak{F}^4(\alpha)\right) \left(\mathfrak{F}(\alpha), \mathfrak{F}^3(\alpha)\right).$$

Setting $\alpha = \beta + z$, where β is defined over $\mathbb{F}_q$,

$$\begin{aligned}
&\left(\mathfrak{F}(\alpha), \mathfrak{F}^4(\alpha)\right)\left(\alpha, \mathfrak{F}^3(\alpha)\right) + \left(\alpha, \mathfrak{F}^4(\alpha)\right)\left(\mathfrak{F}(\alpha), \mathfrak{F}^3(\alpha)\right) \\
=&\left(\beta + z^q, \beta\right)\left(\beta + z, \beta\right) + \left(\beta + z, \beta\right)\left(\beta + z^q, \beta\right) + O(z^{q^3}) \\
=&\left(z^q, \beta\right)\left(z, \beta\right) + \left(z, \beta\right)\left(z^q, \beta\right) + O(z^{q^3}) \\
=&O(z^{q^3}).
\end{aligned}$$

This completes the proof of our identity.

5 The Curve Associated to ${}^2G_2$

5.1 Description of ${}^2G_2$ and its Representations

We begin by describing the group G_2. G_2 can be thought of as the group of automorphisms of an octonian algebra. This non-associative algebra can be given by generators a_i for $i \in \mathbb{Z}/7$ where the multiplication for a_i, a_{i+1}, a_{i+3} is given by the relations for the quaternions for i, j, k. The octonians have a natural conjugation which sends a_i to $-a_i$ and 1 to 1. Instead of thinking of the automorphisms of the entire octonian algebra, we will think about the automorphisms of the space of pure imaginary octonians (which must be preserved by any automorphism since they are the non-central elements whose squares are central). There are two pieces of structure on the pure imaginary octonians that allow one to reconstruct the full algebra. Firstly, there is a non-degenerate, symmetric pairing (x, y) given by the real part of $x \cdot y$. Also there is an anti-symmetric, bilinear operator $x * y$ which is the imaginary part of $x \cdot y$. There is also an anti-symmetric cubic form given by (x, y, z) is the real part of $x \cdot (y \cdot z)$, or $(x, y * z)$.

These pieces of structure are not unrelated. In particular, since any two octonians x, y satisfy $x(xy) = (xx)y$ we have for x and y pure imaginary that

$$(x, x)\, y = x * (x * y) + x\, (x, y)$$

or

$$x * (x * y) = (x, x)\, y - (x, y)\, x.$$

Therefore (x, x) is the eigenvalue of $y \to x * (x * y)$ on a dimension 6 subspace. Hence $(-, -)$ is determined by $*$.

Hence G has a 7 dimensional representation on V, the pure imaginary octonians. The Borels of G fix a flag $0 \subset L \subset M \subset S \subset S^\perp \subset M^\perp \subset L^\perp \subset V$. Where here we have that $(M, M) = 0$, $M * M = 0$, $S = L^\perp * L$. Also $\Lambda^2 V$ has a subrepresentation W of dimension 14 given by the kernel of $* : \Lambda^2 V \to V$. As we shall see, in characteristic 3, $W \supset W^\perp$ thus giving a 7 dimensional representation $V' = W/W^\perp$.

5.2 Definition of σ

A property of note is that in characteristic 3, $*$ makes the pure imaginary octonians into a Lie Algebra. The Jacobi identity is easily checked on generators. We now let $\mathbb{F} = \overline{\mathbb{F}_3}$. We define $\mathbb{O}$ to be the groupoid of Lie Algebras isomorphic to the pure imaginary octonians over $\mathbb{F}$ with the operation $*$. We claim that for $V \in \mathbb{O}$ that the Lie Algebra of outer derivations of V is another Lie Algebra in $\mathbb{O}$ (this is not too hard to check using the standard basis). This obviously defines a functor $' : \mathbb{O} \to \mathbb{O}$.

We will discuss a little bit of the structure of these outer derivations. First note that since a derivation is a differential automorphism, and since an automorphism must preserve $(-,-)$, that these derivations must be alternating linear operators. This means that we can think of the derivations of V as lying inside of $\Lambda^2 V$. It is not hard to check that $a_0 \wedge a_1 + a_2 \wedge a_5$ defines a derivation. From symmetry and the fact that the dimension of the space of derivations is the dimension of G_2, which is 14, we find that the space of all derivations must be $W = \ker(* : \Lambda^2 V \to V)$. It is not hard to see that the space of inner derivations is equal to $W^\perp \subset W$. Hence $V' = W/W^\perp$ is isomorphic to the space of outer derivations.

We will need to know something about the structure of the operator $\mathrm{ad}(x) : V \to V$ when $x \in V$ has $(x,x) = 0$ in characteristic 3. Since $\mathrm{ad}(x)(\mathrm{ad}(x)(y)) = x\,(x,y)$ we know that $\mathrm{ad}(x)^2$ has a kernel of dimension 6. Therefore $\mathrm{ad}(x)$ has a kernel of dimension at least 3. We claim that the dimension of this kernel is exactly 3. Suppose that $x * y = 0$. By the above this implies that $(x,y) = 0$. Hence, as octonians, it must be the case that $x \cdot y = 0$. This means that if Nm is the multiplicative norm on the octonians and if we take lifts $\overline{x}, \overline{y}$ of x and y to the rational octonians, that $\mathrm{Nm}(\overline{x} \cdot \overline{y})$ is a multiple of 9. Since we can choose $\overline{x}$ so that $\mathrm{Nm}(\overline{x})$ is not a multiple of 9, $\mathrm{Nm}(\overline{y})$ is a multiple of 3. Hence $(y,y) = 0$. Hence $\ker(\mathrm{ad}(x))$ is a null-plane for $(-,-)$, and hence cannot have dimension more than 3. Notice that this implies that $\ker(\mathrm{ad}(x)) = \mathrm{ad}(x)(\langle x \rangle^\perp)$.

As before we have a functor $\mathrm{Fr} : \mathbb{O} \to \mathbb{O}$.

Once again our definition of ${}^2G_2$ will depend upon finding a natural transformation between $''$ and Fr. This time, we will find a natural equivalence in the other direction (since the inverse will be easier to write down). We will find a natural transformation $\rho : \mathrm{Fr} \Rightarrow ''$. This amounts to finding a Frobenius-linear isomorphism from V to V''. The basic idea will be to think of derivations of V as differential automorphisms. In particular, given a derivation d, we can find an element $1 + \epsilon d + O(\epsilon^2)$ of $\mathrm{End}(V)[[\epsilon]]$ that is an automorphism of V.

Let $x \in V$ and let d be a derivation of V. We let e_x be an automorphism of the form $1 + \epsilon \mathrm{ad}(x) + \epsilon^2/2 \mathrm{ad}(x)^2 + O(\epsilon^3)$, and let e_d be an automorphism of the form $1 + \delta d + O(\delta^2)$. We claim that the commutator of e_x and e_d is $1 + (\text{inner derivations}) + \epsilon^3 \delta r + O(\epsilon^4 \delta)$ where the O assumes that $\delta \ll \epsilon$ and where r is some derivation of V. We claim that this defines a map $(\rho_V(x))(d) = r$ so that $\rho_V(x)$ gives a well-defined outer automorphism of V', and that $\rho_V : V \to V''$ is Frobenius-linear. We prove this in a number of steps below.

Claim 1. *The commutator $[e_x, e_d]$ is of the form specified, namely*

$$1 + (\textit{inner derivations}) + \epsilon^3\delta r + O(\epsilon^4\delta)$$

for some derivation r.

Proof. We note that any pure ϵ or pure δ terms must vanish from the commutator, leaving only terms involving both ϵ and δ. To compute the terms of size at least $\epsilon^2\delta$, we compute

$$\begin{aligned}&(1 + \epsilon\mathrm{ad}(x) + \epsilon^2/2\mathrm{ad}^2(x))(1 + \delta d) \cdot (1 - \epsilon\mathrm{ad}(x) + \epsilon^2/2\mathrm{ad}^2(x))(1 - \delta d)\\ =&1 + \epsilon\delta\mathrm{ad}(d(x)) + \epsilon^2\delta(\mathrm{ad}(x * (d(x))/2)) + O(\epsilon^3\delta).\end{aligned}$$

So the $\epsilon\delta$ and $\epsilon^2\delta$ terms are inner derivations. The $\epsilon^3\delta$ term must also be a derivation since if we write this automorphism as $1 + a(\epsilon)\delta + O(\delta^2)$ it follows that $a(\epsilon)$ must be a derivation, and in particular that the ϵ^3 term of a is a derivation. □

Claim 2. *Given e_x and d, r does not depend on the choice of e_d. In particular, given a derivation d of V and an automorphism $e_x \in End(V)[[\epsilon]]$, we obtain a well-defined $r \in V'$.*

Proof. d defines e_d up to $O(\delta^2)$, and therefore $[e_x, e_d]$ is defined up to $O(\delta^2)$. □

Claim 3. *For fixed $e_x \in End(V)[[\epsilon]]$, the corresponding map $d \to r$ descends to a well defined linear map $V' \to V'$.*

Proof. It is clear that this produces a linear map from $\mathrm{Der}(V) \to V'$. If d is an inner derivation, $\mathrm{ad}(y)$ then the resulting product is

$$\exp(\delta\mathrm{ad}(e_x(y)))\exp(-\delta\mathrm{ad}(y)) + O(\delta^3) = 1 + \delta\mathrm{ad}(e_x(y) - y) + O(\delta^2).$$

So r is an inner derivation. Hence this descends to a map from V' to V'. □

Claim 4. *If $x \in V$ and $e_x, e'_x \in End(V)[[\epsilon]]$ are two possible automorphisms of the form specified, then e_x and e'_x define maps $V' \to V'$ that differ by an inner derivation of V'. Hence we have a well-defined map $\rho : V \to End(V')/InDer(V')$.*

Proof. Note that $e_x - e'_x$ must be of the form $\epsilon^3 f + O(\epsilon^4)$ for some derivation f. It is then not hard to check that the corresponding maps $V' \to V'$ differ by the inner derivation $d \to [\mathrm{ad}(f), d]$. □

Claim 5. *Let $x \in V$, $e_x \in End(V)[[\epsilon]]$ as above, $y \in V[[\epsilon^{1/n}]]$ with $y = o(\epsilon)$, and $e_y = 1 + ad(y) + ad^2(y)/2 + o(\epsilon^3)$ an automorphism. Then if we compute the above map $V' \to V'$ by taking the commutator of e_d with $e_y e_x$ instead of e_x, we obtain the same element of $End(V')$.*

Proof. This is because

$$e_y e_x e_d e_x^{-1} e_y^{-1} e_d^{-1} = (e_y(e_x e_d e_x^{-1} e_d^{-1}) e_y^{-1})(e_y e_d e_y^{-1} e_d^{-1}).$$

The first term above is $1 + \epsilon^3\delta r + o(\epsilon^3\delta)$ up to inner derivations, and the latter is $1 + o(\epsilon^3\delta)$ up to inner derivations. □

Claim 6. *The map $\rho : V \to End(V')/InDer(V')$ is Frobenius-linear.*

Proof. It is clear that the map $x, d \to r$ is homogeneous of degree 3 in x. We need to show that it is additive in x. By the previous claim, we may substitute $e_x e_y$ for e_{x+y} when computing the image of $x + y$. Additivity follows immediately from the identity

$$e_y e_x e_d e_x^{-1} e_y^{-1} e_d^{-1} = (e_y(e_x e_d e_x^{-1} e_d^{-1}) e_y^{-1})(e_y e_d e_y^{-1} e_d^{-1}).$$

□

Claim 7. *For $x \in V$ with $(x, x) = 0$, $(\rho(x))(d) = ad(x)ad(d(x))ad(x)$.*

Proof. First we claim that we may perform our computation using e_x of the form $e_x = 1 + \epsilon\mathrm{ad}(x) + \epsilon^2/2\mathrm{ad}^2(x) + O(\epsilon^4)$. To show this, it suffices to check that this e_x in an automorphism modulo ϵ^4. This holds because

$$\begin{aligned}
&(y + \epsilon x * y - \epsilon^2 x * (x * y)) * (z + \epsilon x * z - \epsilon^2 x * (x * z)) \\
=&(y + \epsilon x * y - \epsilon^2 (x, y) x) * (z + \epsilon x * z - \epsilon^2 (x, z) x) \\
=&(y * z + \epsilon((x * y) * z + y * (x * z)) \\
&+ \epsilon^2(-(x * (x * y)) * z + (x * y) * (x * z) - y * (x * (x * z))) \\
&+ \epsilon^3(-(x, z)(x * y) * x - (x, y) x * (x * z)) + O(\epsilon^4)) \\
=&(y * z + \epsilon(x * (y * z)) - \epsilon^2(x * (x * (y * z))) \\
&+ \epsilon^3((x, z)(x, y) x - (x, y)(x, z) x) + O(\epsilon^4)) \\
=&(y * z + \epsilon(x * (y * z)) - \epsilon^2(x * (x * (y * z))) + O(\epsilon^4)).
\end{aligned}$$

Using the e_x above, we get that

$$\begin{aligned}
&e_x e_d e_x^{-1} e_d^{-1} \\
=&(1 + \epsilon\mathrm{ad}(x) - \epsilon^2\mathrm{ad}^2(x) + O(\epsilon^4))(1 + \delta d + O(\delta^2)) \\
&\cdot (1 - \epsilon\mathrm{ad}(x) - \epsilon^2\mathrm{ad}^2(x) + O(\epsilon^4))(1 - \delta d + O(\delta^2)) \\
=&1 + \epsilon\delta\mathrm{ad}(d(x)) + \epsilon^2\delta(-\mathrm{ad}(x * d(x))) \\
&+ \epsilon^3\delta(-\mathrm{ad}(x)d\mathrm{ad}^2(x) + \mathrm{ad}^2(x)d\mathrm{ad}(x)) + o(\epsilon^3\delta).
\end{aligned}$$

Hence,

$$\begin{aligned} r &= -\mathrm{ad}(x)d\mathrm{ad}^2(x) + \mathrm{ad}^2(x)d\mathrm{ad}(x) \\ &= \mathrm{ad}(x)[\mathrm{ad}(x), d]\mathrm{ad}(x) \\ &= \mathrm{ad}(x)\mathrm{ad}(d(x))\mathrm{ad}(x). \end{aligned}$$

□

Claim 8. *For $x \in V$ with $(x,x) = 0$, the above map $\rho(x) : V' \to V'$ defined by*

$$d \to ad(x)\,ad(d(x))\,ad(x)$$

is a derivation.

Proof. Note that for $d, e \in V'$,

$$\begin{aligned} &[(\rho(x))(d), e] + [d, (\rho(x))(e)] \\ =& -\mathrm{ad}(e(x))\mathrm{ad}(d(x))\mathrm{ad}(x) - \mathrm{ad}(x)\mathrm{ad}(e(d(x)))\mathrm{ad}(x) \\ &- \mathrm{ad}(x)\mathrm{ad}(d(x))\mathrm{ad}(e(x)) + \mathrm{ad}(d(x))\mathrm{ad}(e(x))\mathrm{ad}(x) \\ &+ \mathrm{ad}(x)\mathrm{ad}(d(e(x)))\mathrm{ad}(x) + \mathrm{ad}(x)\mathrm{ad}(e(x))\mathrm{ad}(d(x)) \\ =&\mathrm{ad}(x)\mathrm{ad}([d,e](x))\mathrm{ad}(x) \\ &+ [\mathrm{ad}(d(x)), \mathrm{ad}(e(x))]\mathrm{ad}(x) - \mathrm{ad}(x)[\mathrm{ad}(d(x)), \mathrm{ad}(e(x))] \\ =&(\rho(x))([d,e]) + [\mathrm{ad}([d(x), e(x)]), \mathrm{ad}(x)] \\ =&(\rho(x))([d,e]) + \mathrm{ad}([[d(x), e(x)], x]). \end{aligned}$$

Which is $(\rho(x))([d,e])$ up to an inner derivation. □

Claim 9. *The map $\rho : V \to End(V')/InDer(V')$ gives a Frobenius linear map $V \to V''$.*

Proof. We already have that the map is Frobenius-linear. Furthermore the above claim implies that the image is a derivation of V' as long as $(x,x) = 0$. Since such vectors span V we have by linearity that the image lies entirely in $\mathrm{Der}(V)$. Hence we have a map $V \to \mathrm{Der}(V')/\mathrm{InDer}(V') = V''$. □

Claim 10. *The map $\rho : V \to V''$ is a map of Lie Algebras.*

Proof. We need to show that $\rho(x * y) = \rho(x) * \rho(y)$. We note that by Claim 5 that we may substitute $[e_x, e_y]$ (with ϵ replaced by $\epsilon^{1/2}$) for e_{x*y} in our computation of $\rho(x * y)$. Equivalently, we map compute $(\rho(x * y))(d)$ as the $\epsilon^6\delta$ term of $[[e_x, e_y], e_d]$. We note that

$$[[e_x, e_y], e_d] = e_y^{-1}e_x^{-1}e_ye_xe_de_x^{-1}e_y^{-1}e_xe_ye_d^{-1}$$

Is conjugate to

$$e_y e_x e_d e_x^{-1} e_y^{-1} e_x e_y e_d^{-1} e_y^{-1} e_x^{-1}.$$

This conjugation by $e_x e_y$ should not effect our final output since it should send inner derivations to inner derivations and not modify our $\epsilon^6\delta$ term by more than $O(\epsilon^7\delta)$. The above equals

$$e_{y(x(d))} e_{y(d)} e_{x(d)} e_d e_{x(y(d))}^{-1} e_{x(d)}^{-1} e_{y(d)}^{-1} e_d^{-1}$$

where

$$\begin{aligned} e_{x(d)} &= e_x e_d e_x^{-1} e_d^{-1} \\ e_{y(d)} &= e_y e_d e_y^{-1} e_d^{-1} \\ e_{x(y(d))} &= e_x e_{y(d)} e_x^{-1} e_{y(d)}^{-1} \\ e_{y(x(d))} &= e_y e_{x(d)} e_y^{-1} e_{x(d)}^{-1}. \end{aligned}$$

Since these are each $1 + O(\delta)$, they commute modulo δ^2, and hence modulo δ^2, the above is

$$e_{y(x(d))} e_{x(y(d))}^{-1}.$$

Up to inner derivations this is

$$1 + \epsilon^6\delta((\rho_V(y))((\rho_V(x))(d)) - (\rho_V(x))((\rho_V(y))(d))) + o(\epsilon^6\delta).$$

Hence $\rho_V(x * y) = \rho_V(x) * \rho_V(y)$ as desired. □

As in the previous case, if we have a $V \in \mathbb{O}$ and a Frobenius^m-linear map $F : V' \to V$, we can define $\mathfrak{F} : V \to V$ a Frobenius^{2m+1}-linear map thus giving V a $\mathbb{F}_{3^{2m+1}}$-structure. We can then construct an endomorphism σ of $G_2(V) = \mathrm{Aut}(V)$ by $\mathrm{Fr}^{-m}(\sigma(g)) \circ F = F \circ g'$, so that $\sigma^2 = \mathfrak{F}$.

5.3 The Deligne-Lusztig Curve

Before we can construct the Deligne-Lusztig curve we need to understand the Borel subgroups of G. Given $V \in \mathbb{O}$ we can consider the algebraic group $G = \mathrm{Aut}(V)$. We consider B a Borel subgroup of G. B is determined by a line $L = \langle x \rangle$ and a plane $M = \langle x, y \rangle \supset L$ that are fixed by it. These have the property that both $(-,-)$ and $*$ are trivial on both L and M. So $(x,x) = (x,y) = (y,y) = 0, x * y = 0$. B will also fix a 3 dimensional space $S = \langle x, y, z \rangle = \ker(\mathrm{ad}(x)) \supset M$. Given a Borel B of G, there should be a corresponding Borel B' of $G' = \mathrm{Aut}(V')$ by applying $'$ to each element of B. Letting $W = \ker(* : \Lambda^2 V \to V)$, we recall that $V' = W/W^\perp$. We note that if $g \in B$ that g' must fix the line containing $x \wedge y \in W$. Note that $(x \wedge y, x \wedge y) = 0$. Furthermore $x \wedge y$ cannot be an inner automorphism because $\mathrm{ad}(a)$ has rank 6 if $(a,a) \neq 0$ and rank 4 if $(a,a) = 0$ while $x \wedge y$ has rank 2. Therefore g'

fixes the line generated by $x \wedge y$ in V'. B' also fixes the plane $\langle x \wedge y, x \wedge z \rangle$. This is of the type described because $(x \wedge y) * (x \wedge z)(a) = 0$ since $(x \wedge z)(a) \in \langle x, z \rangle$ and both x and y are perpendicular to x and z. This means that if B is the Borel fixing L and M, then $\sigma(B)$ is the Borel fixing $F(\langle x \wedge y \rangle)$ and $F(\langle x \wedge y, x \wedge z \rangle)$.

The Deligne-Lusztig curve C is the set of Borels B so that B and $\sigma(B)$ fix the same line. We can produce an explicit embedding $C \hookrightarrow \mathbb{P}(\Lambda^2(V))$ by sending B to the line of $\omega = x \wedge y$. We claim that this embedding gives the curve defined by the following relations:

- The linear relations that tell us that $*(\omega) = 0$.
- The relations that tell us that ω is a rank 2 tensor, and hence a pure wedge of two elements of V. Note that these are the relations used to define the Plucker embedding of a Grassmannian.
- The relation $(\omega, \omega) = 0$
- The relations that tell us that $\langle \omega, F(\omega) \wedge V \rangle$ has dimension at most 6.

These are clearly satisfied for points in the image of our embedding. They also cut them out set-theoretically. The first and second relations guarantee that ω represents a non-zero element of V'. Let $x = F(\omega) \neq 0$. The fourth relation guarantees that $\omega = x \wedge y$ for some y. The first relation says that $x * y = 0$. The third relation implies that $(x, x) = 0$. Hence we have the Borel B that fixes $\langle x \rangle$ and $\langle x, y \rangle$ the unique Borel subgroup so that $x \wedge y$ is parallel to ω. This embedding is smooth because the coordinates of the flag variety can all be written as polynomials in ω for points on C.

These equations also cut out the curve scheme-theoretically. We show this for $m > 0$ by showing that the tangent space to the scheme defined by these equations is one dimensional at every $\bar{\mathbb{F}}$-point. Suppose that we are at some point on C with projective coordinate ω. We wish to consider the space of possible vectors $d\omega$ in the tangent space. Note that $dF(\omega) = 0$ and that therefore, $d\omega$ must lie in the six dimensional space defined by $F(\omega) \wedge V$. We must also have $*(d\omega) = 0$ so $d\omega$ must lie in $F(\omega) \wedge (\ker(\mathrm{ad}(F(\omega))))$. Since $(F(\omega), F(\omega)) = 0$, $(\ker(\mathrm{ad}(F(\omega))))$ is three dimensional. Since $(\ker(\mathrm{ad}(F(\omega))))$ contains $F(\omega)$, $F(\omega) \wedge (\ker(\mathrm{ad}(F(\omega))))$ is two dimensional. Finally when we project down to the tangent space to $\mathbb{P}(\Lambda^2 V)$ at this point (which is $\Lambda^2(V)/\langle \omega \rangle$), we are left with a one-dimensional tangent space.

5.4 Divisors of Note

We begin with some preliminaries. Let C be the Deligne-Lusztig curve corresponding to ${}^2G_2$ defined over the field $\mathbb{F}_{3^{2m+1}}$. Let $q_0 = 3^m$. Let $q = 3q_0^2 = 3^{2m+1}$. Let $q_- = q - 3q_0 + 1$ and $q_+ = q + 3q_0 + 1$. Note that $q_- q_+ = q^2 - q + 1$. We let $G = G_2$ with an endomorphism σ.

We note that C admits a natural G^σ action. We will be considering the projective model of C described above with parameter ω.

We first want to compute the degree of this embedding. We do this by finding a polynomial that vanishes exactly at the $\mathbb{F}_q$ points. We will think of F as a Frobenius^m-linear map $V' \to V$. If B is the Borel fixing the line, plane and space $L = \langle x \rangle, P = \langle x, y \rangle, S = \langle x, y, z \rangle$, $\sigma(B)$ fixes the line and plane $\langle F(x \wedge y) \rangle, \langle F(x \wedge y), F(x \wedge z) \rangle$. For B to correspond to a point on C, then we must have that $F(x \wedge y)$ is a multiple of x. B is defined over $\mathbb{F}_q$ if and only if $B = \sigma(\sigma(B))$. We claim that this happens if and only if $B = \sigma(B)$. The if direction is clear. For only if, note that if $B = \sigma(\sigma(B))$, B, $\sigma(B)$ and $\sigma(\sigma(B))$ must all fix the same line. Therefore $F(x \wedge y)$ is a multiple of $F(F(x \wedge y) \wedge F(x \wedge z))$, which is a multiple of $F(x \wedge F(x \wedge z))$. This happens if and only if $F(x \wedge z)$ is in the span of x and y, which in turn implies that $B = \sigma(B)$. So in fact, $B \in C$ is an $\mathbb{F}_q$-point if and only if $F(x \wedge z)$ is in the plane of x and y.

Notice that since $(x \wedge y) * (x \wedge z) = 0$ that $0 = F(x \wedge y) * F(x \wedge z) = x * F(x \wedge z)$. Therefore $F(x \wedge z)$ is in the span of x, y, z. Therefore we have that $(F(x \wedge z) \wedge x) \wedge (x \wedge y)$ is always a multiple of $(x \wedge y) \wedge (x \wedge z)$ in V''. Furthermore this multiple is 0 if and only if the point is defined over $\mathbb{F}_q$. We next claim that $\rho_V(x)$ is a (non-zero) multiple of $(x \wedge y) \wedge (x \wedge z)$. This is equivalent to saying that $(\rho_V(x))(V')$ is in the span of $x \wedge y$ and $x \wedge z$. This in turn is equivalent to saying that the image $((\rho_V(x))(V'))(V)$ is contained in the span of x, y and z. But note that $\langle x, y, z \rangle = \ker(\mathrm{ad}(x))$. Furthermore, by claim 7,

$$x * ((\rho_V(x))(d))(a) = x * (x * (d(x) * (x * a))) = (x, d(x) * (x * a))\, x.$$

We need to show that $(x, d(x) * (x * a)) = 0$. This is the cubic form applied to $x, d(x), (x * a)$. Hence this is $-(d(x), x * (x * a)) = -(d(x), (x, a)\, x) = -(x, a)\, (d(x), x) = 0$ since d must be an anti-symmetric linear operator. Hence $\rho_V(x)$ is a multiple of $(x \wedge y) \wedge (x \wedge z)$.

We now consider the polynomial

$$\frac{[(F(\omega) \wedge (F(F(\omega) \wedge z))) \wedge \omega] \otimes [\rho_V(F(\omega))]^{\otimes(q_0 - 1)}}{[\omega \wedge (F(\omega) \wedge z)]^{\otimes q_0}}. \tag{6}$$

Both numerator and denominator are elements of the q_0^{th} tensor power of the space of multiples of $(x \wedge y) \wedge (x \wedge z)$. Note that each of numerator and denominator are proportional to the q_0^{th} power of z as an element of S/M. Therefore the number produced is independent of the choice of z. The ω-degree of the above is:

$$[q_0 + q_0^2 + 1] + (q_0 - 1)[3q_0] - q_0[q_0 + 1] = q_0 + q_0^2 + 1 + q - 3q_0 - q_0^2 - q_0 = q - 3q_0 + 1.$$

The polynomial in Equation 6 clearly vanishes exactly when $F(x \wedge z)$ is in M, or on points defined over $\mathbb{F}_q$. We claim that it vanishes to degree 1 on such points (at least for $m > 0$). Consider an analytic coordinate around such a point. Since $m > 0$ the derivative of $F(\omega)$

is 0. Hence we can consider the above with z so that $dz = 0$. We wish to show that the derivative of the above polynomial is non-zero at such a point. To do so it suffices to show that the derivative of $(F(\omega) \wedge (F(F(\omega) \wedge z))) \wedge \omega$ is non-zero. But it should be noted that the derivative of $(F(\omega) \wedge (F(F(\omega) \wedge z)))$ is 0. Hence as ω varies, the wedge has non-zero derivative. This completes the proof. Therefore this polynomial of degree q_- defines a divisor of degree $q^3 + 1$. Hence our embedding must be of degree $q_+(q + 1)$.

We next consider the divisor defined by the polynomial $(\omega, \mathfrak{F}(\omega))$. This defines a divisor of degree $(q + 1)^2 q_+$. This divisor is clearly G^σ-invariant. On the other hand the degree is too small to contain any orbits except for the orbit of $\mathbb{F}_q$-points. Unfortunately, the size of this orbit does not divide the degree of this divisor. Therefore $(\omega, \mathfrak{F}(\omega))$ must vanish on C. Similarly $(\omega, \mathfrak{F}^2(\omega))$ must vanish on C.

Consider the polynomial $(\omega, \mathfrak{F}^3(\omega))$. We claim that this vanishes to degree exactly $q_+(q+1)$ on the $\mathbb{F}_q$-points of C and nowhere else. For the former, consider an $\mathbb{F}_q$-point $\omega = \omega_0$. Then near this point $(\omega, \mathfrak{F}^3(\omega))$ agrees with (ω, ω_0) to degree q^3. Since the latter cannot vanish to degree more than $q_+(q + 1)$, being a degree 1 polynomial, this means that $(\omega, \mathfrak{F}^3(\omega))$ cannot vanish identically. Since it is G^σ-invariant, but of too small a degree to contain any orbit but that of the $\mathbb{F}_q$-points, it must vanish to degree n on each $\mathbb{F}_q$-point for some n and nowhere else. Comparing degrees yields $n = q_+(q+1)$. Note also that this implies that the polynomial (ω, ω_0) vanishes at the point defined by ω_0 with multiplicity $q_+(q + 1)$ and nowhere else.

Next, consider the polynomial $(\omega, \mathfrak{F}^4(\omega))$. This vanishes on the $\mathbb{F}_{q^6}$-points of C. This is because for such points,

$$\mathfrak{F}^2(\big(\omega, \mathfrak{F}^4(\omega)\big)) = \big(\mathfrak{F}^2(\omega), \mathfrak{F}^6(\omega)\big) = \big(\mathfrak{F}^2(\omega), \omega\big) = 0.$$

Furthermore, this polynomial agrees with $(\omega, \mathfrak{F}^3(\omega))$ to order q^3 on the $\mathbb{F}_q$-points. Therefore this polynomial vanishes to degree 1 on the $\mathbb{F}_{q^6}$-points and degree $q_+(q+1)$ on the $\mathbb{F}_q$-points. By degree counting, it vanishes nowhere else.

Lastly, consider the polynomial $(\omega, \mathfrak{F}^5(\omega))$. Analogously to the above this vanishes on the $\mathbb{F}_{q^7}$-points and to degree exactly $q_+(q + 1)$ on the $\mathbb{F}_q$-points. The remaining degree is

$$\begin{aligned} &q_+(q + 1)(q^5 + 1) - q^3(q + 1)q_+q_-(q - 1) - (q^3 + 1)q_+(q + 1) \\ =&q_+(q + 1)(q^5 + 1 - q^3 - 1 - q^5 + 3q_0q^4 - 3q_0q^3 + q^3) \\ =&q_+(q + 1)(3q_0q^4 - 3q_0q^3). \end{aligned}$$

Since the remainder of this divisor is G^σ-invariant and the only orbit small enough to fit is the orbit of $\mathbb{F}_{q^6}$-points, this polynomial must vanish to degree 1 on the $\mathbb{F}_{q^7}$-points, to degree $3q_0$ on the $\mathbb{F}_{q^6}$-points, and to degree $q_+(q + 1)$ on the $\mathbb{F}_q$-points.

There are several polynomials of interest. Let $F_3 := (\omega, \mathfrak{F}^3(\omega))$, $F_4 := (\omega, \mathfrak{F}^4(\omega))$, $F_5 := (\omega, \mathfrak{F}^5(\omega))$. We also have the polynomials that vanish exactly at the $\mathbb{F}_{q^n}$-points. $P_1 := F_3^{\frac{1}{q_+(q+1)}}$, $P_6 := \frac{F_4}{F_3}$, $P_7 := \frac{F_5}{F_3 P_6^{3q_0}}$. We note that these are all polynomials (P_1 as above is

defined only up to a $q_+(q+1)^{st}$ root of unity but could be taken, for example, to be the Polynomial in Equation 6). We claim that if the correct root is taken in the choice of P_1, then we have that

$$P_1^{q^3(q-1)} + P_6^{q_-} - P_7 = 0. \tag{7}$$

Or unequivocally,

$$F_3^{q^3(q-1)} = (P_6^{q_-} - P_7)^{q_+(q+1)}.$$

We prove Equation 7 by showing that for proper choice of P_1, that this polynomial vanishes at all of the $\mathbb{F}_{q^7}$-points and all of the $\mathbb{F}_q$-points. Since this will be more points than the degree of the polynomial would allow, the polynomial must vanish on C.

For the $\mathbb{F}_{q^7}$-points, it will suffice to show that

$$F_3^{q^3(q-1)}(Q) = P_6^{q_-q_+(q+1)}(Q) = \left(\frac{F_4(Q)}{F_3(Q)}\right)^{q^3+1}$$

for Q a $\mathbb{F}_{q^7}$-point. For this it suffices to show that $F_3^{q^4+1}(Q) = F_4^{q^3+1}(Q)$. But we have that

$$\begin{aligned}
F_3^{q^4+1}(Q) &= \left(\omega, \mathfrak{F}^3(\omega)\right)^{q^4} \left(\omega, \mathfrak{F}^3(\omega)\right) \\
&= \left(\mathfrak{F}^4(\omega), \mathfrak{F}^7(\omega)\right) \left(\omega, \mathfrak{F}^3(\omega)\right) \\
&= \left(\mathfrak{F}^4(\omega), \omega\right) \left(\mathfrak{F}^7(\omega), \mathfrak{F}^3(\omega)\right) \\
&= \left(\omega, \mathfrak{F}^4(\omega)\right)^{q^3} \left(\omega, \mathfrak{F}^4(\omega)\right) \\
&= F_4^{q^3+1}(Q).
\end{aligned}$$

For the $\mathbb{F}_q$-points we show that $P_6^{q_-}(Q) = P_7(Q) = \frac{F_5(Q)}{F_3(Q)P_6^{3q_0}(Q)}$ for Q and $\mathbb{F}_q$-point of C. This is equivalent to showing that

$$\left(\frac{F_4(Q)}{F_3(Q)}\right)^{q+1} = P_6^{q+1}(Q) = \frac{F_5(Q)}{F_3(Q)}.$$

To do this we will show that the polynomials

$$F_4^{q+1}$$

and

$$F_5 F_3^q$$

agree to order more than $q_+(q+1)^2$. But this follows immediately from the fact that each of these polynomials is a product of $q+1$ of the F_i which are in turn polynomials that

- Vanish to order $q_+(q+1)$ at Q
- Agree with F_3 to order q^3 at Q

This completes our proof of Equation 7.

Part II

Designs

Chapter E

Designs for Path Connected Spaces

1 Introduction

Given a measure space (X, μ) and a set $f_1, \ldots, f_m : X \to \mathbb{R}$, [55] defines an *averaging set* to be a finite set of points, $p_1, \ldots, p_N \in X$ so that

$$\frac{1}{N} \sum_{i=1}^{N} f_j(p_i) = \frac{1}{\mu(X)} \int_X f_j d\mu \tag{1}$$

for all $1 \leq j \leq m$. They show that if X is a path-connected topological space, μ has full support, and the f_i are continuous that such sets necessarily exist. In this Chapter, we study the problem of how small such averaging sets can be. In particular, we define a *design problem* to be the data of X, μ and the vector space of functions on X spanned by the f_j. For a design problem, D, we show that there exist averaging sets (we call them designs) for D with N relatively small.

Perhaps the best studied case of the above is that of spherical designs, introduced in [10]. A spherical design on S^d of strength n is defined to be an averaging set for $X = S^d$ (with the standard measure) where the set of f_j is a basis for the polynomials of degree at most n on the sphere. It is not hard to show that such a design must have size at least $\Omega_d(n^d)$ (proved for example in [10]). It was conjectured by Korevaar and Meyers that designs of size $O_d(n^d)$ existed. There has been much work towards this Conjecture. Wagner proved in [18] that there were designs of size $O_d(n^{12d^4})$. This was improved by Korevaar and Meyers in [37] to $O_d(n^{(d^2+d)/2})$, by Bondarenko, and Viazovska in [5] to $O_d(n^{2d(d+1)/(d+2)})$. In [6], Bondarenko, Radchenko, and Viazovska recently announced a proof of the full conjecture.

In this Chapter, we develop techniques to prove the existence of small designs in a number of contexts. In greatest generality, we prove that on a path-connected topological space there exist designs to fool any set of continuous functions on X of size roughly MK, where M is the number of linearly independent functions, and K is a measure of how badly behaved

these functions are. We also show that if in addition X is a homogeneous space and the linear span of functions we wish to fool is preserved under the symmetry group of X that $K \leq M$. For example, this immediately implies strength-n designs of size $O(n^{2d}/(d!)^2)$ on S^d. It also implies the existence of small Grassmannian designs (see [1] for the definition). Generally, this result proves the existence of designs whose size is roughly the square of what we expect the optimal size should be.

With a slight modification of our technique, we can also achieve better bounds in some more specialized contexts. In particular, in Section 6 we produce designs of nearly optimal size for beta distributions on the unit interval, and in Section 7 we prove the existence of strength-n designs on S^d of size $O_d(n^d \log(n)^{d-1})$ for $d \geq 2$, which is optimal up to a polylog factor.

In Section 2, we describe the most general setting of our work and some of the fundamental ideas behind our technique. In Section 3, we handle our most general case of path-connected spaces. In Section 4, we produce an example in which the upper bound for sizes of designs in the previous section is essentially tight. In Section 5, we study the special case of homogeneous spaces. In Section 6, we provide nearly optimal bounds for the size of designs for beta distributions on the interval. In Section 7, we prove our bounds on the size of spherical designs.

2 Basic Concepts

We begin by defining the most general notion of a design that we deal with in this Chapter.

Definition. *A* design-problem *is a triple (X, μ, W) where X is a measure space with a positive measure μ, normalized so that $\mu(X) = 1$, and W is a vector space of L^1 functions on X.*

Given a design-problem (X, μ, W), a design *of size N is a list of N points (not necessarily distinct) $p_1, p_2, \ldots, p_N \in X$ so that for every $f \in W$,*

$$\int_X f(x) d\mu(x) = \frac{1}{N} \sum_{i=1}^{N} f(p_i). \tag{2}$$

A weighted design *of size N is a set of points $p_1, p_2, \ldots, p_N \in X$ and a list of weights $w_1, w_2, \ldots, w_N \in [0, 1]$ so that $\sum_{i=1}^{N} w_i = 1$ and so that for each $f \in W$,*

$$\int_X f(x) d\mu(x) = \sum_{i=1}^{N} w_i f(p_i) \tag{3}$$

For example, if (X, μ) is the d-sphere with its standard (normalized) measure, and W is the space of polynomials of total degree at most n restricted to X, then our notion of a

design (resp. weighted design) corresponds exactly to the standard notion of a design (resp. weighted design) of strength n on the d-sphere.

Note that a design is the same thing as a weighted design in which all the weights are $\frac{1}{N}$.

Notice that if we set $f(x)$ to be any constant function that the formulas in Equations 2 and 3 will hold automatically. Hence for a design problem it is natural to define the vector space V of functions on X to be the space of functions, f, in $W + \langle 1 \rangle$ so that $\int_X f(x)d\mu(x) = 0$.

Lemma 1. *For a design-problem (X, μ, W) with V as defined above, $p_1, p_2, \ldots, p_N$ is a design (resp. $p_1, p_2, \ldots, p_N, w_1, w_2, \ldots, w_N$ is a weighted design) if and only if for all $f \in V$, $\sum_{i=1}^N f(p_i) = 0$, (resp. $\sum_{i=1}^N w_i f(p_i) = 0$).*

Proof. Since any design can be thought of as a weighted design, it suffices to prove the version of this Lemma for weighted designs. First assume that $\sum_{i=1}^N w_i f(p_i) = 0$ for each $f \in V$. For every $g \in W$, letting $f(x) = g(x) - \int_X g(y)d\mu(y)$, $f \in V$. Hence

$$\begin{aligned} 0 &= \sum_{i=1}^N w_i \left(g(p_i) - \int_X g(y)d\mu(y) \right) \\ &= \sum_{i=1}^N w_i g(p_i) - \left(\sum_{i=1}^N w_i \right) \left(\int_X g(y)d\mu(y) \right) \\ &= \sum_{i=1}^N w_i g(p_i) - \int_X g(x)d\mu(x). \end{aligned}$$

Hence p_i, w_i is a weighted design.

If on the other hand, p_i, w_i is a weighted design and $f \in V$, then $f(x) = g(x) + c$ for some $g \in W$ and constant c. Furthermore $0 = \int_X g(x) + cd\mu(x) = \int_X g(x)d\mu(x) + c$ so $c = -\int_X g(x)d\mu(x)$. Hence

$$\begin{aligned} \sum_{i=1}^N w_i f(p_i) &= \sum_{i=1}^N w_i (g(p_i) + c) \\ &= \sum_{i=1}^N w_i g(p_i) + \left(\sum_{i=1}^N w_i \right) c \\ &= \int_X g(x)d\mu(x) + c \\ &= 0. \end{aligned}$$

□

It will also be convenient to associate with the design problem (X, μ, W) the number $M = \dim(V)$. We note that there is a natural map $E : X \to V^*$, where V^* is the dual space of V. This is defined by $(E(p))(f) = f(p)$. This function allows us to rephrase the idea of a design in the following useful way:

Lemma 2. *Given a design problem (X, μ, W) along with V and E as described above, p_i is a design (resp. p_i, w_i is a weighted design) if and only if $\sum_{i=1}^{N} E(p_i) = 0$ (resp. $\sum_{i=1}^{N} w_i E(p_i) = 0$).*

Proof. Again it suffices to prove only the version of this Lemma for weighted designs. Note that for $f \in V$, that

$$\sum_{i=1}^{N} w_i f(p_i) = \sum_{i=1}^{N} w_i (E(p_i))(f) = \left(\sum_{i=1}^{N} w_i E(p_i) \right) (f).$$

This is 0 for all $f \in V$, if and only if $\sum_{i=1}^{N} w_i E(p_i) = 0$. This, along with Lemma 1, completes the proof. □

To demonstrate the utility of this geometric formulation, we present the following Lemma:

Lemma 3. *Given a design problem (X, μ, W) with V, M, E as above, if $M < \infty$, there exists a weighted design for this problem of size at most $M + 1$.*

Proof. Note that for $f \in V$ that

$$\left(\int_X E(x) d\mu(x) \right) (f) = \int_X f(x) d\mu(x) = 0.$$

Therefore $\int_X E(x) d\mu(x) = 0$. Therefore 0 is in the convex hull of $E(X)$. Therefore 0 can be written as a positive affine linear combination of at most $M + 1$ points in $E(X)$. By Lemma 2, this gives us a weighted design of size at most $M + 1$. □

Unfortunately, our notion of a design problem is too general to prove many useful results about. We will therefore work instead with the following more restricted notion:

Definition. *A* topological design problem *is a design problem, (X, μ, W) in which X is a topological space, the σ-algebra associated to μ is Borel, the functions in W are bounded and continuous, and W is finite dimensional.*

We call a topological design problem path-connected *if the topology on X makes it a path-connected topological space.*

We call a topological design problem homogeneous *if for every $x, y \in X$ there is a measure-preserving homeomorphism $f : X \to X$ so that $f^*(W) = W$ and $f(x) = y$.*

We will also want a measure on the complexity of the functions in W for such a design problem.

Definition. *Let (X, μ, W) be a topological design problem. Associate to it the number*

$$\begin{aligned} K &= \sup_{f \in V \setminus \{0\}} \frac{\sup(f)}{|\inf(f)|} = \sup_{f \in V \setminus \{0\}} \frac{\sup(-f)}{|\inf(-f)|} \\ &= \sup_{f \in V \setminus \{0\}} \frac{-\inf(f)}{\sup(f)} = \sup_{f \in V \setminus \{0\}} \frac{\sup(|f|)}{\sup(f)}. \end{aligned}$$

Notice that since $\frac{\sup(f)}{|\inf(f)|}$ is invariant under scaling of f by positive numbers, and since $V \setminus \{0\}$ modulo such scalings is compact, that K will be finite unless there is some $f \in V \setminus \{0\}$ so that $f(x) \geq 0$ for all x. Since $\int_X f(x) d\mu(x) = 0$ this can only be the case if f is 0 on the support of μ.

Throughout the rest of the Chapter, to each topological design problem, (X, μ, W) we will associate V, E, M, K as described above.

3 The Bound for Path Connected Spaces

In this Section, we prove the following Theorem, which will also be the basis some of our later results.

Theorem 4. *Let (X, μ, W) be a path-connected topological design problem. If $M > 0$, then for every integer $N > (M-1)(K+1)$ there exists a design of size N for this design problem.*

Throughout the rest of this Section, we use X, μ, W, V, E, M, K, N to refer to the corresponding objects in the statement of Theorem 4. Our proof technique will be as follows. First, we construct a polytope P, that is spanned by points of $E(X)$, and contains the origin. Next, we construct a continuous function $F : P \to V^*$ so that every point in the image of F is a sum of N points in $E(X)$, and so that for each facet, T, of P, $F(T)$ lies on the same side of the hyperplane through the origin parallel the one defining T as T does. Lastly, we show, using topological considerations, that 0 must be in the image of F. We begin with the construction of P.

Proposition 5. *For every $\epsilon > 0$, there exists a polytope $P \subset V^*$ spanned by points in $E(X)$ such that for every linear inequality satisfied by the points of P of the form*

$$\langle x, f \rangle \leq c$$

for some $f \in V \setminus \{0\}$, we have

$$\sup_{p \in X} |f(p)| \leq c(K + \epsilon).$$

Proof. Suppose that P is the polytope spanned by $E(p_i)$ for some $p_i \in X$. Then it is the case that $\langle x, f\rangle \leq c$ for all $x \in P$ if and only if this holds for all $x = E(p_i)$, or if $f(p_i) \leq c$ for all i. Hence it suffices to find some finite set of $p_i \in X$ so that for each $f \in V\backslash\{0\}$, $\sup(|f|) \leq \sup_i f(p_i)(K+\epsilon)$. Notice that this condition is invariant under scaling f by a positive constant, so it suffices to check for f on the unit sphere of V.

Notice that by the definition of K, that for each such f, there is a $p \in X$ so that $\sup(|f|) \leq f(p)K$. Notice that for such a p, $\sup(|g|) \leq g(p)(K+\epsilon)$ for g in some open neighborhood of f. Hence these p define an open cover of the unit ball of V, and by compactness there must exist a finite set of p_i so that for each such f, $\sup(|f|) \leq f(p_i)(K+\epsilon)$ for some i. This completes our proof. □

Throughout the rest of this section we will use ϵ and P to refer to a positive real number and a polytope in V^* satisfying the conditions from Proposition 5. We now construct our function F.

Proposition 6. *If $\epsilon < \frac{N}{M-1} - K - 1$, there exists a continuous function $F : P \to V^*$ so that*

- *For each $x \in P$ there are points $q_1, \ldots, q_N \in X$ so that $F(x) = \sum_{i=1}^N E(q_i)$*
- *For each facet, T, defined by the equation $L(x) = c$ for some linear function L on V^* and some $c \in \mathbb{R}^+$, $L(F(T)) \subset \mathbb{R}^+$*

Proof. For a real number x, let $\lfloor x \rfloor$ denote the greatest integer less than x and let $\{x\} = x - \lfloor x \rfloor$ denote the fractional part of x.

Let p_i be points in X so that $P_i = E(p_i)$ are the vertices of P. Let p_0 be some particular point in X. Since X is path-connected we can produce continuous paths $\gamma_i : [0,1] \to X$ so that $\gamma_i(0) = p_0$ and $\gamma_i(1) = p_i$. For $r \in [0,1]$ a real number, we use $[rP_i]$ to denote $E(\gamma_i(r))$. We let $[0] = E(p_0) = [0P_i]$. We also note that $[P_i] := [1P_i] = P_i$ and that $[rP_i]$ is continuous in r.

Next pick a triangulation of P. Our basic idea will be as follows: for any $Q \in P$, if Q is in the simplex in our triangulation defined by $P_{n_0}, P_{n_1}, \ldots, P_{n_d}$ for some n_i and $d \leq M$. We can write Q uniquely as $\sum_{i=0}^d x_i[P_{n_i}]$ for $x_i \in [0,1]$ with $\sum_i x_i = 1$ (here we think of the sum as being a sum of points in V^*). The idea is that $F(Q)$ should be approximately $NQ = \sum_{i=0}^d Nx_i[P_{n_i}]$. If the Nx_i are all integers, this is just a sum of N points. Otherwise, we need to smooth things out some.

Let S be the set of $i \in \{0, \ldots, d\}$ so that $\{Nx_i\} \geq 1 - 1/(3M)$. Define

$$F(x) := \sum_{i=0}^{d} \left(\lfloor Nx_i \rfloor\right)[P_{n_i}] + \sum_{i \in S}[(1 - 3M(1 - \{Nx_i\})) \cdot P_{n_i}] + \left(N - \sum_{i=0}^{d} \lfloor Nx_i \rfloor - |S|\right)[0].$$

We have several things to check. First, we need to check that F is well defined. Next, we need to check that F is continuous. Finally, we need to check that F has the desired properties.

We must first show that F is well defined. We have defined it on each simplex of our triangulation, but we must show that these definitions agree on the intersection of two simplices. It will be enough to check that if Q is in the simplex defined by $P_{n_0}, \ldots, P_{n_d}$ and the simplex defined by $P_{n_0}, \ldots, P_{n_d}, P_{n_{d+1}}$, that our two definitions of $F(Q)$ agree (because then all definitions of $F(Q)$ agree with the definition coming from the minimal simplex containing Q). In this case, if we write $Q = \sum_{i=0}^{d} x_i P_{n_i} = \sum_{i=0}^{d+1} y_i P_{n_i}$, then it must be the case that $x_i = y_i$ for $i \leq d$ and $y_{d+1} = 0$. It is easy to check that our two definitions of F on this intersection agree on Q.

For continuity, we need to deal with several things. Firstly, by the above, since F can be defined independently on each simplex in our decomposition of P in such a way that they glue on the boundaries, we only need to check that F is continuous on any given simplex. In this case, we may write $F(Q) = F(x_0, \ldots, x_d)$. We also note that we can write $F(Q) = N[0] + \sum_{i=0}^{d} F_i(Nx_i)$ where $F_i(y)$ is

$$\begin{cases} (\lfloor y \rfloor) \cdot ([P_{n_i}] - [0]) & \text{if } \{y\} < 1 - 1/(3M) \\ (\lfloor y \rfloor) \cdot ([P_{n_i}] - [0]) + [(1 - 3M(1 - \{y\})) \cdot P_{n_i}] & -[0] \text{ else} \end{cases}.$$

We now have the check continuity of F_i. Note that F_i is clearly continuous except where y is either an integer or an integer minus $1/(3M)$. For integer n, as y approaches n from below, $F_i(y) = (n-1)([P_{n_i}] - [0]) + [(1 - 3M(n - y)) \cdot P_{n_i}] - [0] \to n([P_{n_i}] - [0]) = F_i(n)$. Also as y approaches $n - 1/(3M)$ from below, $F_i(y) = (n-1)([P_{n_i}] - [0]) = F_i(n - 1/(3M))$. Hence F is continuous.

Next we need to check that for any Q that $F(Q)$ is a sum of N elements of $E(X)$. From the definition it is clear that $F(Q)$ is sum of elements of $E(X)$ with integer coefficients that add up to N. Hence, we just need to check that all of these coefficients are positive. This is obvious for all of the coefficients except for $N - |S| - \sum_{i=0}^{d} \lfloor Nx_i \rfloor$. Hence, we need to show

that $N \geq |S| + \sum_{i=0}^{d} \lfloor Nx_i \rfloor$. Since $\sum_{i=0}^{d} x_i = 1$ by assumption,

$$\begin{aligned} N &= \sum_{i=0}^{d} Nx_i \\ &= \sum_{i=0}^{d} \lfloor Nx_i \rfloor + \{Nx_i\} \\ &\geq \sum_{i=0}^{d} \lfloor Nx_i \rfloor + \sum_{i \in S} \{Nx_i\} \\ &\geq \sum_{i=0}^{d} \lfloor Nx_i \rfloor + |S|(1 - 1/(3M)) \\ &= |S| + \sum_{i=0}^{d} \lfloor Nx_i \rfloor - |S|/(3M). \end{aligned}$$

Since N and $|S| + \sum_{i=0}^{d} \lfloor Nx_i \rfloor$ are both integers and $|S|/(3M) \leq (M+1)/(3M) < 1$, this implies that $N \geq |S| + \sum_{i=0}^{d} \lfloor Nx_i \rfloor$.

Finally, suppose that T is some facet of P defined by $L(x) = c > 0$ and that Q lies on T. Since $(V^*)^* = V$ there is a function $f \in V$ so that $L(x) = \langle x, f \rangle$ for all $x \in V^*$. Let Q be in the simplex defined by $P_{n_0}, \ldots, P_{n_d}$ where $P_{n_i} \in T$ and $d \leq M - 1$. We need to show that $L(F(Q)) > 0$. Recall by the construction of P that for any $p \in X$ that $|f(p)| \leq c(K + \epsilon)$. Equivalently $|L(E(p))| \leq c(K + \epsilon)$. Note also that since the P_{n_i} are on T that $L(P_{n_i}) = c$. Now if $Q = \sum x_i P_{n_i}$, $F(Q)$ is a sum of N points of $E(X)$ at least $\sum_i \lfloor Nx_i \rfloor$ of which are one of the P_{n_i}. Note that $N - \sum_i \lfloor Nx_i \rfloor = \sum_i \{Nx_i\} < \sum_{i=0}^{d} 1 = d + 1 \leq M$. Therefore since this term is an integer, $N - \sum_i \lfloor Nx_i \rfloor \leq M - 1$. Hence $F(Q)$ is a sum of $N - M + 1$ of the P_{n_i} (with multiplicity) plus the sum of $M - 1$ other points in $E(X)$. Hence

$$L(F(Q)) \geq (N - M + 1)c - (M - 1)(K + \epsilon)c \geq c[N - (M - 1)(K + 1 + \epsilon)] > 0.$$

This completes our proof. □

To finish the proof of Theorem 4 we will use the following:

Proposition 7. *Let Q be a polytope in a finite dimensional vector space U with 0 in the interior of Q. Let $F : Q \to U$ be a continuous function so that for any facet T of Q defined by the linear equation $L(x) = c$, with $c > 0$, $L(F(T)) \subset \mathbb{R}^+$, then $0 \in F(Q)$.*

Proof. Suppose without loss of generality that Q spans $U = \mathbb{R}^n$. Suppose for sake of contradiction that $0 \notin F(Q)$. Consider the map $f : B^n \to Q$ defined by letting $f(0) = 0$ and otherwise $f(x) = m_x x$ where m_x is the unique positive real number so that $\frac{m_x x}{|x|} \in \partial Q$. Next

consider $g : Q \to S^{n-1}$ defined by $g(x) = \frac{F(x)}{|F(x)|}$. Composing we get a map $g \circ f : B^n \to S^{n-1}$. Since the map extends to the whole ball, $g \circ f : S^{n-1} \to S^{n-1}$ must be contractible. We use our hypothesis on F to show that this map is actually degree 1 and reach a contradiction.

First, we claim that for no $x \in S^{n-1}$ is $g(f(x)) = -x$. This is because $f(x) \in \partial Q$. Let $f(x)$ land in a facet, T, defined by $L(y) = c > 0$. We have that $L(x) > 0$, because $f(x)$ is a positive multiple of x. We also have that $L(g(f(x))) > 0$ because $g(f(x))$ is a positive multiple of a point in $F(T)$. This proves the claim.

Finally, we claim that any map $h : S^{n-1} \to S^{n-1}$ that sends no point to its antipodal point is degree 1. This is because there is a homotopy from h to the identity by moving each $h(x)$ at a constant rate along the arc from $-x$ to $h(x)$ to x. □

Finally, we can prove Theorem 4

Proof. We construct the polytope P as in Proposition 5 with $\epsilon < \frac{N}{M-1} - K - 1$, and F as in Proposition 6. Then by Proposition 7 we have that 0 is in the image of F. Since every point in the image of F is a sum of N points of $E(X)$, we have a design of size N by Lemma 2. □

4 Tightness of the Bound

In this Section, we demonstrate that in the generality in which it is stated, the lower bound for N in Theorem 4 is tight. First off, we note that although it is possible that K is infinite, this might be indicative of the non-existence of designs of any size.

Proposition 8. *Let $\alpha \in (0, 1)$ be an irrational number. Consider the topological design problem*

$$(X, \mu, W) = ([0, 1], \alpha \cdot \delta(x - 1) + (1 - \alpha) \cdot \delta(x), \textit{Polynomials of degree at most 4}).$$

Then there is no unweighted design for (X, μ, W) of any size.

Proof. Note that for $f(x) = x^2(1 - x)^2$, $\int_X f(x)d\mu(x) = 0$. Note that for this f, $\sup(f) > 0$ and $\inf(f) = 0$, so $K = \infty$. If we have a design $p_1, \ldots, p_N$ then it must be the case that $\sum_i f(p_i) = 0$. Therefore, since $f(x) \geq 0$ for all $x \in X$, this implies that $f(p_i) = 0$ for all i. Therefore $p_i \in \{0, 1\}$ for all i. Next consider $g(x) = x$. $\int_X g(x)d\mu(x) = \alpha$. Therefore, we must have that $\frac{1}{N} \sum g(p_i) = \alpha$. But for each i we must have $g(p_i)$ is either 0 or 1. Therefore this sum is a rational number and cannot be α, which is irrational. □

We show that even when K is finite that a path-connected topological design problem may require that its designs be nearly the size mentioned in Theorem 4. In particular, we show

Proposition 9. *Let $m > 1$ be an integer and $k \geq 1$, $\epsilon > 0$ real numbers. Then there exists a path-connected topological design problem with $M = m$ and $K \leq k + \epsilon$ that admits no design of size $(m-1)(k+1)$ or less.*

Proof. First note that by increasing the value of k by $\epsilon/2$ and decreasing ϵ by a factor of 2, it suffices to construct such a design problem that admits no design of size less than $(m-1)(k+1)$. We construct such a design problem as follows.

Let $X = [0,1]$ and let μ be the Lebesgue measure. Let $F : X \to \mathbb{R}$ be a continuous function with the following properties:

- $F(x) = k$ for $x \in [0, 1/(2k)]$
- $F(x) = -1$ for $x \in [1/2, 1]$
- $F(x) \in [-1, k]$ for $x \in X$
- $\int_X f(x) d\mu(x) = 0$

Notice that such F are not difficult to construct. Next pick $\delta > 0$ a sufficiently small real number (we will discuss how small later). Let ϕ_i for $1 \leq i \leq m-1$ be continuous real-valued function on X so that

- $\phi_i(x) \geq 0$ for all x
- $\text{supp}(\phi_i) \subset [0, 1/(4k)]$
- The supports of ϕ_i and ϕ_j are distinct for $i \neq j$
- $\sup(\phi_i) = 1$
- $\int_X \phi_i(x) d\mu(x) = 2\delta$

It is not hard to see that this is possible to arrange as long as δ is sufficiently small. Let

$$f_i(x) = \delta - \phi_i(x) + \phi_i(2(1-x)).$$

It is easy to see that $\int_X f_i(x) d\mu(x) = 0$. We let W be the span of F and the f_i.

Since all elements of W already have 0 integral, we have that $V = W$ so $M = \dim(W)$. The F and the f_i are clearly linearly independent, and hence $M = m$.

We now need to bound K. Consider an element of V of the form $G = aF + \sum a_i f_i$. It is easy to see that G's values on $[1/(2k), 1 - 1/(4k)]$ are sandwiched between its values on the rest of X. Hence G attains its sup and inf on $[0, 1/(2k)] \cup [1 - 1/(4k), 1]$. Let $s = \sum_i a_i$. We then have that $G(x) = ak + s\delta - \sum a_i \phi_i(x)$ on $[0, 1/(2k)]$ and $G(x) = -a + s\delta + \sum a_i \phi_i(2(1-x))$ on $[1/2, 1]$. Therefore,

$$\sup(G) = \max(ak + s\delta - \min(a_i, 0), -a + s\delta + \max(a_i, 0)),$$

$$\inf(G) = \min(ak + s\delta - \max(a_i, 0), -a + s\delta + \min(a_i, 0)).$$

Suppose for sake of contradiction that $\frac{\sup(G)}{|\inf(G)|} > k + \epsilon$. This means that $\sup(G) + (k + \epsilon)\inf(G) > 0$. If $\sup(G) = ak + s\delta - \min(a_i, 0)$ this is at most

$$\begin{aligned}
&ak + s\delta - \min(a_i, 0) + (k + (k/(k+1))\epsilon)(-a + s\delta + \min(a_i, 0)) \\
&+ \epsilon/(k+1)(ak + s\delta - \max(a_i, 0)) \\
\leq &(k+1)s\delta - \epsilon/(k+1)\max(a_i, 0) \\
\leq &(k+1)(m-1)\max(a_i, 0)\delta - \epsilon/(k+1)\max(a_i, 0)
\end{aligned}$$

which is negative for δ sufficiently small.

If on the other hand, $\sup(G) = -a + s\delta + \max(a_i, 0)$, then $\sup(G) + (k + \epsilon)\inf(G)$ is at most

$$\begin{aligned}
&- a + s\delta + \max(a_i, 0) + (1 + \epsilon/(k+1))(ak + s\delta - \max(a_i, 0)) \\
&+ (k - 1 + k\epsilon/(k+1))(-a + s\delta + \min(a_i, 0)) \\
\leq &(k + 1 + \epsilon)s\delta - \epsilon\max(a_i, 0)/(k+1) \\
\leq &(k + 1 + \epsilon)(m-1)\max(a_i) - \epsilon\max(a_i, 0)/(k+1)
\end{aligned}$$

which is negative for δ sufficiently small, yielding a contradiction.

Hence, if we picked δ sufficiently small $\frac{\sup(G)}{|\inf(G)|} \leq k + \epsilon$ for all $G \in V$, so $K \leq k + \epsilon$.

Next suppose that we have a design $x_1, \ldots, x_N$ for this design problem. Since $\sum f_j(x_i) = 0$ and since f_j is negative only on the support of ϕ_j, we must have at least $m - 1$ of the x_i each in a support of one of the ϕ_j, and hence there must be at least $m - 1$ x_i in $[0, 1/(2k)]$. Next we note that we must also have $\sum F(x_i) = 0$. At least $m - 1$ of these x_i are in $[0, 1/(2k)]$ and therefore F of these x_i equals k. Therefore since $F(x_j) \geq -1$ for each other j, there must be at least $k(m-1)$ other points in our design. Hence N must be at least $k(m-1) + (m-1) = (m-1)(k+1)$. □

5 The Bound for Homogeneous Spaces

In this Section, we show that there is a much nicer bound on the size of designs if we have a homogenous, path-connected, topological design problem.

Theorem 10. *Let (X, μ, W) be a homogeneous topological design problem with $M > 1$. Then for any $N > M(M-1)$, there exists a design for X of size N. Furthermore, there exists a design for X of size at most $M(M-1)$.*

We will show that $K \leq (M-1)$ where the equality is strict unless X has a design of size M. An application of Theorem 4 then yields our result.

We begin with a Lemma

Lemma 11. *If X is a homogenous topological design problem, and if p_i, w_i is a weighted design for X, then $K \leq \frac{1-\max w_i}{\max(w_i)}$.*

Proof. WLOG $w_1 = \max(w_i)$. Suppose for sake of contradiction that $K > \frac{1-w_1}{w_1}$. This means that there is an $f \in V$ so that $\frac{\sup(f)}{|\inf(f)|} > \frac{1-w_1}{w_1}$. This means that there is a $p \in X$ so that $w_1 f(p) + (1-w_1)\inf(f) > 0$. Since X is homogenous, there is a $g : X \to X$ preserving all properties of the design problem so that $g(p_1) = p$. Since g preserves μ and W, $g(p_i), w_i$ must also be a weighted design for X. Therefore, $\sum_i w_i f(g(p_i)) = 0$. But on the other hand this is

$$w_1 f(p) + \sum_{i>1} w_i f(g(p_i)) \geq w_1 f(p) + (1-w_1)\inf(f) > 0.$$

Hence we arrive at a contradiction. □

We note the following interesting pair of Corollaries.

Corollary 12. *If X is a homogeneous topological design problem, and p_i, w_i a weighted design for X, then* $\max(w_i) \leq \frac{1}{K+1}$.

Corollary 13. *If X is a homogeneous topological design problem, X admits no weighted design of size less than $K+1$.*

We will also need one more Lemma

Lemma 14. *If X is a path-connected topological design problem and $M > 0$, X has a weighted design of size at most M.*

Proof. Suppose for sake of contradiction that there is no such weighted design. Then it must be the case that there are no $p_i \in X$ and $w_i \geq 0$ for $1 \leq i \leq M$ so that $\sum_i w_i E(p_i) = 0$. This means that whenever a non-negative linear combination of $M+1$ values of $E(p_i)$ equals 0, the weights must be all 0 or all positive. By Lemma 3 there must be some $M+1$ points with a non-negative linear combination equals 0. As we deform our set of points, it will always be the case that some linear combination equals 0 by a dimension count. Furthermore, the coefficients of this combination will vary continuously. Since by assumption it is never possible to write 0 as a non-negative linear combination with at least one coefficient equal to 0, it must be the case that no matter how we deform the p_i, there will always exist a linear combination equal to 0 with strictly positive coefficients. But this is clearly not the case if all of the p_i are equal to some point p on which not all of the functions in V vanish. □

We can now prove Theorem 10.

Proof. By Lemma 14, there is a weighted design for X of size at most M. If all of the weights are equal, this is a design of size M, and by Lemma 11 $K \leq \frac{1-1/M}{1/M} = M-1$ and the remainder of the result follows from Theorem 4. If the weights of this design are not equal, some weight is larger than $\frac{1}{M}$, and hence $K < \frac{1-1/M}{1/M} = M-1$, and again our result follows from Theorem 4. □

5.1 Examples

We provide several Corollaries of Theorem 10.

Corollary 15. *There exists a spherical design of strength n on the $d-1$-dimensional sphere of size $O(n^{2d-2}/((d-1)!^2))$.*

Corollary 16. *There exists a design of strength n on the Grassmannian, $G(m,k)$ of size $O_{m,k}(n^{2k(m-k)})$.*

5.2 Conjectures

Although we prove a bound of size $O(M^2)$ for homogeneous path-connected topological design problems, it feels like the correct result should be $O(M)$, since that is roughly the number of degrees of freedom that you would need. We can rephrase the problem for homogeneous path-connected spaces a little though. First, we may replace X by $E(X)$, which is a bounded subset of V^*. Next, we note that the L^2 measure on V is preserved by the symmetries of X. Hence, the symmetry group G of X (which is transitive by assumption) is a subgroup of $O(V^*)$, and hence compact. Since X is a quotient of the identity component G_0 of G we may pull our design problem back to one on G_0 (using the pullbacks of μ and W). Since the pullback of μ to G_0 is clearly the Haar measure, and is also finite, G_0 must be compact. Since G_0 is also a path-connected subgroup of $O(V^*)$, it must be a Lie group. Hence we have reduced the problem of finding a design in a path-connected homogenous topological design problem to finding one in a design problem of the following form:

$X = G$ is a compact Lie Group. μ is the normalized Haar measure for G. W is a finite dimensional space of left-invariant functions on G. Since $L^2(G)$ decomposes as a sum $\bigoplus_{\rho_i \in \hat{G}} \phi_i \otimes \phi_i^*$, W must be a sum of the form $\bigoplus_{\rho_i \in \hat{G}} \rho_i \otimes W_i$ where W_i is a subspace of ρ_i^* and all but finitely many ρ_i are 0.

Note that although we have all this structure to work with, proving better bounds even for the circle seems to be non-trivial. This Conjecture applied to S^1 says that given any M distinct non-zero integers n_i that there exist $O(M)$ complex numbers z_j with $|z_j| = 1$ so that $\sum_j z_j^{n_i} = 0$ for all i.

6 Designs on the Interval

Let I be the interval $[-1,1]$. For $\alpha, \beta \geq -\frac{1}{2}$ let $\mu_{\alpha,\beta}$ be the measure $\frac{(1-x)^\alpha (1+x)^\beta \Gamma(\alpha+\beta+2)}{2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)}dx$ on I. Let $\mathcal{P}_n$ be space of polynomials of degree at most n on I. We will prove the following Theorem:

Theorem 17. *The size of the smallest design for $(X, \mu_{\alpha,\beta}, \mathcal{P}_n)$ is of size $\Theta_{\alpha,\beta}(n^{2\max(\alpha,\beta)+2})$.*

In order to prove this Theorem, we will first need to review some basic facts about Jacobi polynomials. We will use [59] as a guide.

Definition. *We define the Jacobi polynomials inductively as follows: For n a non-negative integer and $\alpha, \beta \geq -\frac{1}{2}$, $P_n^{(\alpha,\beta)}(x)$ is the unique degree n polynomial with*

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$$

and so that $P_n^{(\alpha,\beta)}$ is orthogonal to $P_k^{(\alpha,\beta)}$ for $k < n$ with respect to the inner product $\langle f, g\rangle = \int_I f(x)g(x)d\mu_{\alpha,\beta}(x)$.

Hence the $P_n^{(\alpha,\beta)}$ are a set of orthogonal polynomials for the measure $\mu_{\alpha,\beta}$. The normalization is given by [59] Equation (4.3.3)

$$\begin{aligned}
&\int_I (P_n^{(\alpha,\beta)})^2 d\mu_{\alpha,\beta} \\
&\quad = \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)\Gamma(\alpha+\beta+2)}{(2n+\alpha+\beta+1)\Gamma(n+1)\Gamma(n+\alpha+\beta+1)\Gamma(\alpha+1)\Gamma(\beta+1)} \\
&\quad = \Theta_{\alpha,\beta}(n^{-1}).
\end{aligned} \tag{4}$$

Hence we define the normalized orthogonal polynomials

$$\begin{aligned}
&R_n^{(\alpha,\beta)} \\
&\quad = P_n^{(\alpha,\beta)} \sqrt{\frac{(2n+\alpha+\beta+1)\Gamma(n+1)\Gamma(n+\alpha+\beta+1)\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)\Gamma(\alpha+\beta+2)}} \\
&\quad = P_n^{(\alpha,\beta)} \Theta_{\alpha,\beta}(\sqrt{n}).
\end{aligned}$$

We will also need some more precise results on the size of these polynomials. In particular we have Theorem 8.21.12 of [59] which states that

$$\begin{aligned}
\left(\sin\frac{\theta}{2}\right)^{\alpha} \left(\cos\frac{\theta}{2}\right)^{\beta} P_n^{(\alpha,\beta)}(\cos\theta) &= \frac{N^{-\alpha}\Gamma(n+\alpha+1)}{n!} \sqrt{\frac{\theta}{\sin\theta}} J_\alpha(N\theta) \\
&\quad + \begin{cases} \theta^{1/2} O(n^{-3/2}) & \text{if } cn^{-1} \leq \theta \leq \pi - \epsilon \\ \theta^{\alpha+2} O(n^{\alpha}) & \text{if } 0 < \theta \leq cn^{-1} \end{cases}
\end{aligned} \tag{5}$$

for any positive constants c and ϵ and where $N = n + (\alpha+\beta+1)/2$, and J_α is the Bessel function.

We will also want some bounds on the size of the Bessel functions. From [59] (1.71.10) and (1.71.11) we have that for $\alpha \geq -\frac{1}{2}$

$$J_\alpha(x) \sim c_\alpha(x^\alpha) \text{ as } x \to 0$$

and

$$J_\alpha(x) = O_\alpha(x^{-1/2}).$$

The first of these along with Equation 5 implies that $P_n^{(\alpha,\beta)}$ has no roots within $O_{\alpha,\beta}(n^{-2})$ of 1. Noting that $P_n^{(\alpha,\beta)}(x)$ and $P_n^{(\beta,\alpha)}(-x)$ are constant multiples of each other, $P_n^{(\alpha,\beta)}(x)$ also has no roots within $O_{\alpha,\beta}(n^{-2})$ of -1. Applying Theorem (8.21.13) of [59], we also find that $P_n^{(\alpha,\beta)}$ has roots within $O_{\alpha,\beta}(n^{-2})$ of either endpoint. Applying the second equation above we find that for $x \in I$

$$R_n^{(\alpha,\beta)}(x) = O_{\alpha,\beta}\left((1-x)^{-\alpha/2-1/4}(1+x)^{-\beta/2-1/4}\right). \tag{6}$$

Lemma 18. *Let μ be a normalized measure on I. Let R_n^μ be the sequence of orthogonal polynomials for μ. (i.e. R_n^μ is a polynomial of degree n, and $\{R_0^\mu, R_1^\mu, \ldots, R_n^\mu\}$ is an orthonormal basis for $\mathcal{P}_n$ with the inner product $\langle f, g\rangle_\mu = \int_I f(x)g(x)d\mu(x)$.) Let r_i be the roots of $R_n^\mu(x)$. Let $w_i = \frac{1}{\sum_{j=0}^{n-1}(R_j^\mu(r_i))^2}$. Then (w_i, r_i) is a weighted design for $(I, \mu, \mathcal{P}_{2n-1})$. (Note that this is just Gauss-Jacobi Quadrature)*

Proof. First, note that if $f \in \mathcal{P}_{2n-1}$ vanishes on all of the r_i that $\int_I f(x)d\mu = 0$. This is because f must be of the form $R_n^\mu(x)g(x)$ for some $g \in \mathcal{P}_{n-1}$. Since R_n^μ is orthogonal to $\mathcal{P}_{n-1}$, this is 0. Therefore, the map $\mathcal{P}_{2n-1} \to \mathbb{R}$ given by $f \to \int_I f d\mu$ factors through the map $\mathcal{P}_{2n-1} \to \mathbb{R}^n$ given by $f \to (f(r_i))_{1\le i\le n}$. Therefore, there exist some weights w_i' so that for any $f \in \mathcal{P}_{2n-1}$,

$$\int_I f d\mu = \sum_{i=1}^n w_i' f(r_i).$$

Let $f_i \in \mathcal{P}_{n-1}$ be the unique polynomial of degree $n-1$ so that $f_i(r_j) = \delta_{i,j}$. Writing f_i in terms of the R's we find that

$$\begin{aligned} f_i(x) &= \sum_{j=0}^{n-1} R_j^\mu(x)\left(\int_I f_i(y)R_j^\mu(y)d\mu(y)\right) \\ &= \sum_{j=0}^{n-1} R_j^\mu(x)\left(\sum_{k=1}^n w_k' f_i(r_k)R_j^\mu(r_k)\right). \end{aligned}$$

Evaluating at $x = r_i$ we find that

$$\begin{aligned} 1 &= \sum_{j=0}^{n-1} R_j^\mu(r_i) \left(\sum_{k=1}^{n} w_k' \delta_{i,k} R_j^\mu(r_k) \right) \\ &= \sum_{j=0}^{n-1} R_j^\mu(r_i) w_i' R_j^\mu(r_i) \\ &= w_i' \sum_{j=0}^{n-1} (R_j^\mu(r_i))^2. \end{aligned}$$

Thus $w_i' = w_i$, completing out proof. □

We are now prepared to show that all designs for $(I, \mu_{\alpha,\beta}, \mathcal{P}_n)$ are reasonably large.

Proposition 19. *If $\alpha, \beta \geq -\frac{1}{2}$ then all unweighted designs for $(I, \mu_{\alpha,\beta}, \mathcal{P}_n)$ have size $\Omega_{\alpha,\beta}(n^{2\alpha+2})$.*

Proof. We increase n by a factor of 2, and instead prove bounds on the size of designs for $(I, \mu_{\alpha,\beta}, \mathcal{P}_{2n})$.

Let r_n be the biggest root of $R_n^{(\alpha,\beta)}$. Since $p(x) = \frac{\left(R_n^{(\alpha,\beta)}(x)\right)^2}{(x-r_n)}$ is $R_n^{(\alpha,\beta)}(x)$ times a polynomial of degree less than n, $\int_I p d\mu_{\alpha,\beta} = 0$. Since $p(x)$ is positive outside of $[r_n, 1]$, any design must have a point in this interval. Therefore any design must have at least one point in $[1 - O_{\alpha,\beta}(n^{-2}), 1]$. If such a point is written as $\cos\theta$ then $\theta = O_{\alpha,\beta}(n^{-1})$. For c a sufficiently small constant (depending on α and β), define

$$f(x) = \frac{\left(\sum_{i=cn}^{2cn} R_i^{(\alpha,\beta)}(x)\right)^2}{cn}.$$

It is clear that $f(x) \geq 0$ for all x, and clear from the orthonormality that $\int_I f d\mu_{\alpha,\beta} = 1$. On the other hand, for c sufficiently small and $cn \leq i \leq 2cn$, Equation 5 tells us that

$$R_i^{(\alpha,\beta)}(x) = \Omega_{\alpha,\beta}(n^{\alpha+1/2})$$

on $[1 - r_n, 1]$. Therefore

$$f(x) = \Omega_{\alpha,\beta}(n^{2\alpha+2})$$

on $[1 - r_n, 1]$. Therefore if $p_1, \ldots, p_N$ is a design for $(I, \mu_{\alpha,\beta}, \mathcal{P}_n)$, we may assume that $p_1 \in [1 - r_n, 1]$ and we have that

$$\begin{aligned} 1 &= \frac{1}{N} \sum_{i=1}^{N} f(p_i) \\ &\geq \frac{f(p_1)}{N} \\ &\geq \Omega_{\alpha,\beta}(n^{2\alpha+2} N^{-1}). \end{aligned}$$

Therefore $N = \Omega(n^{2\alpha+2})$. □

In order to prove the upper bound, we will use a slightly more sophisticated version of our previous techniques. First, we will need to define some terminology.

Definition. *Let $f : [0,1] \to \mathbb{R}$ we define $Var(f)$ to be the total variation of f on $[0,1]$. For $\gamma : [0,1] \to X$ and $f : X \to \mathbb{R}$, we define $Var_\gamma(f) = Var(f \circ \gamma)$.*

Definition. *For a design problem (X, μ, W) and a map $\gamma : [0,1] \to X$ we define*

$$K_\gamma = \sup_{f \in V \backslash \{0\}} \left(\frac{Var_\gamma(f)}{\sup_{\gamma([0,1])}(f)} \right).$$

We note that replacing f by $g = \frac{\sup_{\gamma([0,1])}(f) - f}{\sup_{\gamma([0,1])}(f)}$, we have that $g \geq 0$ on $\gamma([0,1])$, $\int_X g = 1$, and $\mathrm{Var}_\gamma(g) = \frac{\mathrm{Var}_\gamma(f)}{\sup_{\gamma([0,1])}(f)}$. Hence we have the alternative definition

$$K_\gamma = \sup_{\substack{g \in W \oplus 1 \\ g \geq 0 \text{ on } \gamma([0,1]) \\ \int_X g d\mu = 1}} \mathrm{Var}_\gamma(g).$$

Or equivalently, scaling g by an arbitrary positive constant,

$$K_\gamma = \sup_{\substack{g \in W \oplus 1 \\ g \geq 0 \text{ on } \gamma([0,1])}} \frac{\mathrm{Var}_\gamma(g)}{\int_X g d\mu}.$$

Proposition 20. *Let (X, μ, W) be a topological design problem with $M > 0$. Let $\gamma : [0,1] \to X$ be a continuous function with K_γ finite. Then for any integer $N > K_\gamma/2$ there exists a design for (X, μ, W) of size N.*

Proof. Let $\frac{2N}{K_\gamma} - 1 > \epsilon > 0$. For every $f \in V \backslash \{0\}$, there exists an $x \in [0,1]$ so that $K_\gamma f(\gamma(x))(1+\epsilon) > \mathrm{Var}_\gamma(f)$. Since this property is invariant under scaling of f by positive real numbers, and since it must also hold for some open neighborhood of f, by compactness, we may pick finitely many x_i so that for any $f \in V \backslash \{0\}$,

$$K_\gamma \max_i f(\gamma(x_i)) > (1-\epsilon)\mathrm{Var}_\gamma(f).$$

Let P be the polytope in V^* spanned by the points $E(\gamma(x_i))$. We will define a function $F : P \to V^*$ with the following properties:

- F is continuous
- For each $x \in P$, $F(x)$ can be written as $\sum_{i=1}^N E(\gamma(y_i))$ for some $y_i \in [0,1]$

- For each facet T of P defined by $L(x) = c > 0$, $L(F(T)) \subset \mathbb{R}^+$

Once we construct such an F, we will be done by Proposition 7.

Suppose that our set of x_i is $x_1 < x_2 < \ldots < x_R$. We first define a continuous function $C : P \to \mathbb{R}^R$ whose image consists of points with non-negative coordinates that add to 1. This is defined as follows. First, we triangulate P. Then for $y \in P$ in the simplex spanned by, say, $\{E(\gamma(x_{i_1})), \ldots, E(\gamma(x_{i_k}))\}$. We can then write y uniquely as $\sum_{j=1}^{k} w_j E(\gamma(x_{i_j}))$ for $w_j \geq 0$ and $\sum_j w_j = 1$. We then define $C(y)$ to be w_j on its i_j coordinate for $1 \leq j \leq k$, and 0 on all other coordinates. This map is clearly continuous within a simplex and its definitions on two simplices agree on their intersection. Therefore, C is continuous.

For $w \in \mathbb{R}^R$ with $w_i \geq 0$ and $\sum_i w_i = 1$, we call w_i a set of weights for the x_i. Given such a set of weights define $u_w : [0, 1] \to [0, N + 1]$ to be the increasing, upper semi-continuous function

$$u_w(x) = x + N \sum_{x_i \leq x} w_i.$$

For integers $1 \leq i \leq N$ define

$$p_i(w) = \inf\{x : u_w(x) \geq j\}.$$

Note that $p_i(w)$ is continuous in w. This is because if $|w - w'| < \delta$ (in the L^1 norm) then $|u_w(x) - u_{w'}(x)| < \delta$ for all x. Therefore, since $u_{w'}(x + \delta) \geq u_{w'}(x) + \delta$ we have that $|p_i(w) - p_i(w')| \leq \delta$. We now define F by

$$F(y) = \sum_{i=1}^{N} E(\gamma(p_i(C(y)))).$$

This function clearly satisfies the first two of our properties, we need now to verify the third. Suppose that we have a face of P defined by the equation $\langle f, y \rangle = 1$ for some $f \in V$. We then have that $\sup_i(f(\gamma(x_i))) = 1$. Therefore $\mathrm{Var}_\gamma(f) < K_\gamma(1 + \epsilon)$. Let this face of P be spanned by $E(\gamma(x_{i_1})), \ldots, E(\gamma(x_{i_M}))$ for $i_1 < i_2 < \ldots < i_M$. It is then the case that $f(\gamma(x_{i_j})) = 1$ for each j. Letting $w = C(y)$, it is also the case that w_k is 0 unless k is one of the i_j.

Note that $\lim_{x \to x_{i_1}^-} u_w(x) < 1$ and $u(x_{i_M}) > N$. This implies that none of the $p_i(w)$ are in $[0, x_{i_1})$ or $(x_{i_M}, 1]$. Additionally, note that

$$\lim_{x \to x_{i_{n+1}}^-} u(x) - u(i_n) = x_{i_{n+1}} - x_{i_n} < 1.$$

This implies that there is at most one p_i in $(x_{i_n}, x_{i_{n+1}})$ for each n. For a point x in this interval we have that $|f(\gamma(x)) - 1|$ is at most half of the total variation of $f \circ \gamma$ on $[i_n, i_{n+1}]$.

All other $p_i(w)$ must be one of the x_{i_j}. Therefore summing over all $p_i(w)$, we get that

$$|N - f(F(y))| = \left|N - \sum_{i=1}^{N} f(\gamma(p_i(w)))\right| \leq \sum_{i=1}^{N} |1 - f(\gamma(p_i(w)))|$$

which is at most half of the variation of $f \circ \gamma$ on $[x_{i_1}, x_{i_M}]$. This in turn is at most $\frac{K_\gamma(1+\epsilon)}{2} < N$. Therefore $f(F(y)) > 0$. This proves that F has the last of the required properties and completes our proof. □

In order to prove the upper bound for Theorem 17, we will apply this proposition to $\gamma : [0,1] \to I$ defined by $\gamma(x) = 2x - 1$. We begin with the case of $\alpha = \beta = -\frac{1}{2}$.

Lemma 21. *For $(I, \mu_{-1/2,-1/2}, \mathcal{P}_n)$ and γ as described above, $K_\gamma = O(n)$.*

Proof. We will use the alternative definition of K_γ. If $f \geq 0$ on $\gamma([0,1]) = I$ and $\int f d\mu_{-1/2,-1/2} = 1$, then f must be a sum of squares of polynomials of degree at most $n/2 + 1$ plus $(1 - x^2)$ times a sum of such polynomials. Since $\int_I f d\mu$ is linear and $\mathrm{Var}_\gamma(f)$ sublinear, it suffices to check for $f = g^2$ or $f = (1 - x^2)g^2$. Note that $\mu_{-1/2,-1/2}$ is the projected measure from the circle to the interval. Therefore, we can pull f back to a function on the circle either of the form $g(\cos\theta)^2$ or $(\sin\theta g(\cos\theta))^2$. In either case, $\int_{S^1} f(\theta)d\theta = 1$ and $f(\theta) = h(\theta)^2$ for some polynomial h of degree $O(n)$. It suffices to bound the variation of f on the circle. In particular it suffices to show that $\int_{S^1} |f'(\theta)|d\theta = O(n)$.

We note that

$$|h|_2^2 = \int_{S^1} h^2(\theta)d\theta = \int_{S^1} f(\theta)d\theta = 1.$$

We also note that

$$\int_{S^1} |f'(\theta)|d\theta = 2\int_{S^1} |h(\theta)h'(\theta)|d\theta \leq 2|h|_2|h'|_2.$$

Hence it suffices to prove that for h a polynomial of degree m that $|h'|_2 = O(m)|h|_2$. This follows immediately after noting that the orthogonal polynomials $e^{ik\theta}$ diagonalize the derivative operator. □

We now relate functions for arbitrary α and β to this

Lemma 22. *Let $\alpha, \beta \geq -\frac{1}{2}$. Let $f \in \mathcal{P}_n$, $f \geq 0$ on I. Then*

$$\int_I f d\mu_{\alpha,\beta} = \Omega_{\alpha,\beta}(n^{-2\max(\alpha,\beta)-1}) \int_I f d\mu_{-1/2,-1/2}.$$

Proof. We rescale f so that $\int_I f d\mu_{-1/2,-1/2} = 1$. We let r_i be the roots of $P_{n+1}^{(-1/2,-1/2)}$. By Lemma 18, there are weights w_i making this a design for $(I, \mu_{-1/2,-1/2}, \mathcal{P}_{2n})$. By Equation 6, we have that

$$w_i = \Omega(n^{-1}).$$

We have that $\sum_i w_i f(r_i) = 1$. Therefore, since $f(r_i) \geq 0$, we have that

$$\sum_i w_i f(r_i)^2 \leq \frac{1}{\min w_i} = O(n).$$

This equals $\int_I f^2 d\mu_{-1/2,-1/2}$. Let $R \subset I$ be $R = [1 - cn^{-2}, 1] \cup [-1, -1 + cn^{-2}]$ for c a sufficiently small positive constant. Let I_R be the indicator function of the set R. Then

$$\begin{aligned}\int_I I_R^2 d\mu_{-1/2,-1/2} &= \int_R d\mu_{-1/2,-1/2} \\ &= O\left(\int_{1-cn^{-2}} (1-x)^{-1/2} dx\right) \\ &= O(\sqrt{c} n^{-1}).\end{aligned}$$

Therefore

$$\int_R f d\mu_{-1/2,-1/2} = \int_I I_R f d\mu_{-1/2,-1/2} \leq |f|_2 |I_R|_2 = O(\sqrt{c}).$$

Hence for c sufficiently small, $\int_R f d\mu_{-1/2,-1/2} \leq \frac{1}{2}$. Therefore,

$$\int_{I \backslash R} f d\mu_{-1/2,-1/2} \geq \frac{1}{2}.$$

Since the ratio of the measures $\frac{\mu_{\alpha,\beta}}{\mu_{-1/2,-1/2}} = (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2}$ is at least $\Omega_{\alpha,\beta}(n^{-2\max(\alpha,\beta)-1})$ on $I \backslash R$ we have that

$$\begin{aligned}\int_I f d\mu_{\alpha,\beta} \geq \int_{I \backslash R} f d\mu_{\alpha,\beta} &= \Omega_{\alpha,\beta}(n^{-2\max(\alpha,\beta)-1}) \int_{I \backslash R} f d\mu_{-1/2,-1/2} \\ &= \Omega_{\alpha,\beta}(n^{-2\max(\alpha,\beta)-1}).\end{aligned}$$

□

We can now extend Lemma 21 to our other measures

Lemma 23. *For $(I, \mu_{\alpha,\beta}, \mathcal{P}_n)$ and γ as above, $K_\gamma = O_{\alpha,\beta}(n^{2\max(\alpha,\beta)+2})$.*

Proof. We use the alternative description of K_γ. Let $f \in \mathcal{P}_n$ with $f \geq 0$ on I and $\int_I f d\mu_{\alpha,\beta} = 1$. By Lemma 22, $\int_I f d\mu_{-1/2,-1/2} = O_{\alpha,\beta}(n^{2\max(\alpha,\beta)+1})$. Therefore using Lemma 21,

$$\mathrm{Var}_\gamma(f) \leq O(n) O_{\alpha,\beta}(n^{2\max(\alpha,\beta)+1}) = O_{\alpha,\beta}(n^{2\max(\alpha,\beta)+2}).$$

Therefore since this holds for all such f, $K_\gamma = O_{\alpha,\beta}(n^{2\max(\alpha,\beta)+2})$. □

Theorem 17 now follows from Proposition 19, Proposition 20 and Lemma 23.

7 Spherical Designs

In this Section, we will focus on the problem of designs on a sphere. In particular for integers $d, n > 0$ let $\mathcal{D}_n^d$ denote the design problem given by the d-sphere with its standard, normalized measure, and W the space of polynomials of total degree at most n. We begin by proving lower bounds:

Theorem 24. *Any weighted design for $\mathcal{D}_n^d$ is of size $\Omega_d(n^d)$.*

Proof. Let U be the space of polynomials of degree at most $n/2$ on S^d. Note that $\dim(U) = \Omega_d(n^d)$. We claim that $K \geq M' := \dim(U)$. Pick $x \in S^d$. Let $\phi_1, \ldots, \phi_{M'}$ be an orthonormal basis of U. Let $f(y) = (\sum_i \phi_i(y)\phi_i(x))^2$. It is clear that $\int_{S^d} f d\mu = M'$. Also $f(x) = (\sum_i \phi_i(x)^2)^2$. Let $g(y) = \sum_i \phi_i(y)^2$. g is clearly invariant under the action of $SO(d+1)$, and is therefore constant. Furthermore, $\int_{S^d} g d\mu = M'$. Therefore $g(x) = M'$. Therefore $f(x) = (M')^2$. Since $f \geq 0$ on S^d, $K \geq \frac{f(x)}{\int_{S^d} f d\mu} = M'$.

Therefore since the action of $SO(d)$ makes $\mathcal{D}_n^d$ a homogeneous design problem Corollary 13 implies that any weighted design for $\mathcal{D}_n^d$ must have size at least $M' = \Omega_d(n^d)$. □

We also prove a nearly matching lower bound. Namely:

Theorem 25. *For $N = \Omega_d(n^d \log(n)^{d-1})$, there exists a design for $\mathcal{D}_n^d$ of size N.*

The proof of Theorem 25 again uses Proposition 20, but the choice of γ is far less obvious than it is when applied in Theorem 17. In fact, we will want to introduce a slight generalization of the terminology first.

Definition. *Let G be a topological graph. If $\gamma : G \to X$ and $f : X \to \mathbb{R}$ are functions, define $Var_\gamma(f)$ as follows. For each edge e of G let $\gamma_e : [0,1] \to X$ be the map γ restricted to e. Then*

$$Var_\gamma(f) := \sum_{e \in E(G)} Var_{\gamma_e}(f).$$

Note that for an embedded graph G, we will often simply refer to $\mathrm{Var}_G(f)$.

Definition. *For (X, μ, M) a design problem, G a graph, and $\gamma : G \to X$ a function, define*

$$K_\gamma = \sup_{f \in V \setminus \{0\}} \left(\frac{Var_\gamma(f)}{\sup_{\gamma(G)}(f)} \right).$$

Note that we have alternative definitions of K_γ in the same way as we did before. We will often ignore the function γ and simply define K_G for G and embedded graph in X. We note the following version of Proposition 20:

Proposition 26. *Let (X, μ, W) be a topological design problem. Let G be a connected graph and $\gamma : G \to X$ a continuous function. If K_G is finite, and $N > K_G$ is an integer, then (X, μ, W) admits a design of size N.*

Proof. Note that if we double all of the edges of G that the resulting multigraph admits an Eulerian circuit. This gives us a continuous map $\gamma' : [0, 1] \to X$ that covers each edge of G exactly twice. Therefore for every function f, $\sup_G(f) = \sup_{\gamma([0,1])}(f)$ and $\mathrm{Var}_{\gamma'}(f) = 2\mathrm{Var}_G(f)$. Hence $K_{\gamma'} = 2K_G$, and the result follows from Proposition 20. □

We will now need to prove the following:

Proposition 27. *For $d, n \geq 1$, there exists a connected graph G for the design problem $\mathcal{D}_n^d$ so that $K_G = O_d(n^d \log(n)^{d-1})$. Furthermore this can be done is such a way that the total length of all the edges of G is $n^{O_d(1)}$.*

The basic idea of the proof of Proposition 27 is as follows. First, by projecting S^d down onto its first $d-1$ coordinates, we can think of it as a circle bundle over B^{d-1}. We construct our graphs by induction on d. We pick a number of radii r_i, and place our graphs for various strength designs on the spheres of radius r_i in B^{d-1}. We also add the loops over the points on these graphs given by the corresponding designs. The first step is to show that average value of f over our loops in G is roughly the average value over the sphere (see Lemma 33). Naively, this should hold since the average value of f on the sphere of radius r_i in B^{d-1} should equal the average value of f over the appropriate loops (because the loops are arranged in a design). Our radii will themselves by arranged in an appropriate design, so that the value of f on the sphere will equal the average of the values at there radii. Unfortunately, our component designs will be of insufficient strength for this to hold. This is fixed by showing that the component of f corresponding to high degree spherical harmonics at small radius r_i in B^{d-1} is small (this is shown in Lemma 29). The bound on K_G comes from noting that the variation of f along G is given by the sum of variations on the subgraphs. These in turn are bounded by the size of f on these subgraphs, and the appropriate sum of variations is bounded by the size of f on the whole sphere.

Before we proceed, we will need the following technical results:

Lemma 28. *Let $f \in \mathcal{P}_n$. Then*

$$\sup_{S^d}(f) = O(n^{d/2})|f|_2, \qquad \sup_{S^d}|f'| = O(n^{d/2+1})|f|_2.$$

Proof. Let ϕ_i $(1 \leq i \leq M)$ be an orthonormal basis of the polynomials of degree at most n on S^d, so that each of the ϕ_i is a spherical harmonic. Note that $M = O(n^d)$. Write $f(u) = \sum a_i \phi_i(u)$. For $v \in S^d$, $f(v) = \sum a_i \phi_i(v) \leq \sqrt{\sum_i a_i^2}\sqrt{\sum_i \phi_i(v)^2} = |f|_2\sqrt{\sum_i \phi_i(v)^2}$. Now by symmetry, $\sum_i \phi_i(u)^2$ is a constant function of u. Since its average value is M, $\sum_i \phi_i(v)^2 = M$. Therefore, $f(v) \leq \sqrt{M}|f|_2$.

We also have that

$$|f'(v)| \leq \sum_i a_i |\phi_i'(v)| \leq \sqrt{\sum_i a_i^2} \sqrt{\sum_i |\phi_i'(v)|^2} = |f|_2 \sqrt{\sum_i |\phi_i'(v)|^2}.$$

Now $\sum_i |\phi_i'(u)|^2$ is a constant function of u. Its average value is

$$\begin{aligned} \int \sum_i |\phi_i'(u)|^2 du &= \sum_i \int |\phi_i'(u)|^2 du \\ &= \sum_i \int \phi_i(u) \triangle \phi_i(u) du. \end{aligned}$$

So $\triangle \phi_i(u) = k^2 \phi_i(u)$ for some $k \leq n$. Therefore, this is at most $n^2 M$. Hence, $|f'(v)| = O(n^{d/2+1})$. □

Lemma 29. *For $n \geq d, k \geq 1$ integers and f a polynomial of degree at most n on the d-disk, D, with $\int_D f^2(r)(1-r^2)^{(k-2)/2} \frac{dr}{Vol(D)} = 1$ then $\sup_{[-1,1]} f = O\left(\sqrt{\frac{2}{d\beta(d/2,k/2)}}\right) O\left(\frac{n}{d+k-1}\right)^{(d+k-1)/2}$.*

Proof. Notice that

$$\begin{aligned} \int_D (1-r^2)^{(k-2)/2} \frac{dr}{\mathrm{Vol}(D)} &= d \int_0^1 r^{d-1} (1-r^2)^{(k-2)/2} dr \\ &= d/2 \int_0^1 s^{(d-2)/2} (1-s)^{(k-2)/2} ds \\ &= d\beta(d/2, k/2)/2. \end{aligned}$$

Let μ be the measure $\frac{2dr}{\mathrm{Vol}(D) d\beta(d/2,k/2)}$. Note that μ is the projected measure from the $d+k-1$-sphere onto the d-disk. We have that $\int_D f^2(r) d\mu = \frac{2}{d\beta(d/2,k/2)}$. Rescaling f so that

$$\int_{-1}^1 f(r)^2 d\mu = 1$$

we need to show that for such f, $\sup_{[-1,1]} f = O\left(\frac{n}{d+k-1}\right)^{(d+k-1)/2}$.

Pulling f back onto the $(d+k-1)$-sphere, we get that $\int_{S^(d+k-1)} f^2(x) dx = 1$, where dx is the normalized measure on S^{d+k-1}. We need to show that for $x \in S^d$ that $f(x) = O\left(\frac{n}{d+k-1}\right)^{(d+k-1)/2}$. Let ϕ_i $(1 \leq i \leq M)$ be an orthonormal basis of the space of polynomials of degree at most n on S^{d+k-1}. We can write $f(y) = \sum_i a_i \phi_i(y)$. It must be the case that $\sum_i a_i^2 = 1$ and $f(x) = \sum_i a_i \phi_i(x)$. By Cauchy-Schwarz this is at most $\sqrt{\sum_{i=1}^M \phi_i^2(x)}$. Consider the polynomial $\sum_{i=1}^M \phi_i^2(y)$. This is clearly invariant under $SO(d+k)$ (since it is

independent of the choice of basis ϕ_i). Therefore this function is constant. Furthermore its average value on $S^(d+k-1)$ is clearly M. Therefore $f(x) \leq \sqrt{M}$.

On the other hand we have that

$$M = \binom{n+d+k-1}{d+k-1} + \binom{n+d+k-2}{d+k-1} = O\left(\frac{n}{d+k-1}\right)^{d+k-1}.$$

This completes our proof. □

Lemma 30. *Let f be a real-valued polynomial of degree at most n on S^1. Suppose that $f \geq 0$ on S^1. Write $f(\theta) = f(\cos(\theta), \sin(\theta))$. We can write f in terms of a Fourier Series as*

$$f(\theta) = \sum_{k=-n}^{n} a_k e^{ik\theta}.$$

Then a_0 is real and $a_0 \geq |a_k|$ for all k.

Proof. The fact that f can be written in such a way comes from noting that $e^{\pm ik\theta}$ are the spherical harmonics of degree k on S^1. Since f is real valued it follows that $a_{-k} = \bar{a_k}$ for all k. We have that

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)d\theta = \frac{1}{2\pi}\int_0^{2\pi} |f(\theta)e^{-ik\theta}|d\theta \geq \left|\frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-ik\theta}d\theta\right| = |a_k|.$$

□

Lemma 31. *If f is a polynomial of degree at most n that is non-negative on S^1, then $Var_{S^1}(f) = O(n)\int_{S^1} f$.*

Proof. Consider $f = f(\theta)$ as above. For an angle ϕ, let

$$g_\phi(\theta) = f(\phi + \theta) + f(\phi - \theta).$$

Clearly g_ϕ is non-negative, and $\int_{S^1} g_\phi = 2\int_{S^1} f$. Furthermore, we have that

$$\begin{aligned}
\int_0^{2\pi} \mathrm{Var}_{S^1}(g_\phi)d\phi &= \int_0^{2\pi}\int_0^{2\pi} |f'(\phi+\theta) - f'(\phi-\theta)|d\theta d\phi \\
&= \int_0^{2\pi}\int_0^{2\pi} |f'(\vartheta) - f'(\rho)|d\rho d\vartheta \\
&\geq 2\pi\int_0^{2\pi} |f'(\vartheta)|d\vartheta \\
&= 2\pi\mathrm{Var}_{S^1}(f).
\end{aligned}$$

Where above we use the fact that $\int_0^{2\pi} f'(\rho)d\rho = 0$ and that the absolute value function is convex. Hence for some ϕ, $\mathrm{Var}_{S^1}(g_\phi) \geq \mathrm{Var}_{S^1}(f)$. Therefore, we may consider g_ϕ instead of f. Noting that $g_\phi(\theta) = g_\phi(-\theta)$, we find that g_ϕ can be written as $p(\cos\theta)$ for some polynomial p of degree at most n. Our result then follows from Lemma 21. □

Lemma 32. *Let $d \geq 0$ be an integer. Consider the design problem given by $X = [0,1]$, $\mu = r^d dr/(d+1)$, and W the set of polynomials of degree at most n in r^2. Then there exists a weighted design for this problem, (w_i, r_i) where $w_i = \Omega_d(r_i^d \sqrt{1-r_i^2} n^{-1})$, $\min(r_i) = \Omega(n^{-1})$, and $\max(r_i) = 1 - \Omega(n^{-2})$.*

Proof. For any such polynomial $p(r^2)$ we have that

$$\int_0^1 p(r^2) r^d/(d+1) dr = \frac{1}{2(d+1)} \int_0^1 p(s) s^{(d-1)/2} ds.$$

Therefore, if we have a weighted design (w_i, s_i) for the design problem

$$([0,1], \frac{s^{(d-1)/2} ds}{2(d+1)}, \mathcal{P}_n),$$

then $(w_i, \sqrt{s_i})$ will be a weighted design for our original problem. We use the design implied by Lemma 18. The bound on the w_i is implied by Equation 6. The bounds on the endpoints are implied by our observation that there are no roots of $P_n^{((d-1)/2,0)}$ within $O_d(n^{-2})$ of either endpoint. □

We are now ready to prove Proposition 26. We will construct graphs G_n^d that for the design problem $\mathcal{D}_n^d$ have $K_{G_n^d} = O_d(n^d \log(n)^{d-1})$. We will work by induction on d. For $d = 1$, we let $G_n^d = S^1$. This suffices by Lemma 31. From this point on, all of our asymptotic notation will potentially depend on d.

In order to construct these graphs for larger d, we will want to pick a convenient parametrization of the d-sphere. Consider $S^d \subset \mathbb{R}^{d+1}$ as $\{x : |x| = 1\}$. We let r be the coordinate on the sphere $\sqrt{\sum_{i=1}^{d-1} x_i^2}$. We let $u \in S^{d-2}$ be the coordinate so that $(x_1, x_2, \ldots, x_{d-1}) = ru$. We let θ be the coordinate so that $(x_d, x_{d+1}) = \sqrt{1-r^2}(\cos\theta, \sin\theta)$. Note that u is defined except where $r = 0$ and θ is defined except where $r = 1$. Note that in these coordinates, the normalized measure on S^d is given by $\frac{r^{d-2} dr du d\theta}{2\pi(d-1)}$. We also note that if ϕ_i^m are an orthonormal basis for the degree m spherical harmonics on S^{d-2}, that an orthonormal basis for the polynomials of degree at most n on S^d is given by

$$(1-r^2)^{k/2} e^{ik\theta} r^m \phi_i^m(u) P_\ell^{k,m,d}(r^2)$$

Where k, m, ℓ are integers with $m, \ell \geq 0$ and $|k| + m + 2\ell \leq n$ and where the $P_\ell^{k,m,d}(r^2)$ are orthogonal polynomials for the measure

$$r^{m+d-2}(1-r^2)^{k/2} dr/(d-1)$$

on $[0,1]$ and functions in r^2, or, equivalently, $P_\ell^{k,m,d}(s)$ are the orthogonal polynomials for the measure $s^{(m+d-3)/2}(1-s)^{k/2} ds/(2(d-1))$ on $[0,1]$.

We construct G_n^d as follows. This construction should work as stated for $d > 2$, and with only slight modification for $d = 2$ (which we shall deal with after giving the rest of our argument).

Let (w_i, r_i) $(1 \leq i \leq h)$ be the design for the measure $r^{d-2}dr/(d-1)$ on $[0,1]$ for polynomials of degree at most $2n$ in r^2 as described in Lemma 32.

Let $N = An^{d-2}(\log(n))^{d-2}$ for A a sufficiently large constant. For each r_i, let $N_i = [r_i^{d-2}N]$ and $k_i = \left[\frac{Br_i n \log(n)}{\log(nr\log(n))}\right]$, where B is a sufficiently large constant that is still a sufficiently small multiple of A. We inductively construct $G_i = G_{k_i}^{d-2}$. By the inductive hypothesis for the design problem $\mathcal{D}_{k_i}^{d-2}$, $K_{G_i} < (N_i)$ if A was sufficiently large compared to B. Therefore, by Proposition 26 there is a design $u_{i,j}$, $1 \leq j \leq N_i$ for the design problem $\mathcal{D}_{k_i}^{d-2}$ so that each of the $u_{i,j}$ lies on G_i. Let r_1 be the smallest of the r_i. By rotating $G_i, u_{i,j}$ if necessary we can guarantee that $r_i u_{i,1} = (r_1, \sqrt{r_i^2 - r_1^2}, 0, \ldots, 0)$ for all i.

We now define our graph $G = G_n^d$ as follows in (r, u, θ) coordinates. First we define H to be the union of:

- The circles $(r_i, u_{i,j}, \theta)$ for $\theta \in [0, 2\pi]$ for $1 \leq i \leq h$ and $1 \leq j \leq N_i$
- The graphs $(r_i, u, 0)$ for $u \in G_i$ for $1 \leq i \leq h$

We note that H is not connected. Its connected components correspond to the r_i, since each G_i connects all of the circles at the corresponding $u_{i,j}$. We let $G = H \cup H'$, where H' is the image of H under the reflection that swaps the coordinates x_2 and x_d. We note that H union the circle in H' corresponding to $u_{1,1}$ is connected. Since this circle is parameterized as $(r_1, \sqrt{1-r_1^2}\sin\theta, 0, 0, \ldots, 0, \sqrt{1-r_1^2}\cos\theta)$ intersects each of the $u_{i,1}$ in H. Similarly H' union the circle over $u_{1,1}$ in H is connected. Hence G is connected. It is also clear that the total length of all the edges of G is $n^{O(1)}$. We now only need to prove that $K_G = O(n^d \log(n)^{d-1})$. We note that it suffices to prove that $K_H = O(n^d \log(n)^{d-1})$ since $K_G \leq K_H + K_{H'} = 2K_H$.

Let $v_i = \frac{w_i}{N_i}$. We note that $v_i = \Omega(n^{-1}N^{-1}\sqrt{1-r_i^2})$. We claim that the circles in H with weights given by v_i form an approximate design in the following sense.

Lemma 33. *Let C be any real number. Then if B is sufficiently large, and $f \in \mathcal{P}_{4n}$ we have that*

$$\left| \int_{S^d} f - \sum_{i,j} v_i \frac{1}{2\pi} \int_0^{2\pi} f(r_i, u_{i,j}, \theta) d\theta \right| = O(n^{-C})|f|_2. \tag{7}$$

Proof. We note that after increasing C by a constant, it suffices to check our Lemma for f in an orthonormal basis of $\mathcal{P}_{2n}$. Hence we consider

$$f(r, u, \theta) = (1-r^2)^{k/2} e^{ik\theta} r^m \phi^m(u) P_\ell^{k,m,d}(r^2).$$

Note that unless $k = 0$, both of the terms on the left hand side of Equation 7 are 0. Hence we can assume that $k = 0$ and

$$f(r, u, \theta) = f(r, u) = r^m \phi^m(u) P_\ell^{m,d}(r^2).$$

We need to show that

$$\left| \int r^{m+d-2} \phi^m(u) P_\ell^{m,d}(r^2) \frac{dr du}{d-1} - \sum_{i,j} v_i r_i^m P_\ell^{m,d}(r_i^2) \phi^m(u_{i,j}) \right| = O(n^{-C}).$$

First we note that if $m = 0$, $\phi^m(u) = 1$. In this case

$$\begin{aligned}
\sum_{i,j} v_i r_i^m P_\ell^{m,d}(r_i^2) \phi^m(u_{i,j}) &= \sum_i N_i v_i P_\ell^{m,d}(r_i^2) \\
&= \sum_i w_i P_\ell^{m,d}(r_i^2) \\
&= \int_0^1 r^{d-2} P_\ell^{m,d}(r^2) dr/(d-1) \\
&= \int_{S^d} f.
\end{aligned}$$

Where we use above the fact that w_i, r_i is a weighted design. Hence we are done for the case $m = 0$.

For $m > 0$, the integral of f over S^d is 0. Furthermore for $k_i \geq m$, $\sum_j \phi^m(u_{i,j}) = 0$ (since the $u_{i,j}$ are a design). Therefore in this case, the left hand side of Equation 7 is

$$\left| \sum_{k_i < m} v_i r_i^m P_\ell^{m,d}(r_i^2) \sum_j \phi^m(u_{i,j}) \right|.$$

By results in the proof of Lemma 29, we have that $\phi^m(u_{i,j}) = n^{O(1)}$. Furthermore $v_i = O(1)$ and there are $n^{O(1)}$ many pairs of i, j in the sum. Therefore, this is at most

$$n^{O(1)} \max_{i:k_i<m} |r_i^m P_\ell^{m,d}(r_i^2)|.$$

The fact that $|f|_2 = 1$ implies that

$$\begin{aligned}
1 &= \int_0^1 r^{2m+d-2} (P_\ell^{m,d}(r^2))^2 dr/(d-1) \\
&= \int_0^1 s^{m+(d-3)/2} (P_\ell^{m,d}(s))^2 ds/(2(d-1)) \\
&\geq \frac{1}{2^{m+(d+1)/2}(d-1)} \int_{-1}^1 (1-x^2)^{m+(d-3)/2} (P_\ell^{m,d}(2x-1))^2 dx.
\end{aligned}$$

Therefore, since the degree of P is at most n, by Lemma 29 on the 1-disc we have that

$$\max_{[0,1]} P_\ell^{m,d} = O\left(\frac{n}{m}\right)^{m+(d-1)/2} = n^{O(1)} O\left(\frac{n}{m}\right)^m$$

This means that if $m > k_i$ that $r_i^m P_\ell^{m,d}(r_i^2)$ is at most

$$n^{O(1)} O\left(\frac{nr_i}{m}\right)^m.$$

Since for B sufficiently large, $O\left(\frac{nr_i}{k_i}\right)$ would be less than $\frac{1}{2}$, this is at most

$$n^{O(1)} O\left(\frac{nr_i}{k_i}\right)^{k_i}.$$

Hence we need to know that for B sufficiently large,

$$n^{O(1)} O\left(\frac{nr_i}{k_i}\right)^{k_i} = O(n^{-C}). \tag{8}$$

This is because if $nr_i < \log(n)$ the left hand size of Equation 8 is at most

$$n^{O(1)} O(\log(n)^{-1/2})^{\Omega(B\log(n)/\log\log(n))} = n^{O(1)-\Omega(B)}.$$

Where we use the fact that $nr_i = \Omega(1)$. If on the other hand $nr_i \geq \log(n)$, then $k_i = \Omega(B\log(n))$ and the left hand side of Equation 8 is

$$n^{O(1)} O(B^{-1})^{\Omega(B\log(n))} = n^{O(1)-\Omega(B)}.$$

This completes our proof. □

For f a polynomial on S^d let

$$A(f) := \sum_{i,j} v_i \frac{1}{2\pi} \int_0^{2\pi} f(r_i, u_{i,j}, \theta) d\theta.$$

Let

$$A_i(f) := \sum_j v_i \frac{1}{2\pi} \int_0^{2\pi} f(r_i, u_{i,j}, \theta) d\theta.$$

$$A_{i,j}(f) := v_i \frac{1}{2\pi} \int_0^{2\pi} f(r_i, u_{i,j}, \theta) d\theta.$$

Lemma 34. *For $f \in \mathcal{P}_{2n}$, $f \geq 0$ on H,*

$$A(f^2) = n^{O(1)} A(f)^2$$

Proof. Since $f(r_i, u_{i,j}, \theta)$ is a non-negative polynomial of degree at most $2n$ on the circle,

$$\frac{1}{2\pi} f(r_i, u_{i,j}, \theta)^2 d\theta = O(n) \left(\frac{1}{2\pi} f(r_i, u_{i,j}, \theta) d\theta \right)^2.$$

So $A_{i,j}(f^2) = O(n) A_{i,j}(f)^2$.

$$A(f) = \sum_{i,j} v_j A_{i,j}(f)$$

$$A(f^2) = O(n) \sum_{i,j} v_i A_{i,j}(f)^2 \leq O(n) \sum_{i,j} v_i (A(f)/v_i)^2 = n^{O(1)} A(f)^2.$$

Where the last equality holds since $v_i = \Omega(n^{-1} N^{-1} \sqrt{1 - r_i^2})$, and $1 - r_i^2 = \Omega(n^{-2})$ for all i. □

We now prove a more useful version of Lemma 33.

Lemma 35. *If B is sufficiently large, and if f is a polynomial of degree at most $2n$ on S^d that is non-negative on H then*

$$\left| \int_{S^d} f - A(f) \right| \leq \frac{A(f)}{2}.$$

Proof. By Lemma 33 applied to f^2, we have that

$$|f|_2^2 = n^{O(1)} A(f)^2 + O(n^{-C}) |f^2|_2.$$

On the other hand, we have that $\sup_{S^d}(|f|) = n^{O(1)} |f|_2$. Therefore,

$$|f^2|_2^2 \leq |f|_2^2 \sup_{S^d}(|f|)^2 \leq n^{O(1)} |f|_2^4.$$

Hence, we have that

$$|f|_2^2 = n^{O(1)} A(f)^2 + n^{O(1)-C} |f|_2^2.$$

For C sufficiently large, this implies that

$$|f|_2^2 = n^{O(1)} A(f)^2,$$

or that

$$|f|_2 = n^{O(1)} A(f).$$

Therefore, for C sufficiently large, we have that

$$\left| \int_{S^d} f - A(f) \right| \leq O(n^{-C}) |f|_2 \leq n^{O(1)-C} A(f) \leq \frac{A(f)}{2}.$$

□

Corollary 36. *Assuming B is sufficiently large, if f is a polynomial of degree at most $2n$ on S^d and f is non-negative on H then*

$$A(f) \leq 2 \int^{S^d} f.$$

We will now try to bound K_H based on a variant of one of our existing criteria. In particular, we would like to show that if f is a degree n polynomial with $\int f = 1$ and $f \geq 0$ on H that $\mathrm{Var}_G(f) = O(n^d \log(n)^{d-1})$. Replacing f by $\frac{f+1}{A(f+1)}$ and noting by Corollary 36 that $A(f+1) \leq 4$, we can assume instead that $f \geq \frac{1}{4}$ on H and that $A(f) = 1$.

We first bound the variation of f on the circles over $u_{i,j}$. Define

$$f_{i,j}(\theta) := f(r_i, u_{i,j}, \theta).$$

We will prove the following:

Proposition 37. *Let B be sufficiently large. Let f be a degree n polynomials with $f \geq 1/4$ on H and $A(f) = 1$. Then*

$$Var_{S^1} f_{i,j} = O(n^d \log(n)^{d-1}) A_{i,j}(f).$$

This would follow immediately if $f_{i,j}$ was degree at most $n \log(n)\sqrt{1-r^2}$. We will show that the contribution from higher degree harmonics is negligible.

We define for integers k $a_k(r,u)$ to be the $e^{ik\theta}$ component of f at (r,u,θ). We note that $a_k(r,u) = (1-r^2)^{|k|/2} P_k(\vec{r})$, where $\vec{r} = ru$ is a coordinate on the $(d-1)$-disc.

We first show that $|a_k(r,u)|$ is small for $k > n\log(n)\sqrt{1-r^2}$.

Lemma 38. *Let C be a real number and B sufficiently large compared to C. Let f be a degree n polynomial with $f \geq 0$ on H and $A(f) = 1$. Then for $|k| > n\log(n)\sqrt{1-r_i^2}$, $|a_k(r_i,u)| = O(n^{-C})$.*

Proof. We have that $|a_k|_2 \leq |f|_2 = n^{O(1)}$ by Lemma 34. Therefore,

$$\int_{D^{d-1}} (1-r^2)^{|k|} P_k^2(\vec{r}) dr = n^{O(1)}.$$

Applying Lemma 29, we find that

$$|P_k(\vec{r})| \leq n^{O(1)} O\left(\frac{n}{|k|}\right)^{|k|/2}.$$

Therefore,

$$|a_k(r_i,u)| \leq n^{O(1)} O\left(\frac{n\sqrt{1-r_i^2}}{|k|}\right)^{|k|/2} \leq n^{O(1)} O\left(\frac{1}{\log(n)}\right)^{|k|/2}.$$

Since $|k| = \Omega(\log(n))$ (because $\sqrt{1-r_i^2} = \Omega(n^{-1})$), this is $O(n^{-C})$. □

Proof of Proposition 37. Let $f_{i,j}^l(\theta)$ be the component of $f_{i,j}(\theta)$ coming from Fourier coefficients of absolute value at most $n\log(n)\sqrt{1-r_i^2}$. By Lemmas 28 and 38, we have that for B sufficiently large, $f_{i,j} - f_{i,j}^l$ is less than $1/8$ everywhere and has Variation $O(1)$. But since $f_{i,j}^l$ is non-negative and has bounded Fourier coefficients, we have by Lemma 31 that

$$\mathrm{Var}_{S^1} f_{i,j}^l = O\left(n\log(n)\sqrt{1-r_i^2}\right)\int_{S^1} f_{i,j}^l = O\left(n\log(n)\sqrt{1-r_i^2}\right)\int_{S^1} f_{i,j}.$$

This means that

$$\begin{aligned}
\mathrm{Var}_{S^1}(f_{i,j}) &= O\left(\frac{n\log(n)\sqrt{1-r_i^2}}{v_i}\right) A_{i,j}(f) \\
&= O\left(\frac{N_i n\log(n)\sqrt{1-r_i^2}}{w_i}\right) A_{i,j}(f) \\
&= O\left(\frac{(r_i n\log(n))^{d-2} n\log(n)\sqrt{1-r_i^2} n}{\sqrt{1-r_i^2}}\right) A_{i,j}(f) \\
&= O\left(n^d \log(n)^{d-1}\right) A_{i,j}(f).
\end{aligned}$$

□

We now bound the variation of f on the G_i in H.

Proposition 39. *Suppose that B is sufficiently large. For $f \in \mathcal{P}_n$, $f \geq \frac{1}{4}$ on H, $A(f) = 1$, $Var_{G_i}(f) \leq A_i(f) O(n^d \log(n)^{d-1})$.*

Again this would be easy if we knew that the restriction of f to the appropriate sphere was low degree. Our proof will show that the contribution from higher degree harmonics is small.

Let $f_i(u) = f(r_i, u, 0)$ be f restricted to the $(d-2)$-sphere on which G_i lies. We claim that the contribution to f from harmonics of degree more than k_i is small. In particular we show that:

Lemma 40. *Let C be a real number. Suppose that B is sufficiently large. For $f \in \mathcal{P}_n$, $f \geq 0$ on H, $A(f) = 1$. Let $f_i^h(u)$ be the component of f_i coming from spherical harmonics of degree more than k_i. Then $|f_i^h|_2 = O(n^{-C})$.*

Proof. Perhaps increasing C by a constant, it suffices to show that for ϕ a spherical harmonic of degree $m > k_i$ that the component of ϕ in f_i is $O(n^{-C})$. We will want to use slightly different coordinates on S^d than usual here. Let $s = (x_d, x_{d+1})$ be a coordinate with values lying in the 2-disc. The component of f corresponding to the harmonic $\phi(u)$ is given by

$$\phi(u)(1-s^2)^{m/2} Q(s)$$

for Q some polynomial of degree at most n. Considering the L^2 norm of f, we find that

$$\int_{D^2}(1-s^2)^m Q^2(s)ds \leq \pi|f|_2^2 \leq n^{O(1)}A(f)^2 = n^{O(1)}.$$

Applying Lemma 29 to $Q(s)$, we find that $|Q(s)| = n^{O(1)}O\left(\frac{n}{m}\right)^{m/2}$. Hence the component of ϕ at r_i is $r_i^{m/2}Q(r_i,0)$, which is at most

$$n^{O(1)}O\left(\frac{nr_i}{m}\right)^{m/2}.$$

Since $m > k_i$, $\frac{nr_i}{m} < \frac{1}{2}$, the above is at most

$$n^{O(1)}O\left(\frac{nr_i}{k_i}\right)^{k_i/2} = O(n^{-C})$$

by Equation 8. □

We can now prove Proposition 39.

Proof of Proposition 39. Let $f_i^l(u)$ be the component of $f_i(u)$ coming from spherical harmonics of degree at most k_i. By Lemmas 40 and 28, we have that for B sufficiently large, $f_i^l \geq 0$ on G_i and that $|\mathrm{Var}_{G_i}(f) - \mathrm{Var}_{G_i}(f_i^l)| \leq v_i/4 \leq A_i(f)$. Hence it suffices to prove that $\mathrm{Var}_{G_i}(f_i^l) = A_i(f)O(n^d\log(n)^{d-1})$. Since for polynomials of degree at most k_i on S^{d-2}, $K_{G_i} = O(k_i^{d-2}\log(k_i)^{d-2})$, we have that $\mathrm{Var}_{G_i}(f_i^l) = O(k_i^{d-2}\log(k_i)^{d-2})\int_{S^{d-2}} f_i^l$. Since the $u_{i,j}$ form a spherical design this is

$$O(k_i^{d-2}\log(k_i)^{d-2})\frac{1}{N_i}\sum_j f_i^l(u_{i,j}).$$

Again, for B sufficiently large, this is

$$O(k_i^{d-2}\log(k_i)^{d-2})\frac{1}{N_i}\sum_j f(r_i,u_{i,j},0).$$

Now consider $F(\theta) = \frac{1}{N_i}\sum_j f(r_i,u_{i,j},\theta)$. We have that F is a polynomial of degree at most n and that $F \geq 1/4$. Let F^l be the component of F consisting of Fourier coefficients with $|k| \leq n\log(n)\sqrt{1-r_i^2}$. By Lemmas 28 and 38, if B is sufficiently large, $|F-F^l| < 1/8$. It is clear that

$$A_i(f) = w_i\frac{1}{2\pi}\int_0^{2\pi}F(\theta)d\theta = \Theta(w_i)\frac{1}{2\pi}\int_0^{2\pi}F^l(\theta)d\theta.$$

Note that by Lemma 31

$$\begin{aligned}
F(0) &= O(1) + F^l(0) \\
&\leq \inf_{S^1}(F^l) + \mathrm{Var}_{S^1}(F^l - \inf_{S^1}(F^l)) \\
&= O(n\log(n)\sqrt{1-r_i^2})\int_{S^1} F^l \\
&= O(n\log(n)\sqrt{1-r_i^2})\int_{S^1} F.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
\mathrm{Var}_{G_i}(f) &= O(k_i^{d-2}\log(k_i)^{d-2})F(0) \\
&= O(n\log(n)\sqrt{1-r_i^2}k_i^{d-2}\log(k_i)^{d-2})A(f)/w_i \\
&= A(f)O\left(\frac{n\log(n)\sqrt{1-r_i^2}k_i^{d-2}\log(k_i)^{d-2}}{w_i}\right) \\
&= A(f)O\left(\frac{n\log(n)\sqrt{1-r_i^2}(r_i n\log(n))^{d-2}}{r_i^{d-2}n^{-1}\sqrt{1-r_i^2}}\right) \\
&= A(f)O(n^d\log(n)^{d-1}).
\end{aligned}$$

□

We can finally prove Proposition 27.

Proof. We proceed by induction on d. For $d = 1$ the S^1 suffices as discussed. Assuming that we have the graph for $d-2$ we construct G as described above. Clearly G is connected and has total length $n^{O(1)}$. We need to show that $K_H = O(n^d\log(n)^{d-1})$. To do so it suffices to show that for any $f \in \mathcal{P}_n$ with $f \geq 1/4$ on H and $A(f) = 1$ that $\mathrm{Var}_H(f) = O(n^d\log(n)^{d-1})$. We have that

$$\begin{aligned}
\mathrm{Var}_H(f) &= \sum_{i,j}\mathrm{Var}_{S^1}(f_{i,j}) + \sum_i \mathrm{Var}_{G_i} f_i \\
&= O(n^d\log(n)^{d-1})\left(\sum_{i,j}A_{i,j}(f) + \sum_i A_i(f)\right) \\
&= O(n^d\log(n)^{d-1})(A(f) + A(f)) \\
&= O(n^d\log(n)^{d-1}).
\end{aligned}$$

This completes the proof.

For $d = 2$, there is obviously no connected graph on the S^{d-2}-sphere, so we need to modify our construction slightly. Fortunately, there are strength n designs on S^{d-2} of size $O(n^{d-2})$. Hence we can still define $w_i, r_i, u_{i,j}$ as before. We now ignore the G_i from our construction. Notice that all of our assumptions that f was non-negative on H except for those used in Proposition 39 could have been replaced by the assumption that f was non-negative on the $u_{i,j}$. Furthermore, for each i there are only two j, both of whose circles can be made to intersect the $u_{1,1}$ circle from H', and so we no longer need to the G_i to maintain connectivity of G. □

Theorem 25 now follows from Proposition 27 and Proposition 26.

Part III

Polynomial Threshold Functions

1 Introduction

This Part consists of three Chapters on polynomial threshold functions (abbreviated PTFs). We call a function, f, a polynomial threshold function if it is of the form $f(x) = \text{sgn}(p(x))$ for some polynomial p. We say that f is a degree-d polynomial threshold function if the degree of p is at most d. Polynomial threshold functions are a natural generalization of linear threshold functions (the $d = 1$ case). Polynomial threshold functions find application in fields such as circuit complexity [3], communication complexity [56], and learning theory [36].

Much of the study of polynomial threshold functions involves studying their average behavior over some probability distribution, or in finding ways to approximate this average behavior. There are two natural probability distributions that show up in this context. The first is that of the hypercube or Bernoulli distribution, namely the uniform distribution over $\{0,1\}^n$. The second is the Gaussian distribution. The Bernoulli distribution is perhaps more useful from a computer science context, though the Gaussian distribution is arguably more natural and is almost certainly easier to work with. In fact the Gaussian distribution can often be thought of as a special case of the Bernoulli distribution, as the central limit theorem tells us that one can approximate a Gaussian as an average of a large number of Bernoulli variables.

The Gaussian distribution is usually easier to work with for several reasons. For one thing, the Gaussian distribution is continuous. Importantly this means that polynomials of Gaussians are never too likely to have values concentrated in some small range. In particular, for Gaussians we have the following anticoncentration result:

Theorem (Carbery and Wright)**.** *Let p be a degree d polynomial, and Y a standard Gaussian. Suppose that $E[p(Y)^2] = 1$. Then, for $\epsilon > 0$,*

$$Pr(|p(Y)| < \epsilon) = O(d\epsilon^{1/d}).$$

This clearly does not hold for the Bernoulli distribution, although there are recent techniques to work around it. In particular the invariance principle of [47] states that for low-influence polynomials (i.e. those for which no one coordinate has significant impact on the final value) that Bernoulli and Gaussian inputs behave similarly. For polynomials of high influence, there are a number of regularity Lemmas (for example the one in [15] or [**?**]) that after fixing several of the influential variables, allow one to approximate a polynomial threshold function by a regular one.

The Gaussian setting also has much greater symmetry than the Bernoulli case. We will talk more about how to take advantage of this shortly. For these reasons, all of the work done in this Part is for the Gaussian case. Perhaps another project would be to translate some of these results into the Bernoulli case.

One advantage of the Gaussian setting is that there are convenient notions of perturbing the input value. In the Bernoulli setting, there are the operations of negating a random coordinate, or negating each coordinate with probability ϵ, but neither of these operations is terribly well behaved. In the Gaussian setting, there is the possibility of taking actual derivatives. Additionally, one can replace a random Gaussian X with another random Gaussian Z having high correlation with X in each coordinate. In doing so, Z can be thought of as a noisy version of X. One way of producing such Z is to let Y be an independent Gaussian and to let

$$Z = (1 - \epsilon)X + \sqrt{2\epsilon - \epsilon^2}Y$$

for some parameter ϵ. Note that the coefficients are chosen so that Z is properly normalized. This is very important since $p(X)$ does not necessarily behave very much like $p(X + \epsilon Y)$ as the latter will likely have a noticeably larger norm. This issue shows up in Corollary 7 of Chapter G, where it should be noted that the difference of normalization leads to a dependence on the dimension of the underlying space as well as ϵ. Given this normalization, a more suggestive way to write the above equation would be

$$Z = \cos(\theta)X + \sin(\theta)Y$$

for an appropriate value of θ. This formula suggests that Z is somehow a rotation of X, and this idea provides a powerful technique for proving things about the expected relationship between Z and X. In particular, define $X_\phi := \cos(\phi)X + \sin(\phi)Y$ so that $X = X_0$ and $Z = X_\theta$. Suppose that we wish to bound some probability:

$$\Pr_{X,Y}(\langle \text{Something involving } X \text{ and } Z \rangle).$$

It should be noted that this probability will be the same if X and Y are replaced by any two independent Gaussians. In particular, we can replace them with X_ϕ and $X_{\phi+\pi/2}$. This makes our new X and Z equal to X_ϕ and $X_{\phi+\theta}$. Hence we can instead compute:

$$\Pr_{X,Y}(\langle \text{Something involving } X_\phi \text{ and } X_{\phi+\theta} \rangle).$$

(Note that X_ϕ and $X_{\phi+\theta}$ depend on X and Y). The trick here is to fix the values of X and Y and to average the above over ϕ. If you can obtain a bound that holds for any initial X and Y, then this will hold as a bound on the original probability. This technique proves to be the key insight used in Chapter G as well as a key component of the proof in Chapter H. In particular, to see examples of this technique we direct the reader to the proofs of Theorem 1 in Chapter G and of Lemma 9 in Chapter H.

When considering the use of this trick, I like to consider the XY-plane whose points consist of random variables which are linear combinations of X and Y. X and Z are unit vectors in this plane separated by an angle of θ, and the X_ϕ trace out the unit circle in this plane. Once we fix the values of X and Y, we can restrict a degree-d polynomial of $\mathbb{R}^n$ to

a degree-d polynomial on the XY-plane. So our question about some polynomial applied to X and Z becomes a question about a polynomial evaluated at two points of this plane. Typically just looking at these points is not very useful, but the trick above allows us to average over rotations of the XY-plane. This often reduces our problem to proving general statements about polynomials defined on a circle.

Another idea that I have found useful in this work is something I like to call strong anticoncentration. Although the anticoncentration result of Carbery and Wright above is very useful in a number of contexts, it has an unfortunate dependence on the degree of the polynomial. In some sense, this dependence is a necessary evil, since the polynomial X^d is in fact somewhat concentrated around 0. On the other hand, we can still say something useful about when such a polynomial is small. In particular $p(X) = X^d$ is small only when its derivative is also nearly as small. In particular, I use the heuristic that for any polynomial p, it should be the case that $|p(X)| < \epsilon|p'(X)|$ with probability not much bigger than ϵ. This should hold because changing the value of X by ϵ should adjust the value of $p(X)$ by roughly $\epsilon|p'(X)|$. This idea allows me to state a strong version of anticoncentration. In particular, with probability roughly $1 - \epsilon$ it should hold that

$$|p(X)| \geq \epsilon|p'(X)| \geq \epsilon^2|p''(X)| \geq \ldots \geq \epsilon^d|p^{(d)}(X)|.$$

So $|p(X)| \geq \epsilon^d|p^{(d)}(X)|$. It should be noted that $|p^{(d)}(X)|$ is independent of X and can be thought of as a rough approximation to $|p|_2$. This idea is one of the key technical tools hidden in the proof of Lemma 9 of Chapter G, and a rigorous formulation of the above appears as Corollary 10 of Chapter H.

Another topic that will be discussed in Chapters F and H is that of k-independent families of random variables. We say that a collection of random variables $X_1, \ldots, X_n$ is k-independent if any k of the X_i are independent of each other. k-independence is frequently useful in that k-independent families can be constructed to have much smaller entropy than the corresponding fully-independent family of random variables would have. In particular, this allows us to generate a k-independent family from a small seed in a computationally efficient manner. Additionally, k-independent families of random variables will often share many of the important properties of the fully independent family making them an important tool.

Chapter F

k-Independence Fools Polynomial Threshold Functions

1 Introduction

In this Chapter, we study the problem of fooling polynomial threshold functions by means of limited independence.

We say that a random variables X fools a family of functions with respect to some distribution Y if for every function, f, in the family

$$|\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]| = O(\epsilon).$$

In this paper we will be interested in the case where the family is of all degree-d polynomial threshold functions in n-variables, and Y is either an n-dimension Gaussian distribution, and in particular the case where X is an arbitrary family of k-independent Gaussian random variables. In particular, we prove that

Theorem 1. *Let $d > 0$ be an integer and $\epsilon > 0$ a real number, then there exists a $k = O_d\left(\epsilon^{-2^{O(d)}}\right)$, so that for any degree d polynomial p and any k-independent family of Gaussians X and fully independent family of Gaussians Y*

$$|E[sgn(p(X))] - E[sgn(p(Y))]| = O(\epsilon).$$

There has been a significant amount of recent work on the problem of fooling low degree polynomial threshold functions of Gaussian or Bernoulli random variables, especially via limited independence. It was shown in [11] that $\tilde{O}(\epsilon^{-2})$-independence is sufficient to fool degree-1 polynomial threshold functions of Bernoulli random variables, and show that this is tight up to polylogarithmic factors. In [13] it was shown that $\tilde{O}(\epsilon^{-9})$-independence sufficed for degree-2 polynomial threshold functions of Bernoullis and that $O(\epsilon^{-2})$ and $O(\epsilon^{-8})$ suffices

for degree 1 and 2 polynomial threshold functions of Gaussians. The degree 1 case was also extended by [4], who show that limited independence fools threshold functions of polynomials that can be written in terms of a small number of linear polynomials. Finally, in [45] a more complicated pseudorandom generator for degree-d polynomial threshold functions of Bernoulli variables is developed with seed length $2^{O(d)}\log(n)\epsilon^{-8d-3}$. As far as we are aware, our result is the first result to show that degree-d polynomial threshold functions are fooled by k-independence for any k depending only on ϵ and d for any $d \geq 3$.

2 Overview

We prove Theorem 1 first by proving our result for multilinear polynomials, and then finding a reduction to the general case. In particular we prove

Proposition 2. *Let $d > 0$ be an integer and $\epsilon > 0$ a real number, then there exists a $k = O_d\left(\epsilon^{-2^{O(d)}}\right)$, so that for any degree d multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$ and any k-independent family of Gaussians X and fully independent family of Gaussians Y*

$$|E[sgn(p(X))] - E[sgn(p(Y))]| = O(\epsilon).$$

We define the notation $A \approx_\epsilon B$ to mean $|A - B| = O(\epsilon)$.

This result is proven using the FT-Mollification method which I have been developing along with Jelani Nelson and Ilias Diakonikolas. This technique first appeared in [33] as a way to show that k-independence fooled linear threshold functions of sums of p-stable random variables. FT-Mollification is a general method for showing that k-independence fools various functions. The idea is as follows. Suppose that we are trying to show that for some function f, and a family of independent random variables Y that for any k-independent family, X, with the same individual distributions as Y we have that

$$\mathrm{E}[f(X)] \approx_\epsilon \mathrm{E}[f(Y)].$$

If f were a polynomial of degree at most k, our work would be done (and in fact the expectations would be the same). In the general case, this problem is closely related to that of approximation f by such a polynomial. If f happens to be a smooth function, we can often do this by approximating f by one of its Taylor polynomials. Furthermore, the Taylor error is bounded above by a polynomial, allowing us to deal with it conveniently. If f is not smooth, the idea is to approximate it by a smooth $\tilde{f}$. The basic strategy is then to show that

$$\mathrm{E}[f(Y)] \approx_\epsilon \mathrm{E}[\tilde{f}(Y)] \approx_\epsilon \mathrm{E}[\tilde{f}(X)] \approx_\epsilon \mathrm{E}[f(X)].$$

The middle equality will be proved by approximation $\tilde{f}$ by one of its Taylor polynomials. The first approximation will hold because $\tilde{f}$ is a good approximation of f. This will typically

depend on proving some anticoncentration result. If f is discontinuous, there will necessarily be a large error between f and $\tilde{f}$ near the discontinuity, and we instead must show that Y is not near this discontinuity with much probability. The final approximation follows from similar considerations, only with the added twist that we must now show anticoncentration for k-independent families.

The method used for obtaining $\tilde{f}$ is an important part of this technique. We need to approximate f by a smooth function $\tilde{f}$. Furthermore, in order to keep the Taylor error is small, we need $\tilde{f}$ to have sharp bounds on its higher derivatives. To produce an approximation, we use the standard method of mollification. Namely, we let $\tilde{f}$ be a convolution of f with some smooth function ρ. To get our derivative bounds, we need that ρ have sharp bounds on its derivatives. To obtain this, we use the Fourier transform of a compactly supported function.

This technique seems to work best when f can be written as some relatively simple function of a small number of polynomials of X. Convolving in a large number of dimensions tends to introduce error, so instead if we can write $f = g(P(X))$, it is useful to produce $\tilde{f}$ as $\tilde{g}(P(X))$. In its original incarnation in [33], g was a function of one variable. Later when FT-Mollification was used in [13], g was a function of four variables. In this Chapter, g is a function of some large number of variables. There is still ongoing work with this technique to figure out how to effectively deal with multidimensional FT-Mollification.

The idea of smoothing a function by convolving it with the Fourier transform of a compactly supported function is not new. It is used for example in several proofs of Jackson's Theorem (for example as in [41]). As far as I know, the idea of taking a Taylor polynomial of the resulting function to get a polynomial that approximates f for x not near a discontinuity and with $|x|$ small is new, as is the application to questions of k-independence.

The structure of this Chapter is as follows.

In Section 3 we prove bounds on the moments of multilinear Gaussian polynomials. These results are essentially a reworking of the main result of [39].

In Section 4, we use these bounds to prove a structure Theorem for multilinear polynomials. In particular, we prove that we can write $p(X)$ in the form $h(P_1(X), P_2(X), \ldots, P_N(X))$ where h is a polynomial and $P_i(X)$ are multilinear polynomials with relatively small higher moments. More specifically, the polynomials P_i will be split into d different classes, with the i^{th} class consisting of n_i polynomials each of whose m_i^{th} moments are $O_d(m_i)^{m_i/2}$. This decomposition allows us to write $f(X) = \mathrm{sgn}(P(X))$ as $\mathrm{sgn}(h(P_1(X), \ldots, P_N(X)))$.

From here we make use of the FT-Mollification method. We deal with the error coming from Taylor approximation in Section 6, and the errors from anticoncentration in Section 7.

Our application of FT-Mollification is complicated by the fact that our moment bounds on the P_j are not uniform in j. To deal with this, we will construct $\tilde{h}$ to have different degrees of smoothness in different directions, and the parameter C_i will describe the amount of smoothness along the i^{th} set of coordinates (corresponding the the i^{th} class of the P_j). This forces us to come up with modified techniques for producing $\tilde{h}$ and dealing with the Taylor polynomial and Taylor error.

In Section 8, we reduce the general case to the case of multilinear polynomials by approximating $p(X)$ by a multilinear polynomial in some larger number of variables.

Finally, in Section 9, we discuss the actual requirements for k and the possibility of extended our results to the Bernoulli setting.

3 Moment Bounds

In this Section, we prove a bound on the moments of arbitrary degree-d multilinear polynomials of Gaussians. Our bound is essentially a version of the main Theorem of [39] readapted to work in a slightly more general context. It should be noted that this result is the only reason that we restrict ourselves for most of this Chapter to the case of multilinear polynomials, as it will make our bound easier to state and work with.

The bound, which we shall state shortly, will tell us that a polynomial, p, will have small higher moments (growing like $k^{k/2}$) unless it has some large correlation with a product of smaller degree polynomials.

Throughout this Section, we will refer to two slightly different notions that of a multilinear polynomial and that of a multilinear form. For our purposes, a multilinear polynomial $p(X)$ (X has n coordinates) will be a polynomial so that the degree of p with respect to any of the coordinates of X is at most 1. A multilinear form will be a polynomial $q(X^1, X^2, \ldots, X^m)$ (here each of the X^i may themselves have several coordinates) so that q is linear (homogeneous degree 1) in each of the X^i. We call such a q symmetric if it is symmetric with respect to interchanging the X^i. Finally, we note that to every homogeneous multilinear polynomial p of degree d, there is an associated multilinear form $q(X^1, \ldots, X^d)$, which is the unique symmetric multilinear form so that $p(X) = q(X, \ldots, X)$.

Before we can state our results we need a few more definitions.

Definition. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a homogeneous degree-d multilinear polynomial. Let $X_i, 1 \leq i \leq n$ be independent standard Gaussians. For integers $1 \leq \ell \leq d$ define $M_\ell(p)$ in the following way. Consider all possible choices of: a partition of $\{1, \ldots, n\}$ into sets $S_1, S_2, \ldots, S_\ell$; a sequence of integers $d_i \geq 1, 1 \leq i \leq \ell$ so that $d = \sum_{i=1}^{\ell} d_i$; a sequence of multilinear polynomials $p_i, 1 \leq i \leq \ell$ so that p_i depends only on the coordinates in S_i, p_i is homogeneous of degree d_i, and $E[p_i(X)^2] = 1$. We let $M_\ell(p)$ be the supremum over all choices of S_i, d_i, p_i as above of*

$$\left(E\left[p(X) \prod_{i=1}^{\ell} p_i(X) \right] \right)^{1/2}.$$

Note that by Cauchy-Schwarz we have that $M_\ell(p) \leq \mathrm{E}[p(X)^2]^{1/2}$. M_ℓ can be thought of as a measure of the largest possible correlation that p can have with a product of ℓ polynomials

of smaller degree and total degree d. We now define a similar quantity more closely related to what is used in [39].

Definition. *Let $q : (\mathbb{R}^n)^d \to \mathbb{R}$ by a degree-d multilinear form. Let $X^i, 1 \leq i \leq d$ be independent standard n-dimensional Gaussians. For integers $1 \leq \ell \leq d$ define $M_\ell(q)$ in the following way. Consider all possible choices of: a partition of $\{1, \ldots, d\}$ into non-empty subsets $S_1, \ldots, S_\ell$, with $S_i = \{c_{i,1}, \ldots, c_{i,d_i}\}$; and a set of multilinear forms q_i of degree-d_i with*

$$E[q_i(X^{c_{i,1}}, \ldots, X^{c_{i,d_i}})^2] \leq 1.$$

We define $M_\ell(q)$ to be the supremum over all such choices of S_i and q_i of

$$\left(E\left[q(X^1, \ldots, X^d) \prod_{i=1}^{\ell} q_i(X^{c_{i,1}}, \ldots, X^{c_{i,d_i}}) \right] \right)^{1/2}.$$

We now state the moment bound whose proof will take up the rest of this Section.

Proposition 3. *Let p be a homogenous, degree-d, multilinear polynomial, and X a family of independent standard Gaussians, and $k \geq 2$. Then*

$$E[|p(X)|^k] = \Theta_d \left(\sum_{\ell=1}^{d} M_\ell(p) k^{\ell/2} \right)^k.$$

This is essentially a version of Theorem 1 of [39] for multilinear polynomials rather than multilinear forms. We state said Theorem below:

Theorem ([39] Theorem 1)**.** *For q a degree-d, multilinear form and X^i independent standard n-dimensional Gaussians and k an integer at least 2,*

$$E[|q(X^1, \ldots, X^d)|^k] = \Theta_d \left(\sum_{\ell=1}^{d} M_\ell(q) k^{\ell/2} \right)^k.$$

Proof of Proposition 3. The basic idea of the proof is the relate $M_\ell(p)$ to $M_\ell(q)$ and $\mathrm{E}[|p|^k]$ to $\mathrm{E}[|q|^k]$ for q the symmetric multilinear form associated to a multilinear polynomial p.

Let q be the associated symmetric multilinear form associated to p. We claim that for each ℓ that $M_\ell(p) = \Theta_d(M_\ell(q))$. Suppose that p_1 and p_2 are degree d multilinear polynomials, and q_1 and q_2 the associated symmetric multilinear forms. It is easy to see (by using the standard basis of coefficients) that $\mathrm{E}[p_1(X)p_2(X)] = d!\mathrm{E}[q_1(X^1, \ldots, X^d)q_2(X^1, \ldots, X^d)]$. Similarly, it is easy to see that if p is a homogeneous, degree-d, multilinear polynomial, and p_i are

homogeneous, degree-d_i, multilinear polynomials on distinct sets of coordinates, and q, q_i their associated symmetric multilinear forms we have

$$\begin{aligned}&\mathrm{E}\left[p(X)\prod p_i(X)\right]\\&\quad=\frac{1}{\prod d_i!}\mathrm{E}\left[q(X^1,\ldots,X^d)\prod q_i(X^{d_1+\ldots+d_{i-1}+1},\ldots,X^{d_1+\ldots+d_{i-1}+d_i})\right].\end{aligned}$$

This means that $M_\ell(p) = O_d(M_\ell(q))$ since given the appropriate S_i, d_i, p_i we can use the symmetrizations of the p_i to get as good a bound for $M_\ell(q)$ up to a constant factor. To show the other direction we need to show that $M_\ell(q)$ is not changed by more than a constant factor if we require that the q_i are supported on disjoint sets of coordinates. But we note that if you randomly assign each coordinate to a q_i and take the part that only depends on those coordinates, you lose a factor of at most d^d on average.

Hence we have that

$$\mathrm{E}[|q(X^1,\ldots,X^d)|^k] = \Theta_d\left(\sum_{\ell=1}^{d} M_\ell(p)k^{\ell/2}\right)^k.$$

We just need to show that the moments of p and the moments of q are the same up to a factor of $\Theta_d(1)^k$. This can be shown using the main Theorem of [8], which in our case states that there is some constant C_d, depending only on d, so that for any such p, q and x,

$$\Pr(|p(X)| > x) \leq C_d\Pr(|q(X^1,\ldots,X^d)| > x/C_d)$$

and

$$\Pr(|q(X^1,\ldots,X^d)| > x) \leq C_d\Pr(|p(X)| > x/C_d).$$

Our result follows from noting that for any random variable Y that

$$\mathrm{E}[|Y|^k] = \int_0^\infty kx^{k-1}\Pr(|Y| > x)dx.$$

□

4 Structure

In this Section, we prove the following structure theorem for degree-d multilinear polynomials.

Proposition 4. *Let p be a degree-d multilinear polynomial where the sum of the squares of its coefficients is at most 1. Let $m_1 \leq m_2 \leq \ldots \leq m_d$ be integers. Then there exist integers $n_1, n_2, \ldots, n_d$, $n_i = O_d(m_1m_2\cdots m_{i-1})$ and multilinear polynomials $h_1, \ldots, h_d$, $P_{i,j}, 1 \leq i \leq d, 1 \leq j \leq n_i$ so that:*

1. $p(Y) = \sum_{i=1}^{d} h_i(P_{i,1}(Y), P_{i,2}(Y), \ldots, P_{i,n_i}(Y))$.
2. *$P_{i,j}$ is homogeneous*
3. *h_i is homogeneous degree i*
4. *If $P_{i,a_1} \cdots P_{i,a_i}$ appears as a term in $h_i(P_{i,j})$, then the sum of the degrees of the P_{i,a_i} is at most d*
5. *The sum of the squares of the coefficients of h_i is $O_d(1)$*
6. *The sum of the squares of the coefficients of $P_{i,j}$ is at most 1*
7. *Each variable occurs in at most one monomial in h_i*
8. *If Y is a standard Gaussian and $k \leq m_i$, then $E[|P_{i,j}(Y)|^k] = O_d(\sqrt{k})^k$.*

This will allow us to write p in terms of other polynomials each with smaller moments. In particular, we already know just from the fact that p has degree at most d that its k^{th} moment is at most $O(k)^{kd/2}$. The above proposition allows us to write p in terms of $P_{i,j}$ which have moments on the order of $O(k)^{k/2}$, at least for k which are not too large.

The basic idea of the proof follows from a proper interpretation of Proposition 3. Essentially Proposition 3 says that the higher moments of p will be small unless p has some significant component consisting of a product of polynomials $P_1, \ldots, P_\ell$ of lower degree. The basic idea is that if such polynomials exist, we can split off these P_i as new polynomials in our decomposition, leaving $p - P_1 \cdots P_\ell$ with smaller size than p. We repeatedly apply this procedure to p and all of the other polynomials that show up in our decomposition. Since each step decreases the size of the polynomial being decomposed, and produces only new polynomials of smaller degree, this process will eventually terminate. Beyond these ideas, the proof consists largely of bookkeeping to ensure that we have the correct number of P's and that they have an appropriate number of small moments.

Proof. We define a dot product on the space of multilinear polynomials $\langle P, Q \rangle = \mathrm{E}[P(Y)Q(Y)]$ where Y is a standard Gaussian. Note that the square of the corresponding norm is just $|P|^2$, which equals the sum of the squares of the coefficients of P.

We will prove our Proposition by coming up with a sequence of such decompositions of P as $\sum h_i(P_{i,j})$ satisfying the necessary conditions except for having a slightly relaxed version of the moment bounds on the $P_{i,j}$. In particular, we show by induction on s there exists a decomposition satisfying properties 1-7 above and satisfying 8 so long as $i \leq s$.

The base case of $s = 0$ is can be trivially proven by setting $P_{1,1} = P$, and $h_1 = P_{1,1}$ and letting all of the other relevant polynomials be 0. We now consider the inductive step. We begin with a decomposition of P into polynomials $P_{i,j}$ so that all of the $P_{i,j}$ with $i < s$ have moments that are sufficiently small. If the moments of the $P_{s,j}$ are sufficiently small, then

we are done. If this is not the case, we describe a procedure by which our decomposition is replaced by a sequence of new decompositions, eventually terminating in one in which the moments of the $P_{s,j}$ are not too large. Our procedure will consist of repeatedly applying the following operation:

Suppose that for some $j \leq n_i$ and $k \leq m_s$ that $\mathrm{E}[|P_{s,j}(X)|^k] > \Omega_d(\sqrt{k})^k$. By Proposition 3 this implies that for some $1 \leq \ell \leq d+1-s$ that $k^{k\ell/2} M_\ell > k^{1/2}$, or that $M_\ell > m_s^{(1-\ell)/2}$. Note that this implies that $\ell \geq 2$. By the definition of M_ℓ, this implies that there are homogeneous, multilinear, polynomials $Q_1, \ldots, Q_\ell$ of norm 1, so that $\sum_{i=1}^{\ell} \deg(Q_i) = \deg(P_{s,j})$ and so that $c := \langle P_{s,j}, Q_1 \cdots Q_\ell \rangle \geq m_s^{(1-\ell)/2}$. The idea now is to replace $P_{s,j}$ by $P' := P_{s,j} - cQ_1 \cdots Q_\ell$, and replace occurrences of $P_{s,j}$ in our decomposition by $P' + cQ_1 \cdots Q_\ell$.

To be concrete about what this does to our decomposition, suppose that $P_{s,j}$ appears in the monomial $bP_{s,j}P_{s,j_1} \cdots P_{s,j_{s-1}}$ of h_s. We then replace $P_{s,j}$ by P'. We also introduce a number of new $P'_{s+\ell-1,j}s$ equal to $P_{s,j_1}, \ldots, P_{s,j_{s-1}}, Q_1, \ldots, Q_\ell$, and add the term $bcP_{s,j_1} \cdots P_{s,j_{s-1}} Q_1 \cdots Q_\ell$ to $h_{s+\ell-1}$. This gives us a new decomposition of P with $O_d(1)$ new $P'_{s+\ell-1,j}s$ and with $|P_{s,j}|^2$ decreased by an additive $c^2 > m_s^{1-\ell}$. Note that this process never introduces any new $P_{i,j}$ for $i \leq s$ and that each time it decreases the sum of the squares of the norms of the $P_{s,j}$ by at least m_s^{1-d}. Therefore, this procedure must eventually terminate.

It is clear that the resulting decomposition will have sufficiently small moments for its $P_{i,j}$ for all i, j with $i \leq s$ (as this is the termination condition for the above procedure). It is also clear that the resulting decomposition will satisfy all of the desired properties, except perhaps that it might have n_i too big for some $i > s$, or the sum of the square of the coefficients of h_i might be too big for some $i > s$.

To deal with the former concern, note that whenever we add new $P_{i,j}$ in this procedure, we add $O_d(1)$ such new terms and decrease the sum of the squares of the norms of the $P_{s,j}$ by at least m_s^{s-i}. Therefore, this can happen at most $n_s m_s^{i-s}$ many times. But $n_s = O_d(m_1 \cdots m_{s-1})$, so this is $O_d(m_1 \cdots m_{s-1} m_s^{i-s}) \leq O_d(m_1 \cdots m_{i-1})$, as desired.

To deal with the latter concern, when we add a new monomial to h_i whose coefficient is bc, we have taken a monomial in h_s with leading coefficient b and decreased the sum of the squares of the norms of the $P_{s,j}$ in that monomial by c^2. Hence to increase the sum of the squares of the coefficients of h_i by b^2c^2, we need to have decreased the sum over monomials of h_s of the square of the leading coefficient times the sum of the squares of the norms of the corresponding $P_{s,j}$ by the same amount. Since the latter quantity was originally at most $O_d(1)$ by the inductive hypothesis, we increase the sum of the squares of the coefficients of h_i by at most $O_d(1)$.

□

5 FT-Mollification

We let F be a degree-d polynomial threshold function, $F = \text{sgn}(p)$, where p is a degree-d multilinear polynomial in n variables whose sum of squares of coefficients equals 1. We pick positive integers $m_1, \ldots, m_d$ (their exact sizes will be determined later). For later convenience, we assume the m_i are all even. We then have a decomposition of F given by Proposition 4 as

$$\begin{aligned} F(X) &= \text{sgn}\left(\sum_{i=1}^{d} h_i(P_{i,1}(X), \ldots, P_{i,n_i}(X))\right) \\ &= f(P_1(X), P_2(X), \ldots, P_d(X)) \\ &= f(P(X)) \end{aligned}$$

where $P_i(X)$ is the vector-valued polynomial $(P_{i,1}(X), \ldots, P_{i,n_i}(X))$, P is the vector of all of them, and f is the function $f(P_1, \ldots, P_d) = \text{sgn}(\sum h_i(P_i))$. Furthermore, we have that for $k \leq dm_i$, the k^{th} moment of any coordinate of any coordinate of P_i is $O_d(\sqrt{k})^k$ (giving up an additional $\sqrt{d}$ factor in the O_d). We also have that h_i is a degree i multilinear polynomial the sum of the squares of whose coefficients is at most 1.

Our basic strategy is to approximate f by a smooth function $\tilde{f}$, and letting $\tilde{F}(X) = \tilde{f}(P(X))$. We will then proceed to prove:

$$\text{E}[F(Y)] \approx_\epsilon \text{E}[\tilde{F}(Y)] \approx_\epsilon \text{E}[\tilde{F}(X)] \approx_\epsilon \text{E}[F(X)]. \tag{1}$$

We will produce $\tilde{f}$ from f using the technique of mollification. Namely we will have $\tilde{f} = f * \rho$ for an appropriately chosen smooth function ρ. However, we will need this ρ to have several other properties so we will go into some depth here to construct it. Note: the following Lemma is very similar to Theorem IV.1 of [13].

Lemma 5. *Given an integer $n \geq 0$ and a constant C, there is a function $\rho_C : \mathbb{R}^n \to \mathbb{R}$ so that*

1. *$\rho_C \geq 0$.*
2. *$\int_{\mathbb{R}^n} \rho_C(x)dx = 1$.*
3. *For any unit vector $v \in \mathbb{R}^n$, and any non-negative integer k, $\int_{\mathbb{R}^n} |D_v^k \rho_C(x)|dx \leq C^k$, where D_v^k is the k^{th} directional derivative in the direction v.*
4. *For $D > 0$, $\int_{|x|>D} |\rho(x)|dx = O\left(\left(\frac{n}{CD}\right)^2\right)$.*

Proof. ρ_C is obtained by taking the square of the Fourier transform of a compactly supported, continuous function of width approximately C.

We prove our result $C = 2$ and we note that we can obtain other values of C by setting $\rho_C(x) = (C/2)^n \rho_2(Cx/2)$. We begin by defining

$$B(\xi) = \begin{cases} 1 - |\xi|^2 & \text{if } |\xi| \leq 1 \\ 0 & \text{else} \end{cases}$$

We then define (letting $|B|_2$ be the L^2 norm of B)

$$\rho_2(x) = \rho(x) = \frac{|\hat{B}(x)|^2}{|B|_2^2}.$$

Where $\hat{B}$ denotes the Fourier transform of B. Clearly ρ is non-negative. Also clearly

$$\int_{\mathbb{R}^n} \rho(x)dx = \frac{|\hat{B}|_2^2}{|B|_2^2} = 1$$

by the Plancherel Theorem.

For the third property we note that

$$D_v^k \rho = \frac{1}{|B|_2^2} \sum_{i=0}^{k} \binom{k}{i} D_v^i(\hat{B}) D_v^{k-i}(\overline{\hat{B}}).$$

Letting ξ be the dual vector corresponding to v we have that

$$\begin{aligned}
\left|D_v^k \rho\right|_1 &\leq \frac{1}{|B|_2^2} \sum_{i=0}^{k} \binom{k}{i} \left| D_v^i(\hat{B}) D_v^{k-i}(\overline{\hat{B}}) \right|_1 \\
&\leq \frac{1}{|B|_2^2} \sum_{i=0}^{k} \binom{k}{i} \left| D_v^i(\hat{B}) \right|_2 \left| D_v^{k-i}(\hat{B}) \right|_2 \\
&\leq \frac{1}{|B|_2^2} \sum_{i=0}^{k} \binom{k}{i} \left| \xi^i B \right|_2 \left| \xi^{k-i} B \right|_2 \\
&\leq \frac{1}{|B|_2^2} \sum_{i=0}^{k} \binom{k}{i} |B|_2^2 \\
&= \sum_{i=0}^{k} \binom{k}{i} \\
&= 2^k.
\end{aligned}$$

For the last property we note that it is enough to prove that

$$\int_{\mathbb{R}^n} |x|^2 \rho(x) dx = O(n^2).$$

We have that

$$\begin{aligned}\int_{\mathbb{R}^n} |x|^2 \rho(x) dx &= \frac{1}{|B|_2^2} \sum_{i=1}^{n} |x_i \hat{B}|_2^2 \\ &= \frac{\sum_{i=1}^{n} \left|\frac{\partial B}{\partial \xi_i}\right|_2^2}{|B|_2^2}.\end{aligned}$$

Now $\frac{\partial B}{\partial \xi_i}$ is $2\xi_i$ on the unit ball and 0 outside. Hence the sum of the squares of these is $2|\xi|^2$ on $|\xi| < 1$ and 0 outside. Hence since both numerator and denominator above are integrals of spherically symmetric functions, their ratio is equal to

$$\int_{\mathbb{R}^n} |x|^2 \rho(x) dx = \frac{2\int_0^1 r^{n+1} dr}{\int_0^1 r^{n-1}(1-r^2)^2 dr}.$$

Using integration by parts, the denominator is

$$\begin{aligned}\int_0^1 r^{n-1}(1-r^2)^2 dr &= \frac{4}{n}\int_0^1 r^{n+1}(1-r^2) dr \\ &= \frac{8}{n(n+2)}\int_0^1 r^{n+3} dr \\ &= \frac{8}{n(n+2)(n+4)}.\end{aligned}$$

Hence

$$\int_{\mathbb{R}^n} |x|^2 \rho(x) dx = \frac{n(n+4)}{4} = O(n^2).$$

□

We are now prepared to define $\tilde{f}$. We pick constants $C_1, \ldots, C_d$ (to be determined later). We let

$$\rho(P_1, \ldots, P_d) = \rho_{C_1}(P_1) \cdot \rho_{C_2}(P_2) \cdots \rho_{C_d}(P_d). \tag{2}$$

Above the ρ_{C_i} is defined on $\mathbb{R}^{n_i}$. We let $\tilde{f}$ be the convolution $\tilde{f} = f * \rho$.

6 Taylor Error

In this Section, we prove the middle approximation of Equation 1 for appropriately large k. The basic idea will be to approximate $\tilde{f}$ by its Taylor series, T. $T(P(X))$ will be a polynomial of degree at most k and hence $\mathrm{E}[T(P(Y))] = \mathrm{E}[T(P(X))]$. Furthermore, we will bound the Taylor error by some polynomial R and show that $\mathrm{E}[R(P(Y))] = \mathrm{E}[R(P(X))]$ is $O(\epsilon)$ for appropriate choices of m_i, C_i. In particular, we let T be the polynomial consisting of all of the terms of the Taylor expansion of $\tilde{f}$ whose total degree in the P_i coordinates is less than m_i for all i. Note that a polynomial of this form is about the best we can do since we only have control over the size of moments up to the m_i^{th} moment on the i^{th} block of coordinates. Our error bound will be the following

Proposition 6. *For f, P_i, C_i, m_i and T as given above, then for any value of P,*

$$|T(P) - \tilde{f}(P)| \leq \prod_{i=1}^{d} \left(1 + \frac{C_i^{m_i} |P_i|^{m_i}}{m_i!}\right) - 1.$$

First we prove a Lemma.

Lemma 7. *If g is a multivariate function, $\tilde{g} = g * \rho_C$ and T is the polynomial consisting of all terms in the Taylor expansion of $\tilde{g}$ of degree less than m, then*

$$|\tilde{g}(x) - T(x)| \leq \frac{|g|_\infty C^m |x|^m}{m!}$$

for all x.

Proof. This bound is essentially the standard bound on Taylor error applied to the restriction of $\tilde{g}$ to the line between 0 and x.

Let v be the unit vector in the direction of x. Let L be the line through 0 and x. We note that the restriction of T to L is the same as the first $m-1$ terms of the Taylor series for $\tilde{g}|_L$. Using standard error bounds for Taylor polynomials, we find that

$$|g(x) - T(x)| \leq \frac{|D_v^m \tilde{g}|_\infty |x|^m}{m!}.$$

But

$$\begin{aligned} |D_v^m \tilde{g}|_\infty &= |g * D_v^m \rho_C|_\infty \\ &\leq |g|_\infty |D_v^m \rho_C|_1 \\ &\leq |g|_\infty C^m. \end{aligned}$$

Plugging this into the above yields our result. □

Proof of Proposition 6. he basic idea of the proof will be to repeatedly apply Lemma 7 to one batch of coordinates at a time. We begin by defining some operators on the space of bounded functions on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d}$. For such g, define $g^{\tilde{i}}$ to be the convolution of g with ρ_{C_i} along the i^{th} set of coordinates. Define g^{T_i} to be the Taylor polynomial in the i^{th} set of variables of $g^{\tilde{i}}$ obtained by taking all terms of total degree less than m_i. Note that for $i \neq j$ the operations $\tilde{i}$ and T_i commute with the operations $\tilde{j}$ and T_j since they operate on disjoint sets of coordinates. Note that $\tilde{f} = f^{\tilde{1}\tilde{2}\cdots\tilde{d}}$ and $T = f^{T_1 T_2 \cdots T_d}$. For $1 \leq i \leq d$ let $f_i = f^{\tilde{1}\tilde{2}\cdots\tilde{i}}$ and $T_i = f^{T_1 T_2 \cdots T_i}$.

We prove by induction on s that

$$|T_s(P) - f_s(P)| \leq \prod_{i=1}^{s} \left(1 + \frac{C_i^{m_i} |P_i|^{m_i}}{m_i!}\right) - 1.$$

As a base case, we note that the $s = 0$ case of this is trivial.

Assume that

$$|T_s(P) - f_s(P)| \leq \prod_{i=1}^{s} \left(1 + \frac{C_i^{m_i} |P_i|^{m_i}}{m_i!}\right) - 1.$$

We have that

$$|T_{s+1}(P) - f_{s+1}(P)| \leq |T_s^{T_{s+1}}(P) - T_s^{\widetilde{s+1}}(P)| + |T_s^{\widetilde{s+1}}(P) - f_s^{\widetilde{s+1}}(P)|.$$

Note that

$$T_s^{\widetilde{s+1}}(P) - f_s^{\widetilde{s+1}}(P) = (T_s - f_s)^{\widetilde{s+1}}(P).$$

Therefore since $\widetilde{s+1}$ involves only convolution with a function of L^1 norm 1 we have that

$$|T_s^{\widetilde{s+1}}(P) - f_s^{\widetilde{s+1}}(P)| \leq |T_s(P) - f_s(P)|_{\infty,s+1}$$

where the subscript denotes the L^∞ norm over just the $s+1^{st}$ set of coordinates. By the inductive hypothesis this is at most

$$\prod_{i=1}^{s} \left(1 + \frac{C_i^{m_i} |P_i|^{m_i}}{m_i!}\right) - 1.$$

On the other hand, applying Lemma 7 we have that

$$|T_s^{T_{s+1}}(P) - T_s^{\widetilde{s+1}}(P)| \leq \frac{C_{s+1}^{m_{s+1}} |P_{s+1}|^{m_{s+1}}}{m_{s+1}!} |T_s|_{\infty,s+1}.$$

By the inductive hypothesis

$$|T_s|_{\infty,s+1} \leq |f_s|_{\infty,s+1} + |T_s - f_s|_{\infty,s+1} \leq \prod_{i=1}^{s} \left(1 + \frac{C_i^{m_i} |P_i|^{m_i}}{m_i!}\right).$$

Combining the above bounds, we find that

$$\begin{aligned}|T_{s+1} - f_{s+1}| \leq & \prod_{i=1}^{s}\left(1 + \frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right) - 1 \\ & + \frac{C_{s+1}^{m_{s+1}}|P_{s+1}|^{m_{s+1}}}{m_{s+1}!}\prod_{i=1}^{s}\left(1 + \frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right) \\ = & \prod_{i=1}^{s+1}\left(1 + \frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right) - 1.\end{aligned}$$

□

We can now prove the desired approximation result

Proposition 8. *If $\tilde{F}, P, T$ as above with $m_i = \Omega_d(n_i C_i^2)$, $m_i \geq \log(2^d/\epsilon)$ for all i, and if $k \geq dm_i$ for all i, and if X and Y are k-independent families of standard Gaussians then*

$$E[\tilde{F}(Y)] \approx_\epsilon E[\tilde{F}(X)].$$

Proof. We note that since $T \circ P$ is a polynomial of degree at most k we have that $\mathrm{E}[T(P(X))] = \mathrm{E}[T(P(Y))]$. Hence we just need to show that

$$\mathrm{E}[|F - T|(P(X))], \mathrm{E}[|F - T|(P(Y))] = O(\epsilon).$$

We will show this only for X as the proof for Y is analogous. By Proposition 6 we have that $|F - T|$ is bounded by

$$\prod_{i=1}^{d}\left(1 + \frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right) - 1.$$

This is a sum over non-empty subsets $S \subseteq \{1, 2, \ldots, d\}$ of

$$\prod_{i \in S}\left(\frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right).$$

Since there are only $2^d - 1$ such S, it is enough to show that each term individually has expectation $O(\epsilon/2^d)$. On the other hand, we have by AM-GM that this term is at most

$$\frac{1}{|S|}\sum_{i \in S}\left(\frac{C_i^{m_i}|P_i|^{m_i}}{m_i!}\right)^{|S|}.$$

Now the expectation of $|P_i|^{m_i|S|}$ is at most $n_i^{m_i|S|/2}$ times the average of the $m_i|S|^{th}$ moments of the coordinates of P_i. Since these moments are given by the expectations of polynomials (m_i

is even) of degree at most k, the are determined by the k-independence of X. By assumption these moments are at most $O_d(\sqrt{m_i|S|})^{m_i|S|}$. Therefore, the $m_i|S|^{th}$ the moment of $|P_i|$ is at most $O_d(\sqrt{n_i m_i|S|})^{m_i|S|}$. Hence the error is at most

$$\begin{aligned} O(2^d)\max_{i,s}\left\{O_d\left(\frac{C_i\sqrt{n_i m_i s}}{m_i}\right)^{m_i s}\right\} &= O(2^d)\max_{i,s}\left\{O_d\left(\frac{C_i\sqrt{n_i}}{\sqrt{m_i}}\right)^{m_i s}\right\} \\ &\leq O(2^d)e^{-\min_i m_i} \\ &= O(\epsilon). \end{aligned}$$

□

7 Approximation Error

In this Section, we will prove the first and third approximations in Equation 1. We begin with the first, namely

$$\mathrm{E}[F(Y)] \approx_\epsilon \mathrm{E}[\tilde{F}(Y)].$$

Our basic strategy will be to bound

$$|\mathrm{E}[F(Y)] - \mathrm{E}[\tilde{F}(Y)]| \leq \mathrm{E}[|F(Y) - \tilde{F}(Y)|].$$

In order to get a bound on this we will first show that $F - \tilde{F}$ is small except where $p(Y)$ is small, and then use anti-concentration results to show that this happens with small probability. This will be true because ρ is small away from 0. We begin by proving a Lemma to this effect.

Lemma 9. *Let ρ be the function defined in Equation 2. Then for any $D > 0$ we have that*

$$\int_{\substack{(x_1,\ldots,x_d)\in\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_d} \\ \exists i:|x_i|>Dn_i\sqrt{d}/C_i}} |\rho(x)|dx = O(D^{-2})$$

This will hold essentially because of the concentration property held by each ρ_{C_i}.

Proof. We integrate over the region where $|x_i| > Dn_i\sqrt{d}/C_i$ for each i. This is a product over $j \neq i$ of $\int_{\mathbb{R}^{n_j}} \rho_{C_j}(x_j)$ times $\int_{|x|>Dn_i\sqrt{d}/C_i} |\rho(x)|dx$. By Lemma 5, the former integrals are all 1, and the latter is $O(D^{-2}/d)$. Summing over all possible i yields $O(D^{-2})$. □

Recall that f was sgn$\circ h$, where $h = \sum h_i$ given in the decomposition of p from Proposition 4. Recall that $\tilde{f} = f * \rho$. We want to bound the error in approximating f by $\tilde{f}$. The following, is a direct consequence of Lemma 9.

Lemma 10. *Suppose $x = (x_1, \ldots, x_d) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$. Suppose also that for some $D > 0$ and for all $y = (y_1, \ldots, y_n) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$ so that $|x_i - y_i| \leq Dn_i\sqrt{d}/C_i$ that $h(x)$ and $h(y)$ have the same sign, then*

$$|f(x) - \tilde{f}(x)| = O\left(\min\{1, D^{-2}\}\right).$$

Proof. To show that the error is $O(1)$, we note that $\rho \geq 0$ and $\int \rho(x)dx = 1$. Hence, $\tilde{f}(x) = (f * \rho)(x) \in [\inf(f), \sup(f)] \subseteq [-1, 1]$. Therefore, $|f - \tilde{f}| \leq |f| + |\tilde{f}| \leq 2$.

For the latter, we note that $\tilde{f}(x) = \int_y f(y)\rho(x - y)dy$. We note that since the total integral of ρ is 1 that

$$f(x) - \tilde{f}(x) = \int_y (f(x) - f(y))\rho(x - y)dy.$$

We note that by assumption unless $|x_i - y_i| > Dn_i\sqrt{d}/C_i$ for some i that the integrand is 0. But outside of this, the integrand is at most $2\rho(x - y)$. By Lemma 9 the total integral of this is $O(D^{-2})$. □

We now know that f is near $\tilde{f}$ at points x not near the boundary between the $+1$ and -1 regions. Since we cannot directly control the size of these regions, we want to relate this to the region where $|h(x)|$ is small. This should work since unless x is very large, h will have derivatives that aren't too big. In particular, we prove the following.

Lemma 11. *Let $x \in \mathbb{R}^n$. Suppose that we have $B_i \geq 0$ so that $|P_{i,j}(x)| \leq B_i$ for all i, j. Then we have that*

$$|F(x) - \tilde{F}(x)| \leq$$
$$O_d\left(\min\left\{1, \max\left\{\left(\frac{|p(x)|}{\sum_{i=1}^d n_i^2 B_i^{i-1}/C_i}\right)^{-2}, \max_i\left\{\left(\frac{B_i C_i}{n_i}\right)^{-2}\right\}\right\}\right\}\right).$$

Proof. The bound of $O(1)$ follows immediately from Lemma 10. For the other bound, let

$$D = \min\left\{\frac{|p(x)|}{d2^d \sum_{i=1}^d n_i^2 B_i^{i-1}/C_i}, \min_i\left\{\frac{B_i C_i}{n_i\sqrt{d}}\right\}\right\}.$$

By Lemma 10, it suffices to show that for any $Q = (Q_1, \ldots, Q_n) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$ so that $|Q_i - P_i(x)| \leq Dn_i\sqrt{d}/C_i$ for all i, that $h(P(x)) = p(x)$ and $h(Q)$ have the same sign. To do this we write $h = h_1 + \cdots + h_d$ and we note that

$$|h(P(x)) - h(Q)| \leq \sum_{i=1}^d |P_i(x) - Q_i||h_i'(z)|.$$

Where $h_i'(z)$ is the directional derivative of h_i in the direction from $P_i(x)$ to Q_i, and z is some point along this line. First, note that $|Q_i - P_i(x)| \leq B_i$. Therefore, each coordinate of z is at most $2B_i$. Now note that h_i is a sum of at most n_i monomials of degree i with coefficients at most 1. The derivative of each monomial at z is at most $\sqrt{d}2^d B_i^{i-1}$. Therefore $|h_i'(z)| \leq \sqrt{d}2^d n_i B_i^{i-1}$. Therefore

$$\begin{aligned}
|h(P(x)) - h(Q)| &\leq \sum_{i=1}^{d} |P_i(x) - Q_i||h_i'(z)| \\
&\leq \sum_{i=1}^{d} (Dn_i\sqrt{d}/C_i)(\sqrt{d}2^d n_i B_i^{i-1}) \\
&\leq D\sum_{i=1}^{d} d2^d n_i^2 B_i^{i-1}/C_i \\
&\leq |h(P(x))|.
\end{aligned}$$

Therefore, $h(P(x))$ and $h(Q)$ have the same sign, so our bound follows by Lemma 10. □

We take this bound on the approximation error and prove the following Lemma on the error of expectations.

Lemma 12. *Let Z be a random variable valued in $\mathbb{R}^n$. Let $B_i > 1$ be real numbers. Let $M = \sum_{i=1}^{d} n_i^2 B_i^{i-1}/C_i$. Then*

$$\begin{aligned}
&|E[F(Z)] - E[\tilde{F}(Z)]| \leq \\
&O_d(Pr(|P_{i,j}(Z)| > B_i \text{ for some } i,j) + M + Pr(|p(Z)| \leq \sqrt{M})).
\end{aligned}$$

Furthermore, $E[F(Z)]$ is bounded above by

$$E[\tilde{F}(Z)] + O_d(Pr(|P_{i,j}(Z)| > B_i \text{ for some } i,j) + M) + 2Pr(-\sqrt{M} < p(Z) < 0),$$

and below by

$$E[\tilde{F}(Z)] + O_d(Pr(|P_{i,j}(Z)| > B_i \text{ for some } i,j) + M) - 2Pr(0 < p(Z) < \sqrt{M}).$$

Proof. We note that $|F(Z) - \tilde{F}(Z)| = O(1)$. Also note that $\frac{1}{M} \leq \frac{B_i C_i}{n_i}$ for all i. The first inequality follows by noting that Lemma 11 implies that unless $|P_{i,j}(Z)| > B_i$ for some i,j that the following hold:

1. If $|p(z)| < \sqrt{M}$, $|F(Z) - \tilde{F}(Z)| \leq 2$.
2. If $|p(z)| \geq \sqrt{M}$, $|F(Z) - \tilde{F}(Z)| = O_d(M)$.

The other two inequalities follow from noting that if $p(z) < 0$, then $F(Z) \leq \tilde{F}(Z)$ and if $p(Z) > 0$ then $F(Z) \geq \tilde{F}(Z)$. □

We are almost ready to prove the first of our approximation results, but we first need a theorem on the anticoncentration of Gaussian polynomials. In particular a consequence of [7] Theorem 8 is:

Theorem 13 (Carbery and Wright)**.** *Let p be a degree d polynomial, and Y a standard Gaussian. Suppose that $E[p(Y)^2] = 1$. Then, for $\epsilon > 0$,*

$$Pr(|p(Y)| < \epsilon) = O(d\epsilon^{1/d}).$$

We are now prepared to prove our approximation result.

Proposition 14. *Let $p, F, \tilde{F}, h, m_i, n_i, C_i$ be as above and let $\epsilon > 0$. Let $B_i = \Omega_d(\sqrt{\log(n_i/\epsilon)})$ be some real numbers. Suppose that $m_i > B_i^2$ and that $C_i = \Omega_d(n_i^2 B_i^{i-1} \epsilon^{-2d})$ for all i. Then, if the implied constants for the bounds on B_i and C_i are large enough,*

$$|E[F(Y)] - E[\tilde{F}(Y)]| = O(\epsilon).$$

Proof. We bound the error using Lemma 12. We note that the probability that $|P_{i,j}(Y)| \geq B_i$ can be bounded by looking at the $\log(dn_i/\epsilon) = \ell^{th}$ moment, yielding a probability of $\frac{O_d(\sqrt{\ell})^\ell}{B_i^\ell} \leq e^{-\ell} = \frac{\epsilon}{dn_i}$. Taking a union bound over all j gives a probability of at most $\frac{\epsilon}{d}$. Taking a union bound over i yields a probability of at most ϵ.

Next we note that

$$M = \sum_{i=1}^{d} n_i^2 B_i^{i-1}/C_i = O_d(\epsilon^{2d}).$$

Hence if our constants were chosen to be large enough, by Theorem 13

$$\Pr(|p(Y)| < \sqrt{M}) = O(\epsilon).$$

This proves our result. □

If we could prove Proposition 14 for X instead of Y, we would be done. Unfortunately, Theorem 13 does not immediately apply for families that are merely k-independent. Fortunately, we can work around this to prove Proposition 2. In particular, we will use the inequality versions of Lemma 12 to obtain upper and lower bounds on $\mathrm{E}[F(X)]$ in terms of $\mathrm{E}[\mathrm{sgn}(p(Y) + c)]$, and make use of anticoncentration for $p(Y)$.

Proof of Proposition 2. Let $B_i = \Omega_d(\sqrt{\log(1/\epsilon)})$ with sufficiently large constants. Define m_i and C_i so that $C_i = \Omega_d\left(\left(\prod_{j=1}^{i-1} m_j\right)^2 B_i^{i-1}\epsilon^{-2d}\right)$ and $m_i \geq \Omega_d\left(\left(\prod_{j=1}^{i-1} m_j\right) C_i^2\right), \log(2^d/\epsilon), B_i^2,$

all with sufficiently large constants. Note that this is achievable by setting $C_i = \Omega_d\left(\epsilon^{-7^i d}\right)$, $m_i = \Omega_d\left(\epsilon^{-3\cdot 7^i d}\right)$. Let $k = d\max_i m_i$. Note k can be as small as $O_d(\epsilon^{-4d\cdot 7^d})$. Using these parameters, define $n_i, h_i, P_{i,j}, f, \tilde{f}, \tilde{F}$ as described above. Note that since $n_i = O_d\left(\prod_{j=1}^{i-1} m_i\right)$, that $C_i = \Omega_d(n_i^2 B_i^{i-1}\epsilon^{-2d})$ and $m_i = \Omega_d(n_i C_i^2)$. Let Y be a family of independent standard Gaussians and X a family of k-independent standard Gaussians.

Propositions 6 and 14 imply that

$$\mathrm{E}[F(Y)] \approx_\epsilon \mathrm{E}[\tilde{F}(Y)] \approx_\epsilon \mathrm{E}[\tilde{F}(X)].$$

We note that the M in Lemma 12 is $O_d(\epsilon^{2d})$ with sufficiently small constant. Also note that by looking at the $\log(dn_i/\epsilon)$ moments of the $P_{i,j}$ that the last probability is $O(\epsilon)$. Therefore, combining the later parts of Lemma 12 with the above fact that $\mathrm{E}[F(Y)] \approx_\epsilon \mathrm{E}[\tilde{f}(X)]$, we obtain that

$$\mathrm{E}[F(X)] \geq \mathrm{E}[F(Y)] + O(\epsilon) - 2\mathrm{Pr}(0 < p(X) < O_d(\epsilon^d)),$$

and

$$\mathrm{E}[F(X)] \leq \mathrm{E}[F(Y)] + O(\epsilon) + 2\mathrm{Pr}(-O_d(\epsilon^d) < p(X) < 0).$$

But this implies that

$$\mathrm{E}[\mathrm{sgn}(p(X) - O_d(\epsilon^d))] \leq \mathrm{E}[F(Y)] + O(\epsilon),$$

and

$$\mathrm{E}[\mathrm{sgn}(p(X) + O_d(\epsilon^d))] \geq \mathrm{E}[F(Y)] + O(\epsilon).$$

On the other hand, applying to above to the polynomials $p \pm O_d(\epsilon^d)$,

$$\mathrm{E}[\mathrm{sgn}(p(Y) - O_d(\epsilon^d))] + O(\epsilon) \leq \mathrm{E}[F(X)] \leq \mathrm{E}[\mathrm{sgn}(p(Y) + O_d(\epsilon^d))] + O(\epsilon).$$

But we have that

$$\mathrm{E}[\mathrm{sgn}(p(Y) - O_d(\epsilon^d))] \leq \mathrm{E}[F(Y)] \leq \mathrm{E}[\mathrm{sgn}(p(Y) + O_d(\epsilon^d))].$$

Furthermore, $\mathrm{sgn}(p(Y) - O_d(\epsilon^d))$ and $\mathrm{sgn}(p(Y) + O_d(\epsilon^d))$ differ by at most 2, and only when $|p(Y)| = O_d(\epsilon^d)$. By Theorem 13 this happens with probability $O_d(\epsilon)$. Therefore we have that all of the expectations above are within $O_d(\epsilon)$ of $\mathrm{E}[F(Y)]$, and hence $\mathrm{E}[F(X)] = \mathrm{E}[F(Y)] + O_d(\epsilon)$. Decreasing the value of ϵ by a factor depending only on d (and increasing k by a corresponding factor) yields our result. □

8 General Polynomials

We have proved our Theorem for multilinear polynomials, but would like to extend it to general polynomials. Our basic idea will be to show that a general polynomial is approximated by a multilinear polynomial in perhaps more variables.

Lemma 15. *Let p be a degree-d polynomial and $\delta > 0$. Then there exists a multilinear, degree-d polynomial p_δ (in perhaps a greater number of variables) so that for every k-independent family of random Gaussians X, there is a (correlated) k-independent family of random Gaussians $\tilde{X}$ so that*

$$Pr(|p(X) - p_\delta(\tilde{X})| > \delta) < \delta.$$

Proof. The basic idea of the proof is to let each coordinate X_i of X be written as a sum $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} X_{i,j}$. Expanding our the polynomial $p(X)$, and approximating terms of the form $\frac{1}{N} \sum_{j=1}^{N} X_{i,j}^2$ by 1, we are left with a multilinear polynomial, plus a polynomial which is likely small.

We will pick some large integer N (how large we will say later). If $X = (X_1, \ldots, X_n)$, we let $\tilde{X} = (X_{i,j}), 1 \leq i \leq n, 1 \leq j \leq N$. For fixed i we let the collection of $X_{i,j}$ be the standard collection of N standard Gaussians subject to the condition that $X_i = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} X_{i,j}$. Equivalently, $X_{i,j} = \frac{1}{\sqrt{N}} X_i + Y_{i,j}$ where the $Y_{i,j}$ are Gaussians with variance $1 - 1/N$ and covariance $-1/N$ with each other.

$\tilde{X}$ is k-independent because given any $i_1, \ldots, i_k$, $j_1, \ldots, j_k$ we can obtain the X_{i_ℓ, j_ℓ} by first picking the X_{i_ℓ} randomly and independently, and picking the Y_{i_ℓ, j_ℓ} independently of those. We note that this yields the same distribution we would get by setting all of the $X_{i_\ell, k}$ to be random independent Gaussians, and letting $X_i = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} X_{i,j}$.

We now need to construct p_δ with the appropriate property. The idea will be to replace each term X_i^k in each monomial in p with some degree-k, multilinear polynomial in the $X_{i,j}$. This will yield a multilinear, degree-d polynomial in $\tilde{X}$. We will want this new polynomial to be within δ' of X_i^k with probability $1 - \delta'$ for δ' some small positive number depending on p and δ. This will be enough since if $\delta' < \delta/(2dn)$ the approximation will hold for all i, k with probability at least $1 - \delta/2$. Furthermore, with probability $1 - \delta/2$, each of the $|X_i|$ will be at most $O(\log(n/\delta))$. Therefore, if this holds and each of the replacement polynomials is off by at most δ', then the value of the full polynomial will be off by at most $O(\log^d(n/\delta)\delta')$ times the sum of the coefficients of p. Hence if we can achieve this for δ' small enough we are done.

Hence we have reduced our problem to the case of $p(X) = p(X_1) = X_1^d$. For simplicity of notation, we use X instead of X_1 and X_j instead of $X_{1,j}$. We note that

$$X^d = N^{-d/2} \left(\sum_{i=1}^{N} X_i \right)^d.$$

Unfortunately, this is not a multilinear polynomial in the X_i. Fortunately, it almost is. Expanding it out and grouping terms based on the multiset of exponents occurring in them we find that

$$X^d = N^{-d/2} \sum_{\substack{a_1 \le \ldots \le a_k \\ \sum a_i = d}} \binom{d}{a_1, a_2, \ldots, a_k} \sum_{\substack{i_1,\ldots,i_k \in \{1,\ldots,N\} \\ i_j \text{ distinct} \\ i_j < i_{j+1} \text{ if } a_j = a_{j+1}}} \prod_{j=1}^{k} X_{i_j}^{a_j}.$$

Letting b_ℓ be the number of a_i that are equal to ℓ we find that this is

$$N^{-d/2} \sum_{\substack{a_1 \le \ldots \le a_k \\ \sum a_i = d}} \binom{d}{a_1, a_2, \ldots, a_k} \prod_{\ell} \frac{1}{b_\ell!} \sum_{\substack{i_1,\ldots,i_k \in \{1,\ldots,N\} \\ i_j \text{ distinct}}} \prod_{j=1}^{k} X_{i_j}^{a_j}.$$

Or rewriting slightly, this is

$$\sum_{\substack{a_1 \le \ldots \le a_k \\ \sum a_i = d}} \binom{d}{a_1, a_2, \ldots, a_k} \prod_{\ell} \frac{1}{b_\ell!} \sum_{\substack{i_1,\ldots,i_k \in \{1,\ldots,N\} \\ i_j \text{ distinct}}} \prod_{j=1}^{k} \left(\frac{X_{i_j}}{\sqrt{N}}\right)^{a_j}.$$

Now, with probability $1-\delta$, $\left|\sum_i \frac{X_i}{\sqrt{N}}\right| = O(\log(1/\delta))$. Furthermore, with probability tending to 1 as N goes to infinity,

$$\left(\sum_i \left(\frac{X_i}{\sqrt{N}}\right)^2\right) = 1 + O(\delta/\log^d(1/\delta)),$$

and $\left(\sum_i \left(\frac{X_i}{\sqrt{N}}\right)^a\right) = O(\delta/\log^d(1/\delta))$ for each $3 \le a \le d$. If all of these events hold, then each term in the above with some $a_j > 2$ will be $O(\delta)$, and any terms with some $a_j = 2$ will be within $O(\delta)$ of

$$\sum_{\substack{i_1,\ldots,i'_k \in \{1,\ldots,N\} \\ i_j \text{ distinct}}} \prod_{j=1}^{k'} \frac{X_{i_j}}{\sqrt{N}}$$

where k' is the largest j so that $a_j = 1$. This gives a multilinear polynomial, that with probability $1-\delta$ is within $O_d(\delta)$ of $p(X)$. Perhaps decreasing δ to deal with the constant in the O_d yields our result. □

We can now prove Theorem 1 by applying Proposition 2 to p_δ.

Proof of Theorem 1. Let p be a normalized, degree-d polynomial. Let k be as required by Proposition 2. Let Y be a family of independent standard Gaussians and X a k-independent family of standard Gaussians. Fix $\delta = (\epsilon/d)^d$. Let $p_\delta, \tilde{X}, \tilde{Y}$ be as given by Lemma 15. We need to show that $\Pr(p(X) > 0) = \Pr(p(Y) > 0) + O(\epsilon)$. By the specification of p_δ in Lemma 15,

$$\Pr(p(X) > 0) \geq \Pr(p_\delta(\tilde{X}) > \delta) - \delta.$$

Applying Proposition 2, to the multilinear polynomial $p_\delta - \delta$, this is at least

$$\Pr(p_\delta(\tilde{Y}) > \delta) + O(\epsilon).$$

Since $\tilde{Y}$ is ℓ-independent for all ℓ (since Y is), it is actually an independent family of Gaussians. Therefore by Theorem 13, $\Pr(|p(Y)| < \delta) = O(d\delta^{1/d}) = O(\epsilon)$. Hence

$$\Pr(p(X) > 0) \geq \Pr(p_\delta(\tilde{Y}) > -\delta) + O(\epsilon).$$

Noting that with probability $1 - \delta$ that $p_\delta(\tilde{Y})$ is at most δ less than $p(Y)$, this is at least

$$\Pr(p(Y) > 0) + O(\epsilon).$$

So

$$\Pr(p(X) > 0) \geq \Pr(p(Y) > 0) + O(\epsilon).$$

Similarly

$$\Pr(p(X) < 0) \geq \Pr(p(Y) < 0) + O(\epsilon).$$

Combining these we clearly have

$$\Pr(p(X) > 0) = \Pr(p(Y) > 0) + O(\epsilon)$$

as desired. □

9 Conclusion

The bounds on k presented in this Chapter are far from tight. At the very least the argument in Lemma 12 could be strengthened by considering a larger range of cases of $|p(x)|$ rather than just whether or not it is larger than $\sqrt{M}$. At very least, this would give us bounds on k of the form $O_d(\epsilon^{-x^d})$ for some x less than 7. I suspect that the correct value of k is actually $O(d^2\epsilon^{-2})$, and in fact such large k will actually be required for $p(x) = \prod_{i=1}^{d}(\sum_{j=1}^{k} x_{i,j})$. On the other hand, this bound is at the moment somewhat beyond our means. It would be nice at least to see if a bound of the form $k = O_d(\epsilon^{-\mathrm{poly}(d)})$ can be proven. The main contribution of this work is prove that there is some sufficient k that depends on only d and ϵ.

An interesting question is whether or not this result generalizes the the Bernoulli setting. All of our arguments should carry over trivially with two exceptions. Firstly, our anticoncentration bounds will not hold in general for polynomials of Bernoulli random variables. Fortunately, there are standard techniques in place for dealing with this problem. The invariance principle of [47] implies similar anticoncentration results for low-influence Bernoulli polynomials. For general polynomials, the regularity Lemma of [15] allows one to write the polynomial threshold function as a bounded depth decision tree followed by either an almost certainly constant function, or a threshold function for a regular polynomial. The second problem is more difficult to get around. In particular, to obtain an analogous structure Theorem for polynomials of Bernoulli variables, it would first be necessary to find an appropriate version of [39] Theorem 1, although with such a result in hand, I see no further trouble generalizing the results of this Chapter.

Chapter G

Noise Sensitivity of Polynomial Threshold Functions

1 Introduction

1.1 Background

One property of interest for a Boolean function is that of sensitivity. That is some measure of how stable the value of the function is under small changes to its input. Various notions of sensitivity have found applications in a number of areas including hardness of approximation [34], hardness amplification [51], quantum complexity [57], and learning theory [35]. We will discuss the last of these in more detail shortly. In this Chapter, we discuss bounds on the sensitivity of polynomial threshold functions.

1.2 Measures of Sensitivity

We will concern ourselves with four different measures of sensitivity of a Boolean function; two for Bernoulli inputs, and two for Gaussian inputs. Our results shall focus only on the latter two measures, but the first two are useful for motivational purposes.

Perhaps the most natural notion of sensitivity is that of average sensitivity. This measures the probability that flipping a randomly chosen bit of a random Bernoulli input changes the value of the function. In particular we define $\mathrm{AS}(f)$ to be

$$\mathrm{E}_x[\text{Number of bits of } x \text{ that when flipped would change the value of } f].$$

A slightly less focused notion is that of noise sensitivity. This is a measure of the probability that randomly changing some fraction of the bits will change the value of f. In particular we define the noise sensitivity with noise rate ϵ to be

$$\mathrm{NS}_\epsilon(f) := \mathrm{Pr}_{x,z}(f(x) \neq f(z))$$

where x is Bernoulli and z is obtained from x by flipping each bit independently with probability ϵ.

The concept of noise sensitivity translates naturally into the Gaussian setting. In particular we define the Gaussian noise sensitivity with noise rate ϵ to be

$$\mathrm{GNS}_\epsilon(f) := \Pr(f(X) \neq f(Z))$$

where X is an n-dimensional Gaussian random variable, and $Z = (1-\epsilon)X + \sqrt{2\epsilon - \epsilon^2}Y$ for Y an independent n-dimensional Gaussian. Notice that Z is chosen here to be a standard Gaussian whose covariance with X is $1-\epsilon$ in each coordinate, agreeing with the noise sensitivity in the Bernoulli case. In fact, we can think of the Gaussian noise sensitivity as a special case of the Bernoulli noise sensitivity by approximating each Gaussian random variable as a properly scaled average of a large number of Bernoulli random variables.

There is a closely related concept called the Gaussian surface area, which, instead of directly measuring the sensitivity of a function, measures the boundary between its $+1$ and -1 regions. We define the Gaussian surface area of a set A to be

$$\Gamma(A) := \liminf_{\delta \to 0} \frac{\mathrm{GaussianVolume}(A_\delta \backslash A)}{\delta},$$

where the Gaussian volume of a region R is $\Pr(X \in R)$ for X a Gaussian random variable, and where A_δ is the set of points x so that $d(x, A) \leq \delta$ (under the Euclidean metric). We note that if A is a sufficiently nice region with a smooth boundary, then its Gaussian surface area should be equal to

$$\int_{\partial A} \phi(x) d\sigma.$$

Here $\phi(x)$ is the Gaussian density, and $d\sigma$ is the surface measure on ∂A. This formulation justifies the name of "surface area". Furthermore, if A is such a region, then its Gaussian surface area should be equal to

$$\lim_{\delta \to 0} \frac{\mathrm{GaussianVolume}((\partial A)_\delta)}{2\delta}.$$

The factor of 2 above comes from the fact that $(\partial A)_\delta$ has volume both inside and outside of A. These equalities are proven for the case of A the set of positive values of a square-free polynomial threshold function in Proposition 10.

For f a Boolean function, we define

$$\Gamma(f) := \Gamma(f^{-1}(1)).$$

The concepts of Gaussian noise sensitivity and surface area are related to each other by noting that the noise sensitivity is roughly the probability that X is close enough to the

boundary that wiggling it will push it over the boundary. In particular it is proved in [40] that for any Boolean function, f,

$$\frac{\mathrm{GNS}_\epsilon(f)}{2} \leq \frac{\arccos(1-\epsilon)}{\sqrt{2\pi}}\Gamma(f).$$

We essentially prove an asymptotic converse to this statement for f a polynomial threshold function in Section 3.

1.3 Previous Work

In this Chapter, we study the sensitivity of polynomial threshold functions. Before stating our major results, we will provide a brief overview of the previous work in this area.

The most ambitious Conjecture in this field is the Gotsman-Linial conjecture [16], which states that the largest average sensitivity of a degree-d polynomial threshold function is obtained by taking the product of linear threshold functions slicing through the middle d layers of hypercube. This would imply that

$$\mathrm{AS}(f) \leq d\sqrt{\frac{2n}{\pi}}.$$

This would in turn imply optimal or nearly optimal bounds for our other measures. In particular, by [14] Theorem 7.1, it would imply that

$$\mathrm{NS}_\epsilon(f) \leq d\sqrt{\frac{2\epsilon}{\pi}}.$$

Furthermore, this bound would be asymptotically correct for small ϵ and properly chosen f.

Treating Gaussian random variables as averages of Bernoulli random variables, we could use this to obtain a bound on the Gaussian noise sensitivity. In particular, by [14] Proposition 9.2, this would imply

$$\mathrm{GNS}_\epsilon(f) \leq d\sqrt{\frac{2\epsilon}{\pi}}.$$

As we shall see, this is larger than the correct asymptotic by a factor of $\sqrt{\pi}$. I believe that this discrepancy comes from the fact that in the Bernoulli case things can be arranged so that the boundary makes particular angles with edges of the hypercube, while the Gaussian case is necessarily isotropic.

As we shall see in Section 3, this bound would imply a bound on the Gaussian surface area of

$$\Gamma(f) \leq \frac{d}{\sqrt{2}},$$

which is again off from the optimal by a factor of $\sqrt{\pi}$.

Unfortunately, we are far from proving these conjectures. The Gotsman-Linial Conjecture is known for $d = 1$, but is much more difficult for larger degrees. Non-trivial bounds on the average sensitivity were proved independently by [14] and [21] (these papers were later merged and appeared as [12]). They each proved bounds of $O(n^{1-1/O(d)})$ on the average sensitivity and $O_d(\epsilon^{1-1/O(d)})$ on the Gaussian and Bernoulli noise sensitivity. These results are proved by obtaining results in the Gaussian setting in such a way that they can be carried over to the Bernoulli setting. The proofs involve a number of advanced techniques including the invariance principle, and concentration and anti-concentration results for polynomials of Gaussians.

Much less was previously known about Gaussian surface area. For $d = 1$, a simple computation yields the optimal bound of $\frac{1}{\sqrt{2\pi}}$. [35] proved a bound of 1 for the Gaussian surface area of balls, and more recently, [48] proved a bound of $O(1)$ for centrally symmetric, degree 2 polynomial threshold functions. Essentially nothing was known about the Gaussian surface area of higher degree polynomial threshold functions.

1.4 Statement of Results

We focus on proving two main results.

Theorem 1. *If f is a degree-d polynomial threshold function, then*

$$GNS_\epsilon(f) \leq \frac{d \arcsin(\sqrt{2\epsilon - \epsilon^2})}{\pi} \sim \frac{d\sqrt{2\epsilon}}{\pi} = O(d\sqrt{\epsilon}).$$

Furthermore, this bound is asymptotically tight as $\epsilon \to 0$ for f the threshold function of any square-free product of homogeneous linear functions.

Theorem 2. *If f is a degree-d polynomial threshold function then $\Gamma(f) \leq \frac{d}{\sqrt{2\pi}}$. This bound is again optimal for f the threshold function of any square-free product of homogeneous linear functions.*

1.5 Application to Agnostic Learning

We briefly describe one of the applications of our work to agnostic learning algorithms for polynomial threshold functions.

1.5.1 Overview of Agnostic Learning

Agnostic learning is a model for machine learning. Suppose that there is an unknown distribution D on $X \times \{-1, 1\}$ (for some set X) with known marginal distribution D_X on X. We think of X as the observable data about some object and the element of $\{-1, 1\}$ as a

bit we are trying to predict. We would like an algorithm that, given a number of samples from D, outputs a function $h : X \to \{-1, 1\}$ so that for $(x, y) \sim D$, $h(x) = y$ with high probability. In other words, h allows us to predict y given x for (x, y) taken from D. For arbitrary D, there is little hope of success as we have no way to predict what value of y should correspond to values of x that we have not seen in our training data. To repose this is a potentially solvable problem, the agnostic learning model requires merely that h is competitive against some class of predictors. We define a concept class to be some set C of functions $X \to \{-1, 1\}$. We let $opt = \inf_{f \in C} \Pr_{(x,y)\sim D}(f(x) \neq y)$. In the agnostic learning model, the objective is to find an h so that $\Pr_{(x,y)\sim D}(h(x) \neq y) \leq opt + \epsilon$.

1.5.2 Connection to Gaussian Surface Area

The connection with Gaussian surface was discovered by [35]. In particular they prove:

Theorem ([35] Theorem 25)**.** *Let C be a collection of Borel sets in $\mathbb{R}^n$ each of which has Gaussian surface area at most s. Suppose furthermore that C is invariant under affine transformations. Then C is agnostically learnable with respect to any Gaussian distribution in time $n^{O(s^2/\epsilon^4)}$.*

Hence our results imply that:

Corollary 3. *The concept class of degree-d polynomial threshold functions is agnostically learnable with respect to any Gaussian distribution in time $n^{O(d^2/\epsilon^4)}$.*

1.6 Outline

Section 2 will be devoted to the proof of Theorem 1, Section 3 to the proof of Theorem 2, and Section 4 will provide some closing notes. The proof of Theorem 2 requires several technical results whose proofs are not very enlightening and are put in the Appendix so as not to interfere with the flow of the Chapter.

2 Proof of the Noise Sensitivity Bound

Proof of Theorem 1. Our basic strategy will be to use the symmetrization trick discussed in the Introduction to this Part. From these ideas, and a few facts about polynomials, the result will fall out.

We begin by letting $\theta = \arcsin(\sqrt{2\epsilon - \epsilon^2})$. We need to bound

$$p := \mathrm{GNS}_\epsilon(f) = \Pr(f(X) \neq f(\cos(\theta)X + \sin(\theta)Y)). \tag{1}$$

In general, we let

$$X_\phi = \cos(\phi)X + \sin(\phi)Y.$$

Hence, $\mathrm{GNS}_\epsilon(f) = \Pr(f(X_0) \neq f(X_\theta))$.

We note that the value of p given in Equation 1 remains the same if X and Y are replaced by any X' and Y' that are i.i.d. standard Gaussian distributions. Note that X_ϕ and $X_{\phi+\pi/2}$ are such i.i.d. Gaussians. Additionally, we note that for any ϕ

$$\begin{aligned}\cos(\theta)X_\phi + \sin(\theta)X_{\phi+\pi/2} =& \cos(\theta)\cos(\phi)X - \sin(\theta)\sin(\phi)X + \\ & \cos(\theta)\sin(\phi)Y + \sin(\theta)\cos(\phi)Y \\ =& \cos(\theta+\phi)X + \sin(\theta+\phi)Y \\ =& X_{\theta+\phi}.\end{aligned}$$

Therefore, we have that for any ϕ,

$$\mathrm{GNS}_\epsilon(f) = \Pr(f(X_\phi) \neq f(X_{\phi+\theta})).$$

Averaging over ϕ, we find that

$$\mathrm{GNS}_\epsilon(f) = \frac{1}{2\pi}\int_0^{2\pi} \Pr(f(X_\phi) \neq f(X_{\phi+\theta}))d\phi. \tag{2}$$

We define the random function $F : \mathbb{R} \to \{-1, 1\}$ by

$$F(\phi) = f(X_\phi).$$

(F depends on X and Y as well as ϕ). Hence

$$\begin{aligned}\mathrm{GNS}_\epsilon(f) &= \frac{1}{2\pi}\int_0^{2\pi} \Pr(F(\phi) \neq F(\phi+\theta))d\phi \\ &= \frac{1}{2\pi}\mathrm{E}_{X,Y}\left[\int_0^{2\pi} \mathbf{1}_{F(\phi)\neq F(\phi+\theta)}d\phi\right].\end{aligned}$$

Here $\mathbf{1}_{F(\phi)\neq F(\phi+\theta)}$ is the indicator function of the event that $F(\phi) \neq F(\phi+\theta)$. We may now consider the value of the integral above for fixed X and Y and then take the expectation. In such a case, F is some Boolean function. We note that $F(\phi)$ can only be different from $F(\phi+\theta)$ if F has a sign change in $[\phi, \phi+\theta]$ (with these values taken modulo 2π). This means that $F(\phi) \neq F(\phi+\theta)$ only on a union of intervals of length θ preceding each sign change of F. Therefore, the value of the above integral is at most θ times the number of sign changes of F on $[0, 2\pi)$. Hence,

$$\mathrm{GNS}_\epsilon(f) \leq \frac{\theta\mathrm{E}_{X,Y}[\text{number of sign changes of } F \text{ on } [0,2\pi)]}{2\pi}. \tag{3}$$

We now make use of the fact that f is a degree d polynomial threshold function. In particular, we will show that for any X and Y, F changes signs at most $2d$ times on $[0, 2\pi)$.

We let $f = \text{sgn}(g)$ for some degree-d polynomial g. We note that the number of sign changes of F is at most the number of zeroes of the function $g(\cos(\phi)X + \sin(\phi)Y)$ (unless this function is identically 0, in which case there are no sign changes). Note that $g(\cos(\phi)X + \sin(\phi)Y) = 0$ if and only if $z = e^{i\phi}$ is a root of the degree-$2d$ polynomial

$$z^d g\left(\left(\frac{z+z^{-1}}{2}\right)X + \left(\frac{z-z^{-1}}{2i}\right)Y\right).$$

Since such a polynomial can have at most $2d$ roots, the expectation in Equation 3 is at most $2d$.

As an alternative proof, note that sign changes of F correspond to joint solutions to the equations $g(aX + bY) = 0$ and $a^2 + b^2 = 1$. By Bezout's Theorem, there are at most $2d$ such roots.

Given this bound on the number of sign changes, we have that

$$\text{GNS}_\epsilon(f) \leq \frac{2d\theta}{2\pi} = \frac{d\theta}{\pi}$$

as desired.

We also note the ways in which the above bound can fail to be tight. Firstly, there may be some probability that F changes signs less than $2d$ times on a full circle. Secondly, the integral of $\mathbf{1}_{F(\phi)\neq F(\phi+\theta)}$ may be less than θ times number of times F changes signs if two of these sign changes are within θ of each other. On the other hand, if f is the threshold function for a product of d homogeneous linear functions, no two of which are scalar multiples of each other, the first case happens with probability 0, and the probability of the second goes to 0 as ϵ does. Therefore, for such functions our bound is asymptotically tight as $\epsilon \to 0$. □

3 Proof of the Gaussian Surface Area Bounds

The basic idea for our proof of Theorem 2 is to relate the Gaussian surface area of f to its noise sensitivity. In particular we claim that:

Lemma 4. *If f is a polynomial threshold function, and if X and Y are independent Gaussians then,*

$$\lim_{\epsilon\to 0} \frac{Pr(f(X) = -1 \text{ and } f(X + \epsilon Y) = 1)}{\epsilon} = \frac{\Gamma(f)}{\sqrt{2\pi}}. \tag{4}$$

The basic idea here is that the surface area determines how likely it is that X will lie within some distance of the boundary between the $+1$ and -1 regions of f. In particular, we expect that the probability that X is distance t from the boundary (and on the appropriate side) to be roughly $\Gamma(f)dt$.

Consider the above probability for X fixed. We may assume that X is close to the boundary (since otherwise the probability that $X + \epsilon Y$ lies on the other side is negligible). Since the boundary is smooth (at most points anyway) it is approximately linear at small scales. Therefore, the probability that $X + \epsilon Y$ is on the opposite side of the boundary is roughly the probability that the projection of ϵY onto the direction perpendicular to the boundary is more than the distance from X to the boundary. Since the size of this projection is just a Gaussian, the probability on the left hand side of Equation 4 is roughly

$$\int_{\epsilon y > t > 0} \phi(y)\Gamma(f)dtdy = \int_0^\infty \epsilon\Gamma(f)y\phi(y)dy = \frac{\Gamma(f)\epsilon}{\sqrt{2\pi}}.$$

The actual proof of this Lemma is technical and not terribly enlightening, and so is put off until the Appendix.

We note that the $\Pr(f(X) = -1 \text{ and } f(X + \epsilon Y) = 1)$ term in Lemma 4 is very nearly a noise sensitivity. Unfortunately, it fails to be for two reasons. Firstly, we require that $f(X) = -1$ and $f(X + \epsilon Y) = 1$ rather than merely asking that they are unequal. Secondly, X and $X + \epsilon Y$ are normalized differently. We solve the first of these problems by noting that the asymptotic probability that $f(X) = -1$ and $f(X + \epsilon Y) = 1$ should equal the asymptotic probability that $f(X) = 1$ and $f(X + \epsilon Y) = -1$ (which follows from Lemma 4 and Lemma 5 below). The second difficulty is overcome by showing that $f(X + \epsilon Y)$ is very likely equal to $f\left(\frac{X+\epsilon Y}{\sqrt{1+\epsilon^2}}\right)$, which is properly normalized. This follows from Lemma 6 below.

Lemma 5. *For f a polynomial threshold function, $\Gamma(f) = \Gamma(-f)$.*

The idea is that they both measure the surface area of the same boundary between $f^{-1}(1)$ and $f^{-1}(-1)$. We put off the proof until the Appendix.

Lemma 6. *If f is a degree d polynomial threshold function in n dimensions, $\epsilon > 0$ and X a random Gaussian variable, then*

$$Pr(f(X) \neq f(X(1+\epsilon))) \leq d\epsilon\sqrt{\frac{n}{4\pi}}.$$

Proof. Note that by conditioning on the line through the origin that X lies on, we may reduce this problem to the case of a one dimensional distribution. Note that f changes sign at most d times along this line. We need to bound the probability that at least one of these sign changes is in between X and $(1+\epsilon)X$. It therefore suffices to prove that for any one of these sign changes, that it lies between X and $(1+\epsilon)X$ with probability at most $\epsilon\sqrt{\frac{n}{4\pi}}$. Note that the probability that X is on the same side of the origin as this sign change is $\frac{1}{2}$. Beyond that, $|X|^2$ satisfies the χ^2 distribution with n degrees of freedom, namely $\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}dx$. Letting $y = \ln(x) = 2\ln(|X|)$ we find that that y has distribution

$$\frac{1}{2^{n/2}\Gamma(n/2)}e^{ny/2}e^{-e^y/2}dy.$$

We want the probability that y is within a particular interval of width $2\ln(1+\epsilon)$. This is at most 2ϵ times the maximum value of the density function. The maximum is achieved when $ny - e^y$ is maximal, or when $y = \ln(n)$. Then the density is

$$\begin{aligned}\frac{1}{2^{n/2}\Gamma(n/2)} n^{n/2} e^{-n/2} &= \sqrt{\frac{n}{4\pi}} \frac{(n/2)^{n/2} e^{-n/2} \sqrt{2\pi(n/2)}}{(n/2)!} \\ &\leq \sqrt{\frac{n}{4\pi}}.\end{aligned}$$

Multiplying this by d, 2ϵ and $\frac{1}{2}$ (the probability that X is on the same side of 0 as the sign change), we get our bound. □

Notice that this bound should be nearly sharp if the polynomial giving f is a product of terms of the form $|X|^2 - r_i$ for r_i approximately n and spaced apart by factors of $(1+\epsilon)^2$.

We can now prove a bound on a quantity more relevant to Gaussian surface area:

Corollary 7. *If f is an n dimensional, degree-d polynomial threshold function, $\epsilon > 0$ and X and Y independent Gaussians, then*

$$Pr(f(X) \neq f(X + \epsilon Y)) \leq \frac{d\epsilon}{\pi} + \frac{d\epsilon^2}{4}\sqrt{\frac{n}{\pi}}.$$

Proof. We let $r = \sqrt{1+\epsilon^2}$, $\theta = \arctan(\epsilon)$, and let $Z = \cos(\theta)X + \sin(\theta)Y$ be a normal random variable. Note that $X + \epsilon Y = rZ$. We then have that

$$\begin{aligned}&\Pr(f(X) \neq f(X+\epsilon Y)) \\ &\qquad \leq \Pr(f(X) \neq f(Z)) + \Pr(f(Z) \neq f(rZ)).\end{aligned}$$

By Theorem 1 and Lemma 6, this is at most

$$\frac{d\theta}{\pi} + d(r-1)\sqrt{\frac{n}{4\pi}} \leq \frac{d\epsilon}{\pi} + \frac{d\epsilon^2}{4}\sqrt{\frac{n}{\pi}}.$$

(since $\theta \leq \tan(\theta) = \epsilon$ and $\sqrt{1+\epsilon^2} \leq 1 + \epsilon^2/2$) □

We are now ready to prove Theorem 2.

Proof of Theorem 2. The result follows immediately from Lemma 4 (for both f and $-f$), Lemma 5, and Corollary 7 after noting that

$$\begin{aligned}\Pr(f(X) = -1, f(X+\epsilon Y) = 1) &\sim \frac{\epsilon\Gamma(f)}{\sqrt{2\pi}} = \frac{\epsilon\Gamma(-f)}{\sqrt{2\pi}} \\ &\sim \Pr(f(X) = 1, f(X+\epsilon Y) = -1)\end{aligned}$$

and thus

$$\Gamma(f) = \sqrt{\frac{\pi}{2}} \lim_{\epsilon \to 0} \frac{\Pr(f(X) \neq f(X + \epsilon Y))}{\epsilon} \leq \frac{d}{\sqrt{2\pi}}.$$

□

4 Conclusion

We have shown nearly tight bounds on the Gaussian surface area and noise sensitivity of polynomial threshold functions. One might hope to generalize these results to work for other distributions, such as the uniform distribution on vertices of the hypercube. Unfortunately, several aspects of this proof are difficult to generalize. Perhaps most significantly, we lose the symmetry that allowed us to prove our original result on noise sensitivity. Another difficulty would be in the relation between noise sensitivity and surface area. In the Gaussian case, the two are essentially equivalent quantities of study. On the other hand, [35] defined a notion of surface area for the hypercube distribution and proved that for even linear threshold functions there could be a gap between noise sensitivity and surface area of as much as $\Theta(\sqrt{\log(n)})$.

On the other hand, our results still provide hope for a proof of the full Gotsman-Linial Conjecture. For one, we have proved what can be considered to be an important special case of this Conjecture. Secondly, many of the results for average sensitivity and Bernoulli noise sensitivity come from generalizing results in the Gaussian setting, and hence there is hope that our results might be a starting point for proofs in these more difficult contexts.

Appendix

Here we prove some technical results about the Gaussian surface area for polynomial threshold functions. In the following, μ denotes the Gaussian measure. Throughout we consider p a non-zero degree d polynomial, with $f = \mathrm{sgn}(p)$.

Lemma 8. *Let p be a non-constant, degree-d polynomial. For $\epsilon > 0$, the probability that $|p(X)| < \epsilon$ for X a random Gaussian is $O(\epsilon^{1/d})$, where the implied constant depends on p.*

It should be noted that this is a rather weak anticoncentration result compared to say, the Theorem of Carbery and Wright as we do little to control the dependence of our constant on p. On the other hand, we present a proof for completeness and because the ideas used will be in the same spirit as in the more difficult proofs to follow.

Proof. The Lemma follows easily for a monic polynomial of one-variable. This is because p will only be small if x is near at least one of the roots of p. Our basic idea is to make a

change of variables and consider our polynomial as a function of just one variable to reduce to this case.

By making a generic orthogonal change of variables, we may assume that the leading coefficient of p in terms of x_1 is a constant. By rescaling, we may assume that p is a degree d monic polynomial in $x = x_1$ whose coefficients are polynomials in $y = (x_2, \ldots, x_n)$ (rescaling changes ϵ by a factor that depends on p but nothing else). We have that $p(x) = (x - r_1(y))(x - r_2(y)) \cdots (x - r_d(y))$ (here $r_i(y)$ are the complex roots of $p(x, y)$ thought of as a polynomial in x for fixed y). Therefore, if $|p(x, y)| \leq \epsilon$, then x must be within $\epsilon^{1/d}$ of at least one of the $r_i(y)$ (i.e. $|x - r_i|$ must be less than $\epsilon^{1/d}$ for some i). Hence after fixing y,

$$\mu\left(\{x : |p(x, y)| \leq \epsilon\}\right) = O(\epsilon^{1/d}).$$

Integrating with respect to y proves the Lemma. □

Lemma 9. *Let p be a non-zero, degree-d polynomial in n variables such that p is not divisible by the square of any polynomial. Let $S = \{x : p(x) = 0\}$. For $\epsilon > \delta > 0$, let $S^\epsilon = \{x \in S : |\nabla p(x)| < \epsilon\}$. Then*

$$\mu((S^\epsilon)_\delta) = O(\epsilon^{1/d^2}\delta).$$

Here $(S^\epsilon)_\delta$ is the set of points within δ of some point of S^ϵ (again the asymptotic constant here and those throughout the proof depend on p).

Proof. This proof is considerably more difficult than the previous one. The basic idea is as follows. First, ignoring the condition on ϵ, we would like to show that $\mu(S_\delta) = O(\delta)$. Consider the set of points x' within δ of some point x with $p(x) = 0$ and $\frac{\partial p(x)}{\partial x_1}$ reasonably large. For such x', it follows that $\left|\frac{\partial p(x')}{\partial x_1}/p(x')\right| = \Omega(\delta^{-1})$. Considering p as a function of x_1 (whose coefficients are polynomials in the other x_i), it is not hard to see that the probability of this happening is $O(\delta)$ (since if the derivative is constant sized, you leave the region where p is small after $O(\delta)$ distance). If $\frac{\partial p}{\partial x_1}$ is small, but $\frac{\partial^2 p}{\partial^2 x_1}$ is reasonably large we can apply the same trick with $\frac{\partial p}{\partial x_1}$ instead of p. In general, we split S into subsets $S_{k,i}$ where k indexes the first degree of derivative for which p is not small at x, and i indexes a direction at which the k^{th} derivative is reasonably large and apply the above ideas to an appropriate derivative of p.

We deal with ϵ by noting that if both p and p' are small, then their resolvent polynomial is also small, and use Lemma 8 to get the ϵ dependence.

We begin by partitioning S^ϵ into parts. First we pick a large constant C (we will specify how large later). For $k \leq d$ and $x \in \mathbb{R}^n$ we let $|p^{(k)}(x)|$ be the sum of the absolute values of the k^{th} order mixed partial derivatives of p at x. We let S^ϵ_k be the set of $x \in S^\epsilon$ so that $(C\delta)^k|p^{(k)}(x)| \geq (C\delta)^i|p^{(i)}(x)|$ for all $1 \leq i \leq d$.

We pick some finite number of orthonormal bases, $B_i = (x_{1,i}, \ldots, x_{n,i})$ so that:

1. When written in basis B_i, p has non-zero x_1^d coefficient.

2. For each k there exists a $C_k > 0$ so that for any x, $|p^{(k)}(x)|$ is at most $C_k \max_i \left|\frac{\partial^k p(x)}{\partial x_{1,i}^k}\right|$.

The first condition holds as long as the B_i are chosen generically. The second condition will hold as long as the B_i are chosen generically and there are sufficiently many of them. We claim that if sufficiently many B_i are chosen generically then there is no non-zero degree-k polynomial, q, so that $\frac{\partial^k q(0)}{\partial x_{1,i}^k} = 0$ for all i. Once we have that $\frac{\partial^k q(0)}{\partial x_{1,i}^k}$ is non-zero for all q, we claim that it is enough to find a C_k so that $|q^{(k)}(0)| \leq C_k \max_i \left|\frac{\partial^k q(0)}{\partial x_{1,i}^k}\right|$ for q an arbitrary normalized (for example by making the sum of squares of coefficients equal to 1) homogeneous degree k polynomial. This is because for any x, we can use this result with q equal to the normalized version of the degree k part of the Taylor expansion of p about x, recentered around 0, to show that $|p^{(k)}(x)| \leq C_k \max_i \left|\frac{\partial^k p(x)}{\partial x_{1,i}^k}\right|$. For normalized, homogeneous, q of degree k we have that $\max_i \left|\frac{\partial^k q(0)}{\partial x_{1,i}^k}\right|$ is non-zero. Hence for any value of q, there is some value of C_k that works for q in some open neighborhood of that value. Hence, since the set of normalized degree k polynomials is compact, there is a C_k that works for all such polynomials. Note that the C_k used here will likely depend on n.

Having picked the B_i, we ensure that $C > 3C_k$ for each k. We then further partition S^ϵ by letting $S^\epsilon_{k,i}$ be the subset of S^ϵ_k so that $|p^{(k)}(x)| \leq C_k \left|\frac{\partial^k p(x)}{\partial x_{1,i}^k}\right|$. We will show that

$$\mu\left(\left(S^\epsilon_{k,i}\right)_\delta\right) = O(\delta\epsilon^{1/d^2})$$

for each i, k. In what follows, we consider this problem for fixed i and k.

We use the coordinates $x = x_{1,i}$ and $y = (x_{2,i}, \ldots, x_{n,i})$. We rescale p so that it is a monic, degree-d polynomial in x with coefficients that are polynomials in y. Consider $(x, y) \in S^\epsilon_{k,i}$. We have that

$$C_k \left|\frac{\partial^k p(x,y)}{\partial x^k}\right| \geq |p^{(k)}(x,y)| \geq (C\delta)^{j-k}|p^{(j)}(x,y)|.$$

(the inequality on the left holds because $(x, y) \in S^\epsilon_{k,i}$ and the one on the right holds because $(x, y) \in S^\epsilon_k$.) In particular, if (x', y') is within δ of (x, y) then by Taylor's Theorem:

$$\begin{aligned}
\left|\frac{\partial^k p(x',y')}{\partial x^k} - \frac{\partial^k p(x,y)}{\partial x^k}\right| &\leq \sum_{j=1}^{d-k} \frac{\delta^j}{j!}|p^{(k+j)}(x,y)| \\
&\leq \left(C_k \sum_{j=1}^{d-k} \frac{C^{-j}}{j!}\right)\left|\frac{\partial^k p(x,y)}{\partial x^k}\right| \\
&\leq \frac{1}{2}\left|\frac{\partial^k p(x,y)}{\partial x^k}\right|.
\end{aligned}$$

Hence $\left|\frac{\partial^k p(x',y')}{\partial x^k}\right| \geq \frac{1}{2}\left|\frac{\partial^k p(x,y)}{\partial x^k}\right|$.

Similarly, we have that

$$\begin{aligned}\left|\frac{\partial^{k-1} p(x',y')}{\partial x^{k-1}}\right| &\leq \sum_{j=0}^{d-k+1} \frac{\delta^j}{j!}|p^{(k-1+j)}(x,y)| \\ &\leq C\delta \sum_{j=0}^{d-k+1} \frac{C^{-j}}{j!}\left|\frac{\partial^k p(x,y)}{\partial x^k}\right| \\ &\leq 2C\delta\left|\frac{\partial^k p(x,y)}{\partial x^k}\right|.\end{aligned}$$

Therefore for $(x',y') \in (S^\epsilon_{k,i})_\delta$ we have that

$$\left|\frac{\left(\frac{\partial^k p(x',y')}{\partial x'^k}\right)}{\left(\frac{\partial^{k-1} p(x',y')}{\partial x'^{k-1}}\right)}\right| \geq \frac{1}{4C\delta}.$$

Letting $g(x,y) = \frac{\partial^{k-1} p(x,y)}{\partial x^{k-1}}$, we have that $\left|\frac{g'(x')}{g(x')}\right| \geq \frac{1}{4C\delta}$. On the other hand if we let $g(x,y) = a_k(x-r_1(y))\cdots(x-r_{d-k+1}(y))$, we have that $\frac{g'(x)}{g(x)} = \frac{1}{x-r_1(y)} + \cdots + \frac{1}{x-r_{d-k+1}(y)}$. Hence for the above to hold, x must be within $4Cd\delta$ of at least one of the $r_j(y)$. Therefore, the measure of the intersection of $(S^\epsilon_{k,i})_\delta$ with the line defined by fixing the value of y is at most $4Cd^2\delta$.

Note that if $(x,y) \in S^\epsilon$ and if (x',y') is within δ of (x,y) that $|p(x',y')| \leq \epsilon\delta + O(r^d\delta^2)$, and $|p'(x',y')| \leq \epsilon + O(\delta r^d)$, where $r = |(x,y)|+1$. Considering $p(x,y) = (x-r_1(y))\cdots(x-r_d(y))$ again as a function of x consider its discriminant $R(y)$. $R(y)$ is a non-zero polynomial of degree at most $d(d-1)$ in y. Furthermore, $R(y)$ can be written as $pq_1 + p'q_2$ for some polynomials q_1 and q_2 of degrees $d(d-2)$ and $d(d-2)+1$. Therefore, if $(x',y') \in (S^\epsilon)_\delta$, then $|R(y)| \leq \epsilon r^{d^2}$.

Now the measure of the set of points with $r > \epsilon^{1/d^3}$ is small enough to ignore. Throwing away this region, we find that if we project $\left(S^\epsilon_{k,i}\right)_\delta$ onto its y-coordinate, it covers only points where $|R(y)| \leq \epsilon^{1+1/d}$, which by Lemma 8 has measure at most $O(\epsilon^{(d+1)/(d^2(d-1))}) = O(\epsilon^{1/d^2})$. Furthermore, by the above, once we fix such a y the measure over x of this set is $O(\delta)$. Therefore, for each k, i, $\mu((S^\epsilon_{k,i})_\delta) = O(\epsilon^{1/d^2}\delta)$. Hence, summing over all k and i,

$$\mu((S^\epsilon)_\delta) = O(\epsilon^{1/d^2}\delta).$$

□

Proposition 10. *Let p be a non-zero polynomial. Let $A = \{x : p(x) < 0\}$. Let $S = \partial\bar{A}$. Then*

$$\Gamma(A) = \lim_{\delta\to 0}\frac{\mu(A_\delta \backslash A)}{\delta} = \lim_{\delta\to 0}\frac{\mu(S_\delta)}{2\delta} = \int_S \phi(x)d\sigma(x).$$

Proof. The idea here is that we can use Lemma 9 to bound the contribution to the volume from x near singular points of S. Away from these points, we can parameterize $A_\delta \backslash A$ by $S \times [0, \delta]$ using $x \to$ (nearest point of S to $x, d(x, S)$). Avoiding the singular set, it is easy to see that this converges to our desired integral.

First, we note that we can remove any square factors from p (as removing them only changes $A_\delta \setminus A$ by a set of measure 0 and does not change S), and thus assume that p is a square-free polynomial. By Lemma 9, when calculating $\mu(A_\delta \backslash A)$ in the computation of $\Gamma(A)$, we may ignore points in $(S^{\delta^{1/6}})_\delta$. We may also ignore points with absolute value more than $r = \log(\delta^{-1})$ since the total measure of the set of such points goes rapidly to 0 as $\delta \to 0$. Let $S_0 = S \backslash S^{\delta^{1/6}}$.

We begin with the following claim. Suppose that $|a| \leq r$ and that b is the closest point in S to a (such a point exists since S is closed and non-empty). Suppose also that $d(a, b) < \delta$ and that $b \in S_0$. Then for δ sufficiently small, b is the only point of S within δ of a for which $d(a, b')$ obtains a local minimum.

To prove this claim, we may begin by assuming that $b = 0$ and p a polynomial of degree d with coefficients of size at most $O(r^d)$. Let us pick a coordinate system where we describe points as (x, y) where x is the distance from the origin in the direction of $p'(0)$, and y is a coordinate in the orthogonal plane. Hence $p'(0) = (t, 0)$ for some $t > \delta^{1/6}$. Furthermore, since b attains a local minimum on S of distance to a, it must be the case that $a = (s, 0)$ for some $\delta > s > 0$.

We have that for $c = (x, y)$ within δ of a that $p(c) = tx + O(r^d|c|^2)$ and that $p'(c) = (t, 0) + O(r^d|c|)$. If such a point were to be a local minimum of distance to a on S we would need that $p(c) = 0$ and $p'(c)$ is parallel to $a - c$. The first implies that $x = O(r^d \delta^{-1/6} y^2)$. The latter implies that

$$\frac{O(r^d|y|)}{|y|} = \frac{t + O(r^d|y|)}{s}.$$

But this is impossible since $t \gg sr^d + |y|r^d$.

For $x \in S_0$ let $n(x) = \frac{p'(x)}{|p'(x)|}$ be the outward pointing normal vector to x. The above claim along with Lemma 9 says that $\mu(A_\delta \backslash A)$ is $o(\delta)$ plus the volume of the region parameterized by $x + tn(x)$ for $x \in S_0, t \in [0, \delta]$. Hence, we have that

$$\mu(A_\delta \backslash A) = \int_{S_0} \int_0^\delta \phi(x + tn(x)) |\det(J(x, t))| dt d\sigma(x) + o(\delta)$$

where $J(x, t)$ is the Jacobian of our parametrization $(x, t) \to x + tn(x)$. We first claim that $J(x, t)$ is $O(r^{3d}\delta^{1/2})$ plus the matrix sending a tangent vector to S to itself and sending dt to $n(x)$. This is because the t derivative is exactly $n(x)$, and the derivative with respect to x is $I + t\frac{\partial n}{\partial x}$. But $|t| \leq \delta$ and $\frac{\partial n}{\partial x}$ is a degree $3d$ polynomial divided by $|p'|^3$, and

therefore is $O(r^{3d}\delta^{-1/2})$. Hence, multiplying by t, we get an error of size $O(r^{3d}\delta^{1/2})$. Hence $|\det(J(x,t))| = 1 + O(r^{3d}\delta^{1/2})$. Furthermore, $\phi(x+tn) = \phi(x)(1+O(r\delta))$. Hence, we have that:

$$\begin{aligned}\mu(A_\delta \backslash A) &= \int_{S_0}\int_0^\delta \phi(x)(1+O(r^{3d}\delta^{1/2}))dtd\sigma(x) + o(\delta) \\ &= \int_{S_0}\int_0^\delta \phi(x)dtd\sigma(x)(1+O(r^{3d}\delta^{1/2})) + o(\delta) \\ &= \delta\int_{S_0}\phi(x)d\sigma(x) + o(\delta).\end{aligned}$$

Finally noting that

$$\lim_{\delta\to 0}\int_{S_0}\phi(x)d\sigma(x) = \int_S \phi(x)d\sigma(x)$$

proves the first two equalities.

The last equality follows from this after noting that if $B = \{x : p(x) > 0\}$, then

$$S_\delta = (A_\delta\backslash\bar{A}) \cup (B_\delta\backslash\bar{B}).$$

□

Lemma 5 now follows immediately

Proof of Lemma 5. Let $f = \operatorname{sgn}(p(x))$. Let $A = \{x : p(x) > 0\}, A' = \{x : p(x) < 0\}$ and $S = \partial\bar{A} = \bar{A}\cap\bar{A}' = \partial\bar{A}'$. Then by Proposition 10 we have that

$$\begin{aligned}\Gamma(f) &= \lim_{\delta\to 0}\frac{\mu(A_\delta\backslash A)}{\delta} \\ &= \int_S \phi(x)d\sigma(x) \\ &= \lim_{\delta\to 0}\frac{\mu(A'_\delta\backslash A')}{\delta} \\ &= \Gamma(-f).\end{aligned}$$

□

Proof of Lemma 4. Again we use Lemma 9 to show we can afford to ignore the case where X is near singular points of p. We use a parametrization of X near the zero set of p similar to that in the proof of Proposition 10, and derive a rigorous version of the intuition we had before.

Recall that if $A = f^{-1}(1)$, then

$$\Gamma(f) = \lim_{\epsilon \to 0} \frac{\mu(A_\epsilon \backslash A)}{\epsilon}.$$

Next note that since the probability that $|\epsilon Y| > \epsilon^{1-1/12d^2}$ goes rapidly to 0 as $\epsilon \to 0$, we can throw away all cases where X is not within $\epsilon^{1-1/12d^2}$ of $S = \partial A$ from the left hand side of Equation 4.

Define S_0 as in the proof of Proposition 10. By Lemma 9 we have that X is within $\epsilon^{1-1/12d^2}$ of $(S \backslash S_0)$ with probability at most $O(\epsilon^{1+1/12d^2})$ and so we may throw these events away. Hence we may restrict ourselves to considering only X whose nearest neighbor in S lies in S_0 and is at distance at most $\epsilon^{1-1/12d^2}$. We may also assume that $|X| < \log(\epsilon^{-1})$.

Letting n be the derivative of p at the point of S nearest to X, we have that $p(X+\epsilon Y) = p(X) + \epsilon n \cdot Y + O(r^d \epsilon^2 Y^2)$. For Y in the range we are considering, $O(r^d \epsilon^2 Y^2) = O(r^d \epsilon^{3/2})$. Hence $p(X) = d(X,S)|n| + O(\epsilon^{3/2})$. Hence $p(X+\epsilon Y) = |n| \left(d(X,S) + \epsilon \left(\frac{n}{|n|} \right) \cdot Y \right) + O(\epsilon^{3/2})$. Note that up to $O(\epsilon^{3/2})$, the sign of this only depends on the dot product of Y with a unit vector and on $d(X,S)$. On the other hand since $|n|\epsilon \gg \epsilon^{3/2}$, the probability that $p(X+\epsilon Y)$ is negative is within $\epsilon^{1/3}$ of the probability that a random one dimensional normal is more than $\frac{d(X,S)}{\epsilon}$.

Hence

$$\begin{aligned}
&\Pr(f(X) = -1 \text{ and } f(X+\epsilon Y) = 1) \\
=&o(\epsilon) + O(\epsilon^{1/3}) \Pr(d(X,S) \leq \epsilon^{1-1/12d^2}) \\
&+ \Pr(f(X) = -1 \text{ and } d(X,S) < \epsilon N) \\
=&o(\epsilon) + \Pr(f(X) = -1 \text{ and } d(X,S) < \epsilon N).
\end{aligned}$$

where N is a random normal variable.

After fixing the value of N this probability is just $\mu(A_{\epsilon N} \backslash A)$. Hence integrating we get

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mu(A_{\epsilon N} \backslash A) dx \\
&\quad = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Gamma(f) \epsilon x (1 + o(1)) dx \\
&\quad = \frac{\Gamma(f)\epsilon}{\sqrt{2\pi}} + o(\epsilon).
\end{aligned}$$

This completes our proof. □

Chapter H

A PRG for Polynomial Threshold Functions

1 Introduction

We discuss the issue of pseudorandom generators for polynomial threshold functions of bounded degree. Namely for some known probability distribution $\mathcal{D}$ on $\mathbb{R}^n$, we would like to find an explicit, easily computable function $G : \{0,1\}^S \to \mathbb{R}^n$ so that for any degree-d polynomial threshold function, f,

$$\left|\mathrm{E}_{Y\sim\mathcal{D}}[f(Y)] - \mathrm{E}_{X\in_u\{0,1\}^S}[f(G(X))]\right| < \epsilon.$$

This problem could be studied for $\mathcal{D}$ any probability distribution, though most work has been done with either the Bernoulli or Gaussian case. In this Chapter, we construct an explicit PRG for the Gaussian case. In particular, for any real numbers $c, \epsilon > 0$ and integer $d > 0$ we construct a PRG that fools degree-d PTFs of Gaussians to within ϵ and has seed length $\log(n)2^{O_c(d)}\epsilon^{-4-c}$. In particular we show that

Theorem 1. *Let $c > 0$ be a constant. For $\epsilon > 0$, and d a positive integer, let $N = 2^{\Omega_c(d)}\epsilon^{-4-c}$ and $k = \Omega_c(d)$ be integers. Let X_i $1 \le i \le N$ be independently sampled from k-independent families of standard Gaussians. Let $X = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i$. Let Y be a fully independent family of standard Gaussians. Then for any degree-d polynomial threshold function, f,*

$$|E[f(X)] - E[f(Y)]| < \epsilon.$$

Note that the constants in the Ω_c's will get large as $c \to 0$. From this we can construct an efficient PRG for PTFs of Gaussians. In particular

Corollary 2. *For every $c > 0$, there exists a PRG that ϵ-fools degree-d PTFs of Gaussians with seed length*

$$\log(n)2^{O_c(d)}\epsilon^{-4-c}.$$

Much of the previous work in constructing pseudo-random generators involves the use of functions of limited independence. It was shown in [11] that $\tilde{O}(\epsilon^{-2})$-independence fools degree-1 PTFs. The degree-2 case was later dealt with in in [13], in which it was shown that $\tilde{O}(\epsilon^{-9})$-independence sufficed (and that $O(\epsilon^{-8})$ sufficed for Gaussians). The author showed that limited independence suffices to fool arbitrary degree PTFs of Gaussians (Chapter F), but the amount of independence required was $O_d(\epsilon^{-2^{O(d)}})$. In terms of PRGs that do not rely solely on limited independence, [45] found a PRG for degree-d PTFs on the hypercube distribution of size $\log(n)2^{O(d)}\epsilon^{-8d-3}$. Hence for d more than constant sized, and ϵ less than some constant, our PRG will always beat the other known examples.

The basic idea of the proof of Theorem 1 will be to show that X fools a function g which is a smooth approximation of f. This is done using the replacement method. In particular, we replace the X_i by fully independent families of Gaussians one at a time and show that at each step a small error is introduced. This is done by replacing g by its Taylor expansion and noting that small degree moments of the X_i are identical to the corresponding moments of a fully independent family.

Naively, if $f = \text{sgn}(p(x))$, we might try to let $g = \rho(p(x))$ for some smooth function ρ so that $\rho(x) = \text{sgn}(x)$ for $|x| > \delta$. If we Taylor expand g to order $T-1$, we find that the error in replacing X_i by a fully random Gaussian is roughly the size of the T^{th} derivative of g times the T^{th} moment of $p(X) - p(X')$, where X' is the new random variable we get after replacing X_i. We expect the former to be roughly δ^{-T} and the latter to be roughly $|p|_2^T N^{-T/2}$. Hence, for this to work, we will need $N \gg (|p|_2\delta^{-1})^2$. On the other hand, for g to be a good approximation of f, we will need that the probability that $|p(Y)| < \delta$ to be small. Using standard anticoncentration bounds, this requires $|p|_2\delta^{-1}$ to be roughly ϵ^{-d}, and hence N will be required to be at least ϵ^{-2d}.

To fix this, we instead use the idea of strong anticoncentration discussed in the Introduction to this Part. Heuristically, with probability roughly $1-\epsilon$ it should hold that

$$|p(X)| \geq \epsilon|p'(X)| \geq \epsilon^2|p''(X)| \geq \ldots \geq \epsilon^d|p^{(d)}(X)|. \tag{1}$$

In our analysis we will use a g that is a function not only of $p(X)$, but also of $|p^{(m)}(X)|$ for $1 \leq m \leq d$. Instead of forcing g to be a good approximation to f whenever $|p(X)| \geq \epsilon^d$, we will only require it to be a good approximation to f at X where Equation 1 holds. This gives us significant leeway, since although the derivative to g with respect to $p(X)$ will still be large at places, this will only happen when $|p'(X)|$ is small, and this in turn will imply that the variance in $p(X)$ achieved by replacing X_i is comparably small.

In Section 2, we will review some basic properties of polynomial threshold functions. In Section 3, we in introduce the notion of the derivative (which we call the noisy derivative) that will be useful for our purposes. We then prove a number of Lemmas about this derivative and in particular prove a rigorous version of Equation 1. In Section 4, we discuss some averaging operators that will be useful in analyzing what happens when one of the X_i is

changed. In Section 5, we use these results and the above ideas to prove Theorem 1. In Section 6, we use this result to prove Corollary 2.

2 Definitions and Basic Properties

We are concerned with polynomial threshold functions, so for completeness we give a definition

Definition. *A degree-d polynomial threshold function (PTF) is a function of the form*

$$f = sgn(p(x))$$

where p is a polynomial of degree at most d.

We are also concerned with the idea of fooling functions so we define

Definition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ we say that a random variable X with values in $\mathbb{R}^n$ ϵ-fools f if*

$$|E[f(X)] - E[f(Y)]| \leq \epsilon$$

where Y is a standard n-dimensional Gaussian.

For convenience we define the notation:

Definition. *We use*

$$A \approx_\epsilon B$$

to denote that

$$|A - B| = O(\epsilon).$$

For a function on $\mathbb{R}^n$ we define its L^k norm by

Definition. *For $p : \mathbb{R}^n \to \mathbb{R}$ define*

$$|p|_k = E_X[|p(X)|^k]^{1/k}.$$

Where the above expectation is over X a standard n-dimensional Gaussian.

We will make use of the hypercontractive inequality. The proof follows from Theorem 2 of [49].

Lemma 3. *If p is a degree-d polynomial and $t > 2$, then*

$$|p|_t \leq \sqrt{t-1}^d |p|_2.$$

In particular this implies the following Corollary:

Corollary 4. *Let p be a degree-d polynomial in n variables. Let X be a family of standard Gaussians. Then*

$$Pr(|p(X)| \geq |p|_2/2) \geq 9^{-d}/2.$$

Proof. This follows immediately from the PaleyZygmund inequality applied to p^2. □

We also obtain:

Corollary 5. *Let p be a degree-d polynomial, $t \geq 1$ a real number, then*

$$|p|_t \leq 2^{O_t(d)}|p|_1$$

Proof. By Lemma 3, it suffices to prove this for $t = 2$. This in turn follows from Corollary 4, which implies that $|p|_1 \geq 9^{-d}/4|p|_2$. □

And

Corollary 6. *If $p(X,Y), q(X,Y)$ are degree-d polynomials in standard Gaussians X and Y then*

$$Pr_Y(|p(X,Y)|_{2,X} < \epsilon|q(X,Y)|_{2,X}) \leq 4 \cdot 9^d Pr_{X,Y}(|p(X,Y)| < 4 \cdot 3^d\epsilon|q(X,Y)|).$$

Where $|r(X,Y)|_{2,X}$ denotes the L^2 norm over X, namely $(E_X[r(X,Y)^2])^{1/2}$.

Proof. Given Y so that $|p(X,Y)|_{2,X} < \epsilon|q(X,Y)|_{2,X}$, by Corollary 4 we have that $|q(X,Y)| \geq |q(X,Y)|_{2,X}/2$ with probability at least $9^{-d}/2$. Furthermore, $|p(X,Y)| \leq 2 \cdot 3^d|p(X,Y)|_{2,X}$ with probability at least $1 - 9^{-d}/4$. Hence with probability at least $9^{-d}/4$ we have that

$$|p(X,Y)| \leq 2 \cdot 3^d|p(X,Y)|_{2,X} < 2 \cdot 3^d\epsilon|q(X,Y)|_{2,X} \leq 4 \cdot 3^d\epsilon|q(X,Y)|.$$

So

$$\Pr_{X,Y}(|p(X,Y)| < 4 \cdot 3^d\epsilon|q(X,Y)|) \geq \frac{9^{-d}}{4}\Pr_Y(|p(X,Y)|_{2,X} < \epsilon|q(X,Y)|_{2,X}).$$

□

3 Noisy Derivatives

In this section we define our notion of the noisy derivative and obtain some of its basic properties.

Definition. *Let X and Y be n-dimensional vectors and θ a real number. Then we let*

$$N_Y^\theta(X) := \cos(\theta)X + \sin(\theta)Y.$$

If X and Y are independent Gaussians, Y can be thought of as a noisy version of X with θ a noise parameter. We next define the noisy derivative.

Definition. *Let X, Y, Z be n-dimensional vectors, $f : \mathbb{R}^n \to \mathbb{R}$ a function, and θ a real number. We define the* noisy derivative of f at X with parameter θ in directions Y and Z *to be*

$$D_{Y,Z}^\theta f(X) := \frac{f(N_Y^\theta(X)) - f(N_Z^\theta(X))}{\theta}.$$

It should be noted that if X, Y, Z are held constant and θ goes to 0, the noisy derivative approaches the difference of the directional derivatives of f at X in the directions Y and Z. The noisy derivative for positive θ can be thought of as sort of a large scale derivative that covers slightly more than just a differential distance.

We also require a notion of the average size of the ℓ^{th} derivative. In particular we define:

Definition. *For X an n-dimensional vector, $f : \mathbb{R}^n \to \mathbb{R}$ a function, ℓ a non-negative integer, and θ a real number, we define*

$$|f_\theta^{(\ell)}(X)|_2^2 := E_{Y_1,Z_1,\ldots,Y_\ell,Z_\ell}[|D_{Y_1,Z_1}^\theta D_{Y_2,Z_2}^\theta \cdots D_{Y_\ell,Z_\ell}^\theta f(X)|^2],$$

where the expectation is taken over independent, standard Gaussians Y_i, Z_i.

Lemma 7. *For p a degree-d polynomial, and θ a real number*

$$|p_\theta^{(d)}(X)|_2$$

is independent of X.

Proof. This follows from the fact that for fixed Y_i, Z_i that

$$D_{Y_1,Z_1}^\theta D_{Y_2,Z_2}^\theta \cdots D_{Y_\ell,Z_\ell}^\theta p(X)$$

is independent of X. This in turn follows from the fact that for any degree-d polynomial q and any Y,Z, $D_{Y,Z}^\theta q(X)$ is a degree-$(d-1)$ polynomial in X. □

We now prove our version of the statement that the value of a polynomial is probably not too much smaller than its derivative.

Proposition 8. *Let $\epsilon, \theta > 0$ be real numbers with $\theta = O(\epsilon)$. Let p be a degree-d polynomial and let X, Y, Z be standard independent Gaussians. Then*

$$Pr_{X,Y,Z}(|p(X)| < \epsilon |D^{\theta}_{Y,Z} p(X)|) = O(d^2 \epsilon).$$

To prove this we use the following Lemma:

Lemma 9. *Let $\epsilon, \theta, p, d, X, Y$ be as above. Then*

$$Pr_{X,Y}\left(|p(X)| < \frac{\epsilon}{\theta}|p(X) - p(N^{\theta}_{Y}(X))|\right) = O(d^2\epsilon).$$

Proof. The basic idea of the proof will be the averaging argument discussed in the Introduction to this Part. We note that the above probability should be the same for any independent Gaussians, X and Y. We let $X_\phi := N^{\phi}_{Y}(X)$. Note that X_ϕ and $X_{\phi+\pi/2}$ are independent of each other. Furthermore, $N^{\theta}_{X_{\phi+\pi/2}}(X_\phi) = X_{\phi+\theta}$. Hence for any ϕ, the above probability equals

$$\Pr_{X,Y}\left(|p(X_\phi)| < \frac{\epsilon}{\theta}|p(X_\phi) - p(X_{\phi+\theta})|\right).$$

We claim that for any values of X and Y, that the average value over $\phi \in [0, 2\pi]$ of the above is $O(d^2\epsilon)$.

Notice that with X and Y fixed, $p(X_\phi)$ is a degree-d polynomial in $\cos(\phi)$ and $\sin(\phi)$. Letting $z = e^{i\phi}$ we have that $\cos(\phi) = \frac{z+z^{-1}}{2}, \sin(\phi) = \frac{z-z^{-1}}{2i}$. Hence $p(X_\phi) = z^{-d}q(z)$ for some polynomial q of degree at most $2d$.

We wish to bound the probability that

$$\frac{\theta}{\epsilon} < \frac{|z^{-d}q(z) - z^{-d}e^{-id\theta}q(ze^{i\theta})|}{|z^{-d}q(z)|} = \frac{|q(z) - e^{-id\theta}q(ze^{i\theta})|}{|q(z)|}.$$

For $\frac{\theta}{\epsilon}$ sufficiently small, we may instead bound the probability that

$$\left|\log\left(\frac{e^{-id\theta}q(ze^{i\theta})}{q(z)}\right)\right| > \frac{\theta}{2\epsilon}.$$

On the other hand, we may factor $q(z)$ as $a\prod_{i=1}^{2d}(z - r_i)$ where r_i are the roots of q. The left hand side of the above is then at most

$$\begin{aligned} d\theta + \sum_{i=1}^{2d}\left|\log\left(\frac{ze^{i\theta} - r_i}{z - r_i}\right)\right| &\leq d\theta + \sum_{i=1}^{d} O\left(\left|\frac{ze^{i\theta} - z}{z - r_i}\right|\right) \\ &\leq d\theta + \theta\sum_{i=1}^{2d} O\left(\frac{1}{|z - r_i|}\right). \end{aligned}$$

Hence it suffices to bound the probability that

$$d + \sum_{i=1}^{2d} O\left(\frac{1}{|z - r_i|}\right) > \frac{1}{2\epsilon}.$$

If $4\epsilon > d^{-1}$, there is nothing to prove. Otherwise, the above holds only if

$$\sum_{i=1}^{2d} O\left(\frac{1}{|z - r_i|}\right) > \frac{1}{4\epsilon}.$$

This in turn only occurs when z is within $O(d\epsilon)$ of some r_i. For each r_i this happens with probability $O(d\epsilon)$ over ϕ, and hence by the union bound, the above holds with probability $O(d^2\epsilon)$. □

Note that a tighter analysis could be used to prove the bound $O(d\log(d)\epsilon)$, but we will not need this stronger result.

Proposition 8 now follows immediately by noting that $|p(X)|$ is less than $\epsilon|D^{\theta}_{Y,Z}p(X)|$ only when either $|p(X)|/2 < \frac{\epsilon}{\theta}|p(X) - p(N^{\theta}_{Y}(X))|$ or $|p(X)|/2 < \frac{\epsilon}{\theta}|p(X) - p(N^{\theta}_{Z}(X))|$. This allows us to prove our version of Equation 1.

Corollary 10. *For p a degree-d polynomial, X a standard Gaussian, $\epsilon, \theta > 0$ with $\theta = O(\epsilon)$, and ℓ a non-negative integer,*

$$Pr_X(|p^{(\ell)}_{\theta}(X)|_2 \leq \epsilon|p^{(\ell+1)}_{\theta}(X)|_2) \leq 2^{O(d)}\epsilon.$$

Proof. Let Y_i, Z_i be standard Gaussians independent of each other and of X for $1 \leq i \leq \ell+1$. Applying Proposition 8 to $D^{\theta}_{Y_2,Z_2} \cdots D^{\theta}_{Y_{\ell+1},Z_{\ell+1}}p(X)$, we find that

$$\begin{aligned}\Pr_{X,Y_1,Z_1}\left(|D^{\theta}_{Y_2,Z_2} \cdots D^{\theta}_{Y_{\ell+1},Z_{\ell+1}}p(X)| \leq 4 \cdot 3^d\epsilon|D^{\theta}_{Y_1,Z_1} \cdots D^{\theta}_{Y_{\ell+1},Z_{\ell+1}}p(X)|\right) \\ = O(d^2 3^d\epsilon).\end{aligned}$$

Noting that

$$|p^{(\ell)}_{\theta}(X)|_2 = |D^{\theta}_{Y_2,Z_2} \cdots D^{\theta}_{Y_{\ell+1},Z_{\ell+1}}p(X)|_{2,(Y_1,Z_1,\ldots,Y_{\ell+1},Z_{\ell+1})}$$

and

$$|p^{(\ell+1)}_{\theta}(X)|_2 = |D^{\theta}_{Y_1,Z_1} \cdots D^{\theta}_{Y_{\ell+1},Z_{\ell+1}}p(X)|_{2,(Y_1,Z_1,\ldots,Y_{\ell+1},Z_{\ell+1})},$$

Corollary 6 tells us that

$$\Pr_X\left(|p^{(\ell)}_{\theta}(X)|_2 < \epsilon|p^{(\ell+1)}_{\theta}(X)|_2\right) = O(d^2 27^d\epsilon).$$

□

4 Averaging Operators

A key ingredient of our proof will be to show that if we replace X by X' by replacing one of the X_i by a random Gaussian, that the variance of $|p^{(\ell)}(X')|$ is bounded in terms of $|p^{(\ell+1)}(X)|$ (see Proposition 12). Unfortunately, for our argument to work nicely we would want it bounded in terms of the expectation of $|p^{(\ell+1)}(X')|$. In order to deal with this issue, we will need to study the behavior of the expectation of $q(X')$ for polynomials q, and in particular when it is close to $q(X)$. To get started on this project, we define the following averaging operator:

Definition. *Let X be an n-dimensional vector, $f : \mathbb{R}^n \to \mathbb{R}$ a function, and θ a real number. Define*

$$A^\theta f(X) := E_Y \left[f(N_Y^\theta(X)) \right].$$

Where the expectation is over Y a standard Gaussian.

Note that the A^θ form the Ornstein-Uhlenbeck semigroup with $A^\theta = T_t$ where $\cos(\theta) = e^{-t}$. The composition law becomes $T_{t_1} T_{t_2} = T_{t_1+t_2}$. We express this operator in terms of θ rather than t, since it fits in with our N^θ notation, which makes it more convenient for our purposes.

We also define averaged versions of our derivatives

Definition. *For p a polynomial, ℓ, m non-negative integers, X a vector, and θ a real number let*

$$|p_\theta^{(\ell),m}(X)|_2^2 := E_{Y_1,\ldots,Y_m} \left[|p_\theta^{(\ell)}(N_{Y_1}^\theta \cdots N_{Y_m}^\theta X)|_2^2 \right] = (A^\theta)^m |p_\theta^{(\ell)}(X)|_2^2.$$

We claim that for X a standard Gaussian, that with fairly high probability $|p_\theta^{(\ell),m}(X)|$ is close to $|p_\theta^{(\ell)}(X)|$. In particular:

Lemma 11. *If p is a degree-d polynomial, and ℓ and m are non-negative integers, X a standard Gaussian, and $\epsilon, \theta > 0$, then*

$$Pr_X \left(||p_\theta^{(\ell),m+1}(X)|_2^2 - |p_\theta^{(\ell),m}(X)|_2^2| > \epsilon |p_\theta^{(\ell),m}(X)|_2^2 \right) \leq 2^{O(d)} \theta \epsilon^{-1}.$$

Proof. If $||p_\theta^{(\ell),m+1}(X)|_2^2 - |p_\theta^{(\ell),m}(X)|_2^2| > \epsilon |p_\theta^{(\ell),m}(X)|_2^2$, then by Corollary 4 with probability $2^{O(d)}$ over a standard normal Y we have that

$$||p_\theta^{(\ell),m}(N_Y^\theta(X))|_2^2 - |p_\theta^{(\ell),m}(X)|_2^2| > \epsilon |p_\theta^{(\ell),m}(X)|_2^2.$$

By Lemma 9, this happens with probability $O(d^2 \theta \epsilon^{-1})$, and hence our original event happens with probability at most $2^{O(d)} \theta \epsilon^{-1}$. □

We bound the variance of these polynomials as X changes.

Proposition 12. *Let p be a degree-d polynomial, and $\theta > 0$, ℓ, m non-negative integers, then for X any vector and Y a standard Gaussian*

$$Var_Y\left[|p_\theta^{(\ell),m}(N_Y^\theta X)|_2^2\right] \leq 2^{O(d)}\theta |p_\theta^{(\ell+1),m}(X)|_2^2 |p_\theta^{(\ell),m+1}(X)|_2^2.$$

We begin by proving a Lemma:

Lemma 13. *Let $q(X,Y)$ be a degree-d polynomial and let X, Y and Z be independent standard Gaussians. Then*

$$Var_Y\left[E_X\left[q(X,Y)^2\right]\right] \leq 2^{O(d)} E_{X,Y,Z}\left[(q(X,Y)-q(X,Z))^2\right] E_{X,Y}\left[q(X,Y)^2\right].$$

Proof. Recall that if A, B are i.i.d. random variables, then $\mathrm{Var}[A] = \frac{1}{2}\mathrm{E}[(A-B)^2]$. We have that

$$\begin{aligned}
\mathrm{Var}_Y\left[\mathrm{E}_X\left[q(X,Y)^2\right]\right] &= \frac{1}{2}\mathrm{E}_{Y,Z}\left[\left(\mathrm{E}_X[q(X,Y)^2] - E_X[q(X,Z)^2]\right)^2\right] \\
&\leq 2^{O(d)} E_{Y,Z}\left[\left|\mathrm{E}_X[q(X,Y)^2 - q(X,Z)^2]\right|\right]^2 \\
&\leq 2^{O(d)} E_{X,Y,Z}\left[|q(X,Y)-q(X,Z)||q(X,Y)+q(X,Z)|\right]^2 \\
&\leq 2^{O(d)} \mathrm{E}_{X,Y,Z}\left[(q(X,Y)-q(X,Z))^2\right] \mathrm{E}_{X,Y,Z}\left[(q(X,Y)+q(X,Z))^2\right] \\
&\leq 2^{O(d)} \mathrm{E}_{X,Y,Z}\left[(q(X,Y)-q(X,Z))^2\right] \mathrm{E}_{X,Y}\left[q(X,Y)^2\right].
\end{aligned}$$

The second line above is due to Corollary 5. The fourth line is due to Cauchy-Schwarz. □

Proof of Proposition 12. For fixed X, consider the polynomial

$$\begin{aligned}
&q((Y_2, Z_2, \ldots, Y_{\ell+1}, Z_{\ell+1}, W_1, \ldots, W_m), Y) \\
&=: D^\theta_{Y_2,Z_2} \cdots D^\theta_{Y_{\ell+1},Z_{\ell+1}} p(N^\theta_{W_1} \cdots N^\theta_{W_m} N^\theta_Y(X)).
\end{aligned}$$

Let $V = (Y_2, Z_2, \ldots, Y_{\ell+1}, Z_{\ell+1}, W_1, \ldots, W_m)$. Notice that

$$\begin{aligned}
|p_\theta^{(\ell),m}(N_Y^\theta X)|_2^2 &= \mathrm{E}_V\left[q(V,Y)^2\right], \\
\theta|p^{(\ell+1),m}(X)|_2^2 &= \mathrm{E}_{V,Y,Z}\left[(q(V,Y)-q(V,Z))^2\right], \\
|p_\theta^{(\ell),m+1}(X)|_2^2 &= \mathrm{E}_{V,Y}\left[q(V,Y)^2\right].
\end{aligned}$$

Our result follows immediately upon applying the above Lemma to $q(V,Y)$. □

We also prove a relation between the higher averages

Lemma 14. *Let d be an integer and $\theta = O(d^{-1})$ a real number. Then there exist constants $c_0, \ldots, c_{2d+1}$ with $|c_m| = 2^{O(d)}$ and $\sum_{m=0}^{2d+1} c_m = 0$ so that for any degree-d polynomial p and any vector X,*

$$\sum_{m=0}^{2d+1} c_m (A^\theta)^m p(X) = 0.$$

Proof. First note that $(A^\theta)^m p(X) = A^{\theta_m} p(X)$ where $\cos(\theta_m) = \cos(\theta)^m$. We note that it suffices to find such c's so that for any Y

$$\sum_{m=0}^{2d+1} c_m p(N_Y^{\theta_m}(X)) = 0.$$

Note that $\cos^2(\theta_m) = 1 - m\theta^2 + O(d\theta^4)$. Hence $\sin(\theta_m) = \sqrt{m}\theta + O(d\theta^3)$.

Recall that once X and Y are fixed, there is a degree-$2d$ polynomial q so that for $z_m = e^{i\theta_m} = 1 + i\sqrt{m}\theta + O(d\theta^2)$, $p(N_Y^{\theta_m}(X)) = z_m^{-d} q(z_m)$. Hence we just need to pick c_m so that for any degree-$2d$ polynomial q we have that

$$\sum_{m=0}^{2d+1} c_m z_m^{-d} q(z_m) = 0.$$

Such c exist by standard interpolation results. In particular, it is sufficient to pick

$$c_m = z_m^d \sqrt{(2d)!} \prod_{i=1, i\neq m}^{2d+1} \frac{\theta}{z_i - z_m}.$$

For this choice of c we have that

$$|c_m| = \sqrt{(2d)!} \prod_{i=1, i\neq m}^{2d+1} \frac{\theta}{|z_i - z_m|} = \sqrt{(2d)!} \prod_{i=1, i\neq m}^{2d+1} \Theta\left(\frac{1}{|\sqrt{i} - \sqrt{m}|}\right)$$

For $1 \leq i \leq 2m$ we have that $|\sqrt{i} - \sqrt{m}| = \Theta(|i-m|/\sqrt{m})$. For $i \geq 2m$, we have that $|\sqrt{i} - \sqrt{m}| = \Theta(\sqrt{i})$. We evaluate $|c_m|$ based upon whether or not $2m \geq 2d+1$. If $2m \geq 2d+1$, then

$$\begin{aligned}
|c_m| &= \sqrt{(2d)!} \prod_{i=1, i\neq m}^{2d+1} \Theta\left(\frac{\sqrt{m}}{|i-m|}\right) \\
&= 2^{O(d)} \sqrt{(2d)!} \frac{m^d}{(m-1)!(2d+1-m)!} \\
&= 2^{O(d)} \sqrt{d^{2d}} d^d \frac{\binom{2d}{m-1}}{(2d!)} \\
&= 2^{O(d)}.
\end{aligned}$$

If $2m < 2d+1$,

$$\begin{aligned}
|c_m| &= \sqrt{(2d)!} \prod_{i=1, i\neq m}^{2m} \Theta\left(\frac{\sqrt{m}}{|i-m|}\right) \prod_{i=2m+1}^{2d+1} \Theta\left(\frac{1}{\sqrt{i}}\right) \\
&= 2^{O(d)} \sqrt{(2d)!} \frac{m^m}{(m-1)!(m-1)!} \sqrt{\frac{(2m)!}{(2d+1)!}} \\
&= 2^{O(d)} \frac{m^m \sqrt{(2m)!}}{(m-1)!(m-1)!} \\
&= 2^{O(d)} \frac{m^m \sqrt{(2m)!}\binom{2m-2}{m-1}}{(2m-2)!} \\
&= 2^{O(d)}.
\end{aligned}$$

Considering $q(X) = 1$, we find that $\sum_{m=0}^{2d+1} c_m = 0$. □

Applying this to the polynomial $|p_\theta^{(\ell)}(X)|$, we find that

Corollary 15. *Let d be an integer and $\theta = O(d^{-1})$ a sufficiently small real number (as a function of d). Then there exist constants $c_0, \ldots, c_{4d+1}$ with $|c_m| = \Theta_d(1)$ and $\sum_{m=0}^{4d+1} c_m = 0$ so that for any degree-d polynomial p, any vector X, and any non-negative integer ℓ,*

$$\sum_{m=0}^{4d+1} c_m |p_\theta^{(\ell),m}(X)|_2^2 = 0.$$

Furthermore $\sum_{m=0}^{4d+1} c_m = 0$ and $|c_m| = 2^{O(d)}$.

In particular,

Corollary 16. *There exists some absolute constant α, so that if $\theta = O(d^{-1})$ and if*

$$\left|\log\left(\frac{|p_\theta^{(\ell),m}(X)|_2^2}{|p_\theta^{(\ell),m+1}(X)|_2^2}\right)\right| < \alpha^d$$

for some ℓ and all $1 \le m \le 4d$, then all of the $|p_\theta^{(\ell),m}(X)|_2^2$ for that ℓ and all $0 \le m \le 4d+1$ are within constant multiples of each other.

Proof. It is clear that $|p_\theta^{(\ell),m}(X)|_2^2$ are within $1 + O(d\alpha^d)$ of each other for $1 \le m \le 4d+1$. By Corollary 15, we have that

$$\begin{aligned}
|p_\theta^{(\ell),0}(X)|_2^2 &= \sum_{m=1}^{4d} \frac{-c_m}{c_0} |p_\theta^{(\ell),m}(X)|_2^2 \\
&= \sum_{m=1}^{4d} \frac{-c_m}{c_0} |p_\theta^{(\ell),1}(X)|_2^2 \left(1 + O(d\alpha^d)\right) \\
&= \sum_{m=1}^{4d} \frac{-c_m}{c_0} |p_\theta^{(\ell),1}(X)|_2^2 + |p_\theta^{(\ell),1}(X)|_2^2 O\left(d\alpha^d \sum_{m=1}^{4d} \left|\frac{c_m}{c_0}\right|\right) \\
&= |p_\theta^{(\ell),1}(X)|_2^2 \left(1 + O\left(d^2\alpha^d 2^{O(d)}\right)\right).
\end{aligned}$$

Hence for α sufficiently small, $|p_\theta^{(\ell),0}(X)|_2^2$ is within a constant multiple of $|p_\theta^{(\ell),1}(X)|_2^2$. □

5 Proof of Theorem 1

We now fix c, ϵ, d, N, k, p as in Theorem 1. Namely $c, \epsilon > 0$, d is a positive integer, N an integer bigger than $B(c)^d \epsilon^{-4-c}$, and k an integer bigger than $B(c)d$ for $B(c)$ some sufficiently large number depending only on c, and p a degree-d polynomial. We fix $\theta = \arcsin\left(\frac{1}{\sqrt{N}}\right) \sim B(c)^{d/2}\epsilon^{-2-c/2}$.

Let $\rho : \mathbb{R} \to [0,1]$ be a smooth function so that $\rho(x) = 0$ if $x < -1$ and $\rho(x) = 1$ if $x > 0$. Let $\sigma : \mathbb{R} \to [0,1]$ a smooth function so that $\sigma(x) = 1$ if $|x| < 1/3$ and $\sigma(x) = 0$ if $|x| > 1/2$. Let α be the constant given in Corollary 16.

For $0 \le \ell \le d$, $0 \le m \le 4d+1$, let $q_{\ell,m}(X)$ be the degree-$2d$ polynomial $|p_\theta^{(\ell),m}(X)|_2^2$. Recall by Lemma 7 that $q_{d,m}$ is constant. We let $g_\pm(X)$ be

$$I_{(0,\infty)}(\pm p(X)) \prod_{\ell=0}^{d-1} \rho\left(\log\left(\frac{q_{\ell,0}(X)}{\epsilon^2 q_{\ell+1,0}(X)}\right)\right) \prod_{m=0}^{4d-1} \sigma\left(\alpha^{-d} \log\left(\frac{q_{\ell,m}(X)}{q_{\ell,m+1}(X)}\right)\right).$$

Where $I_{(0,\infty)}(x)$ above is the indicator function of the set $(0,\infty)$. Namely it is 1 for $x > 0$ and 0 otherwise.

$g_\pm$ approximates the indicator functions of the sets where $p(X)$ is positive or negative. To make this intuitive statement useful we prove:

Lemma 17. *The following hold:*

- $g_\pm : \mathbb{R}^n \to [0,1]$
- $g_+(X) + g_-(X) \le 1$ *for all* X

- *For Y a standard Gaussian $E_Y[1 - g_+(Y) - g_-(Y)] \leq 2^{O(d)}\epsilon$.*

Proof. The first two statements follow immediately from the definition. The third statement follows by noting that $g_+(Y) + g_-(Y) = 1$ unless $q_{\ell,0}(Y) < \epsilon^2 q_{\ell+1,0}(X)$ for some ℓ or $|q_{\ell,m+1}(X) - q_{\ell,m}(X)| < \Omega(\alpha^d)|q_{\ell,m}(X)|$ for some ℓ, m. By Corollary 10 and Lemma 11, this happens with probability at most $2^{O(d)}\epsilon$. □

We also want to know that the derivatives of $g_\pm$ are relatively small.

Lemma 18. *Consider $g_\pm$ as a function of $q_{\ell,m}(X)$ (consider $sgn(p(X))$ to be constant). Then the t^{th} partial derivative of $g_\pm$, $\frac{\partial^t}{\partial q_{\ell_1,m_1}\partial q_{\ell_2,m_2}\cdots\partial q_{\ell_t,m_t}}$ is at most $O_t(1)2^{O(dt)}\prod_{j=1}^t \frac{1}{q_{\ell_j,m_j}}$.*

Proof. The bound follows easily after considering g as a function of the $\log(q_{\ell,m})$. Noting that the t^{th} partial derivatives of q in terms of these logs are at most $O_t(\alpha^{-dt})$, the result follows easily. □

Lemma 19. *Given $c \leq 4$, let X be any vector and let Y and Z be k-independent families of Gaussians with $k \geq 512c^{-1}d$, $N = \epsilon^{-4-c}$ and $\theta = O(d^{-2})$. Then*

$$\left|E_Y\left[g_+(N_Y^\theta(X))\right] - E_Z\left[g_+(N_Z^\theta(X))\right]\right| \leq 2^{O_c(d)}\epsilon N^{-1}.$$

And the analogous statement holds for g_-.

Proof. First note that $\epsilon^2\theta = O(\epsilon^{c/2})$, and hence that $(\epsilon^2\theta)^{16/c} = O(\epsilon N^{-1})$. Let T be an even integer between $32/c$ and $64/c$.

First we deal with the case where $|\log(q_{\ell,m}(X)/q_{\ell,m+1}(X))| > \alpha^d$ for some $0 \leq \ell \leq d, 1 \leq m \leq 4d$, or $q_{\ell,1}(X) < \epsilon^2/10 q_{\ell+1,1}(X)$ for some ℓ. We claim that in either case the $\mathrm{E}\left[g_+(N_Y^\theta(X))\right] = O_d(\epsilon N^{-1})$ and a similar bound holds for Z. If there is such an occurrence, find one with the largest possible ℓ and of the second type if possible for the same value of ℓ.

Suppose that we had an occurrence of the first type. Namely that for some ℓ, m,

$$|\log(q_{\ell,m}(X)/q_{\ell,m+1}(X))| > \alpha^d.$$

Pick such a one with ℓ maximal, and with m minimal for this value of ℓ. Consider then the random variables $q_{\ell,m-1}(N_Y^\theta(X))$ and $q_{\ell,m}(N_Y^\theta(X))$. They have means $q_{\ell,m}(X)$ and $q_{\ell,m+1}(X)$, respectively. By Proposition 12, their variances are bounded by

$$2^{O(d)}\theta q_{\ell+1,m-1}(X)q_{\ell,m}(X), \text{ and } 2^{O(d)}\theta q_{\ell+1,m}(X)q_{\ell,m+1}(X),$$

respectively. Since we chose the smallest such ℓ, all of the $q_{\ell+1,m}(X)$ for $1 \leq m \leq 4d+1$ are close to $q_{\ell+1,1}(X)$ with multiplicative error at most α^d. By Corollary 16, this implies that $q_{\ell+1,0}(X)$ is also close. Hence, the variances of $q_{\ell,m-1}(N_Y^\theta(X))$ and $q_{\ell,m}(N_Y^\theta(X))$ are $O_d(\theta q_{\ell+1,1}(X)q_{\ell,m}(X))$ and $O_d(\theta q_{\ell+1,1}(X)q_{\ell,m+1}(X))$. Since there was no smaller m to choose,

$q_{\ell,m}(X)$ is within a constant multiple of $q_{\ell,1}(X)$. Since we could not have picked an occurrence of the second type with the same ℓ, we have that $q_{\ell,1}(X) \geq \epsilon^2 q_{\ell+1,1}(X)/10$. Hence both of these variances are at most

$$2^{O(d)}\epsilon^{-2}\theta \max(q_{\ell,m}(X), q_{\ell,m+1}(X))^2).$$

Hence by Corollary 5, for either of the random variables $Q_1 = q_{\ell,m}(N_Y^\theta(X))$, or $Q_2 = q_{\ell,m+1}(N_Y^\theta(X))$ with means μ_1, μ_2 the T^{th} moment of $|Q_i - \mu_i|$ (using the fact that Y is at least $4Td$-independent) is at most $2^{O_c(d)}\epsilon N^{-1}\max(\mu_i)^T$. Hence, with probability at least $1 - 2^{O_c(d)}\epsilon N^{-1}$, $|Q_i - \mu_i| < \alpha^d \max(\mu_i)/10$. But if this is the case, then $|\log(Q_1/Q_2)|$ will be more than $\alpha^d/2$, and g_+ will be 0.

Suppose that we had an occurrence of the second type for some ℓ. Again by Corollary 16, we have that $q_{\ell+1,0}(X)$ is within a constant multiple of $q_{\ell+1,1}(X)$ and $q_{\ell+2,0}(X)$ within a constant multiple of $q_{\ell+2,1}(X)$. Let Q_0 be the random variable $q_{\ell,0}(N^\theta(Y))$ and Q_1 the variable $q_{\ell+1,0}(N^\theta(Y))$. We note that they have means equal to $q_{\ell,1}(X)$ and $q_{\ell+1,1}(X)$, respectively. By Proposition 12, their variances are at most $2^{O(d)}\theta q_{\ell+1,1}(X)q_{\ell,1}(X)$ and $2^{O(d)}\theta q_{\ell+2,1}(X)q_{\ell+1,1}(X)$. Since we had an occurrence at this ℓ but not the larger one, these are at most $2^{O(d)}\theta\epsilon^2 q_{\ell+1,1}(X)^2$ and $2^{O(d)}\theta\epsilon^{-2}q_{\ell+1,1}(X)^2$, respectively. Considering the T^{th} moment of Q_1 minus its mean, μ_1, we find that with probability at least $1 - 2^{O_c(d)}\epsilon N^{-1}$ that $|Q_1 - q_{\ell+1,1}| < q_{\ell+1,1}/20$. Considering the T^{th} moment of Q_0 minus its mean, we find that with probability at least $1 - 2^{O_c(d)}\epsilon N^{-1}$ that $|Q_0 - q_{\ell,1}| < \epsilon^2 q_{\ell+1,1}/20$. Together these imply that $\log(Q_0/(\epsilon^2 Q_1)) < -1$ and hence that $g_+(N_Y^\theta(X)) = 0$.

Finally, we assume that neither of these cases occur. We note by Corollary 16 that for each m and ℓ that $q_{\ell,m}(X)$ is within a constant multiple of $q_{\ell,1}(X)$. We define $Q_{\ell,m} = q_{\ell,m}(N_Y^\theta(X))$ and note by Proposition 12 that $\mathrm{Var}(Q_{\ell,m})$ is at most $2^{O(d)}\theta\epsilon^{-2}\mathrm{E}[Q_{\ell,m}]^2$. We wish to show that this along with the k-independence of Y is enough to determine $\mathrm{E}_Y[g_+(Q_{\ell,m})]$ to within $2^{O_c(d)}\epsilon N^{-1}$. We do this by approximating g_+ by its Taylor series to degree $T-1$ about $(\mathrm{E}[Q_{\ell,m}])$. The expectation of the Taylor polynomial is determined by the $4Td$-independence of Y. We have left to show that the expectation of the Taylor error is small. We split this error into cases based on whether $Q_{\ell,m}$ differs from its mean value by more than a constant multiple.

If no $Q_{\ell,m}$ varies by this much, the Taylor error is at most the sum over sequences $\ell_1, \ldots, \ell_T, m_1, \ldots, m_T$ of $\prod_{i=1}^T |Q_{\ell_i,m_i} - \mathrm{E}[Q_{\ell_i,m_i}]|$ times an appropriate partial derivative. Note that by Lemma 18 this derivative has size at most $O_d\left(\prod_{i=1}^T \frac{1}{|\mathrm{E}[Q_{\ell_i,m_i}]|}\right)$. Noting that there are at most $2^{O_c(d)}$ such terms and that we can bound the expectation above as

$$\sum_{i=1}^{T}\left(\frac{|Q_{\ell_i,m_i} - \mathrm{E}[Q_{\ell_i,m_i}]|}{|\mathrm{E}[Q_{\ell_i,m_i}]|}\right)^T.$$

Since T is even, the above is a polynomial in Y of degree $4Td$. Since Y is $4Td$-independent, the expectation of the above is the same as it would be for Y fully independent. By Corollary

5 this is at most

$$2^{O_c(d)} \sum_{i=1}^{T} \left(\frac{\mathrm{Var}_Y[Q_{\ell_i,m_i}]}{\mathrm{E}[Q_{\ell_i,m_i}]^2} \right)^{T/2} \leq 2^{O_c(d)} (\theta\epsilon^{-2})^{T/2} \leq 2^{O_c(d)} \epsilon N^{-1}.$$

If some $Q_{\ell,m}$ differs from its mean by a factor of more than 2, the Taylor error is at most 1 plus the size of our original Taylor term. By Cauchy-Schwarz, the contribution to the error is at most the square root of the expectation of the square of the error term times the square root of the probability that one of the $Q_{\ell,m}$ varies by too much. By an argument similar to the above, the former is $2^{O_c(d)}$. To bound the latter, we consider the probability that a particular $Q_{\ell,m}$ varies by too much. For this to happen $Q_{\ell,m}$ would need to differ from its mean by $2^{O(d)} \left(\theta\epsilon^{-2}\right)^{-1/2}$ times its standard deviation. Using Corollary 5 and the $8Td$-independence of Y, we bound this by considering the $2T^{th}$ moment of $Q_{\ell,m}$ minus its mean value, and obtain a probability of $2^{O_c(d)}(\epsilon N^{-1})^2$. Hence this term produces a total error of $2^{O_c(d)}\epsilon N^{-1}$.

The argument for g_- is analogous. □

Corollary 20. *If $\epsilon, c > 0$, X is the random variable described in Theorem 1 with $N \geq \epsilon^{-4-c}$, $N = O(d^{-2})$, $k \geq 512c^{-1}d$ and Y is a fully independent family of random Gaussians then*

$$|E[g_+(X)] - E[g_+(Y)]| \leq 2^{O_c(d)}\epsilon.$$

The same also holds for g_-.

Proof. We let Y_i be independent random standard Gaussians and let $Y = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y_i$. We let $Z^j = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{j} Y_i + \sum_{i=j+1}^{N} X_i \right)$. Note that $Z^N = Y$ and $Z^0 = X$. We claim that

$$|\mathrm{E}[g_+(Z^j)] - \mathrm{E}[g_+(Z^{j+1})]| = 2^{O_c(d)}\epsilon N^{-1},$$

from which our result would follow. Let $Z_j := \frac{1}{\sqrt{N-1}} \left(\sum_{i=1}^{j} Y_i + \sum_{i=j+2}^{N} X_i \right)$, then this is

$$\begin{aligned} &|\mathrm{E}_{Z_j,Y_j}[g_+(N^\theta_{Y_j}(Z_j))] - \mathrm{E}_{Z_j,X_j}[g_+(N^\theta_{X_j}(Z_j))]| \\ &\qquad \leq E_{Z_j}\left[|\mathrm{E}_{Y_j}[g_+(N^\theta_{Y_j}(Z_j))] - \mathrm{E}_{X_j}[g_+(N^\theta_{X_j}(Z_j))]| \right], \end{aligned}$$

which by Lemma 19 is at most $2^{O_c(d)}\epsilon N^{-1}$. □

We can now prove Theorem 1.

Proof. We prove that for X as given with $N \geq \epsilon^{-4-c}$ and $k \geq 512c^{-1}d$, and Y fully independent that

$$|\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]| \leq 2^{O_c(d)}\epsilon.$$

The rest will follow by replacing ϵ by $\epsilon' = \epsilon/2^{O_c(d)}$. We note that f is sandwiched between $2g_+ - 1$ and $1 - 2g_-$. Now X fools both of these functions to within $2^{O_c(d)}\epsilon$. Furthermore by Lemma 17, they have expectations that differ by $2^{O(d)}\epsilon$. Therefore

$$\mathrm{E}[f(X)] \leq \mathrm{E}[1 - 2g_-(X)] \approx_{2^{O_c(d)}\epsilon} \mathrm{E}[1 - 2g_-(Y)] \approx_{2^{O(d)}\epsilon} \mathrm{E}[f(Y)].$$

We also have a similar lower bound. This proves the Theorem. □

6 Finite Entropy Version

The random variable X described in the previous sections, although it does fool PTFs, has infinite entropy and hence cannot be used directly to make a PRG. We fix this by instead using a finite entropy random variable that approximates X. In order to make this work, we will need the following Lemma.

Lemma 21. *Let X_i be a k-independent family of Gaussians for $1 \leq i \leq N$, so that the X_i are independent of each other and k, N satisfy the hypothesis in Theorem 1 for some c, ϵ, d. Let $\delta > 0$. Suppose that $Z_{i,j}$ are any random variables so that for each i, j, $|X_{i,j} - Z_{i,j}| < \delta$ with probability $1 - \delta$. Then the family of random variables $Z_j = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_{i,j}$ fools degree d polynomial threshold functions up to $\epsilon + O(nN\delta + d\sqrt{nN}\log(\delta^{-1})\delta^{1/d})$.*

The basic idea is that with probability $1 - nN\delta$, we will have that $|X_{i,j} - Z_{i,j}| < \delta$ for all i, j. If that is the case, then for any polynomial p it should be the case that $p(X)$ is close to $p(Z)$. In particular, we show:

Lemma 22. *Let $p(X)$ be a polynomial of degree d in n variables. Let $X \in \mathbb{R}^n$ be a vector with $|X|_\infty \leq B$ $(B > 1)$. Let X' be another vector, so that $|X - X'|_\infty < \delta < 1$. Then*

$$|p(X) - p(X')| \leq \delta |p|_2 n^{d/2} O(B)^d$$

Proof. We begin by writing p in terms of Hermite polynomials. We can write

$$p(X) = \sum_{i \in S} a_i h_i(X).$$

Here S is a set of size less than n^d, $h_i(X)$ is a Hermite polynomial of degree d and $\sum_{i \in S} a_i^2 = |p|_2^2$. The Hermite polynomial h_i has $2^{O(d)}$ coefficients each of size $2^{O(d)}$. Hence any of its partial derivatives at a point of L^∞ norm at most B is at most $O(B)^d$. Hence by the intermediate value theorem, $|h_i(X) - h_i(X')| = \delta O(B)^d$. Hence $|p(X) - p(X')| \leq \delta \sum_{i \in S} |a_i| O(B)^d$. But $\sum_{i \in S} |a_i| \leq |p|_2 \sqrt{|S|}$ by Cauchy-Schwarz. Hence

$$|p(X) - p(X')| \leq \delta |p|_2 n^{d/2} O(B)^d.$$

□

We also need to know that it is unlikely that changing the value of p by a little will change its sign. In particular we have the following anticoncentration result, which is an easy consequence of [7] Theorem 8:

Lemma 23 (Carbery and Wright)**.** *If p is a degree-d polynomial then*

$$Pr(|p(X)| \leq \epsilon |p|_2) = O(d\epsilon^{1/d}).$$

Where the probability is over X, a standard n-dimensional Gaussian.

We are now ready to prove Lemma 21.

Proof. Note that with probability $1-\delta$ that $|X_{i,j}| = O(\log \delta^{-1})$. Hence with probability $1 - O(nN\delta)$ we have that $|X_{i,j}|_\infty = O(\log \delta^{-1})$ and $|X_{i,j} - Z_{i,j}|_\infty < \delta$. Let p be a degree d polynomial normalized so that $|p|_2 = 1$. We may think of p as a function of nN variables rather than just N, by thinking of $p(X)$ instead as $p\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i\right)$. Applying Lemma 22, we have therefore that with probability $1 - O(nN\delta)$ that $|p(X) - p(Z)| < \delta O(\sqrt{nN} \log(\delta^{-1}))^d$.

We therefore have that if Y is a standard family of Gaussians that

$$\begin{aligned}
\Pr(p(Z) < 0) &\leq O(nN\delta) + \Pr(p(X) < \delta O(\sqrt{nN} \log(\delta^{-1}))^d) \\
&\leq \epsilon + O(nN\delta) + \Pr(p(Y) < \delta O(\sqrt{nN} \log(\delta^{-1}))^d) \\
&\leq \epsilon + O(nN\delta + d\sqrt{nN} \log(\delta^{-1})\delta^{1/d}) + \Pr(p(Y) < 0).
\end{aligned}$$

The last step above following from Lemma 23. We similarly get a bound in the other direction, completing the proof. □

We are now prepared to prove Corollary 2.

Proof. Given $\epsilon, c > 0$, let k, N be as required in the statement of Theorem 1. We will attempt to produce an effectively computable family of random variables $Z_{i,j}$ so that for some k-independent families of Gaussians X_i we have that $|X_{i,j} - Z_{i,j}| < \delta$ with probability $1 - \delta$ for each i, j and δ sufficiently small. Our result will then follow from Lemma 21.

Firstly, it is clear that in order to do this we need to understand how to actually effectively compute Gaussian random variables. Note that if u and v are independent uniform $[0, 1]$ random variables, then $\sqrt{-2\log(u)}\cos(2\pi v)$ is a Gaussian. Hence we can let our $X_{i,j}$ be given by

$$X_{i,j} = \sqrt{-2\log(u_{i,j})}\cos(2\pi v_{i,j}),$$

where u_i and v_i are k-independent families of uniform $[0, 1]$ random variables. We let $u'_{i,j}, v'_{i,j}$ be M-bit approximations to $u_{i,j}, v_{i,j}$ (i.e. $u'_{i,j}$ is $u_{i,j}$ rounded up to the nearest multiple of 2^{-M}, and similarly for $v'_{i,j}$), and let $Z_{i,j} = \sqrt{-2\log(u'_{i,j})}\cos(2\pi v'_{i,j})$. Note that we can equivalently

compute $Z_{i,j}$ be letting u'_i, v'_i be k-independent families of variables taken uniformly from $\{2^{-M}, 2 \cdot 2^{-M}, \ldots, 1\}$. Hence, the $Z_{i,j}$ are effectively computable from a random seed of size $O(kNM)$.

We now need to show that $|X_{i,j} - Z_{i,j}|$ is small with high probability. Let $a(u,v) = \sqrt{-2\log(u)}\cos(2\pi v)$. Note that for $u, v \in [0,1]$ that $|a'| = O(1 + u^{-1} + (1-u)^{-1/2})$. Therefore, (unless $u'_{i,j} = 1$) we have that since $X_{i,j} = a(u_{i,j}, v_{i,j})$ and $Z_{i,j} = a(u'_{i,j}, v'_{i,j})$, and since $|u_{i,j} - u'_{i,j}|, |v_{i,j} - v'_{i,j}| \leq 2^{-M}$, we have that

$$|X_{i,j} - Z_{i,j}| = O(2^{-M}(1 + u^{-1} + (1-u)^{-1/2})).$$

Now letting $\delta = \Omega(2^{-M/2})$, we have that $2^{-M}(1 + u^{-1} + (1-u)^{-1/2}) < \delta$ with probability more than $1 - \delta$. Hence for such δ, we can apply Lemma 21 and find that Z fools degree d polynomial threshold functions to within $\epsilon + O(nN\delta + d\sqrt{nN}\log(\delta^{-1})\delta^{1/d})$. If $\delta < \epsilon^{3d}(dnN)^{-3d}$, then this is $O(\epsilon)$ (since for $x > d^{3d}$, we have that $x\log^{-d}(x) > x^{1/3}$). Hence with $k = \Omega_c(d), N = 2^{\Omega_c(d)}\epsilon^{-4-c}$ and $M = \Omega_c(d\log(dn\epsilon^{-1}))$, this gives us a PRG that ϵ-fools degree d polynomial threshold functions and has seed length $O(kNM)$. Changing c by a bit to absorb the $\log \epsilon^{-1}$ into the ϵ^{-4-c}, and absorbing the $d\log d$ into the $2^{O_c(d)}$, this seed length is $\log(n)2^{O_c(d)}\epsilon^{-4-c}$.

□

Bibliography

[1] C. Bachoc, R. Coulangeon, and G. Nebe. Designs in grassmannian spaces and lattices. *Journal of Algebraic Combinatorics*, 16(1):5–19, 2002.

[2] A. Balog and A. Perelli. Exponential sums over primes in an arithmetic progression. *Proc. of the AMS*, 93(4):578–582, 1985.

[3] R. Beigel. The polynomial method in circuit complexity. In *Proceedings of 8th Annual Structure in Complexity Theory Conference*, pages 82–95, 1993.

[4] I. Ben-Eliezer, S. Lovett, and A. Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness. Manuscript available at `http://arxiv.org/PS_cache/arxiv/pdf/0911/0911.3473v3.pdf`.

[5] A. Bondarenko and M. Viazovska. Spherical designs via Brouwer fixed point theorem. *SIAM J. Discrete Math.*, 24:207–217, 2010.

[6] A. V. Bondarenko, D. V. Radchenko, and M. S. Viazovska. On optimal asymptotic bounds for spherical designs. Manuscript available at `http://arxiv.org/abs/1009.4407`.

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

[8] V. H. De La Pena and S. Montgomery-Smith. Bounds for the tail probabilities of u-statistics and quadratic forms. *Bulletin of the American Mathematical Society*, 31:223–227, 1994.

[9] P. Deligne and G. Lusztig. Representations of reductive groups over finite fields. *Annals of Mathematics*, 103:103–161, 1976.

[10] P. Delsarte, J. Goethals, and J. Seidel. Spherical codes and designs. *Geometriae Dedicata*, 6:363–388, 1977.

[11] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. In *50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2009.

[12] I. Diakonikolas, P. Harsha, A. R. Klivans, P. R. Raghu Meka, R. A. Servedio, and L.-Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *42nd Annual Symposium on Theory of Computing (STOC)*, pages 533–542, 2010.

[13] I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *Foundations of Computer Science (FOCS)*, 2010.

[14] I. Diakonikolas, P. Raghavendra, R. Servedio, and L.-Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. Manuscript available at `http://arxiv.org/abs/0909.5011`, 2009.

[15] I. Diakonikolas, R. Servedio, L.-Y. Tan, and A. Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. In *25th Conference on Computational Complexity (CCC)*, 2010.

[16] C. Gotsman and N. Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.

[17] B. H. Gross. Rigid local systems on $\mathbb{G}_m$ with finite monodromy. *Advances in Mathematics*, 224:2531–2543, 2010.

[18] G.Wagner. On averaging sets. *Monatsh. Math.*, 111:69–78, 1991.

[19] K. Halupczok. On the ternary Goldbach problem with primes in arithmetic progressions having a common modulus. *J. Theor. Nombres Bordeaux*, 21(1):203–213, 2009.

[20] J. P. Hansen. Deligne-Lusztig varieties and group codes. In *Lecture Notes in Mathematics*, volume 1518, pages 63–81. Springer, 1992.

[21] P. Harsha, A. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. Manuscript available at `http://arxiv.org/abs/0909.5175"`, 2009.

[22] D. Heath-Brown. The size of Selmer groups for the congruent number problem, ii. *Inventiones mathematicae*, 118:331–370, 1994.

[23] D. Heath-Brown. The density of rational points on cubic surfaces. *Acta Arith.*, 79(1):17–30, 1997.

[24] H. Iwaniec and E. Kowalski. *Analytic Number Theory*. American Mathematical Society, 2004.

[25] D. M. Kane. An asymptotic for the number of solutions to linear equations in prime numbers from specified Chebotarev classes. Manuscript available at `http://www.math.harvard.edu/~dankane/chebGold.pdf`.

[26] D. M. Kane. Canonical projective embeddings of the Deligne-Lusztig curves associated to ${}^2A_2$, ${}^2B_2$ and ${}^2G_2$. unpublished.

[27] D. M. Kane. k-independent Gaussians fool polynomial threshold functions. To appear in the 26th annual IEEE Conference on Computational Complexity (CCC). Manuscript available at `http://www.math.harvard.edu/~dankane/foolingPTFs.pdf`.

[28] D. M. Kane. On a problem relating to the ABC conjecture. unpublished.

[29] D. M. Kane. On the ranks of the 2-Selmer groups of twists of a given elliptic curve. Manuscript available at `http://www.math.harvard.edu/~dankane/selmer.pdf`.

[30] D. M. Kane. Small designs for path connected spaces and path connected homogeneous spaces. unpublished.

[31] D. M. Kane. A small PRG for polynomial threshold functions of Gaussians. unpublished.

[32] D. M. Kane. The Gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. In *Proceedings of the 25th annual IEEE Conference on Computational Complexity (CCC)*, pages 205–210, 2010.

[33] D. M. Kane, J. Nelson, and D. P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

[34] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM J. Computing*, 37:319–357, 2007.

[35] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning geometric concepts via Gaussian surface area. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations Of Computer Science (FOCS)*, pages 541–550, 2008.

[36] A. R. Klivans and R. A. Servedio. Learning DNF in time $2^{O(n^{1/3})}$. *J. Computer and System Sciences*, 68:303–318, 2004.

[37] J. Korevaar and J. L. H. Meyers. Spherical Faraday cage for the case of equal point charges and Chebyshev-type quadrature on the sphere. *Integral Transforms Spec. Funct.*, 1:105–117, 1993.

[38] K. Kramer. Arithmetic of elliptic curves upon quadratic extension. *Transactions Amer. Math. Soc.*, 264:121–135, 1998.

[39] R. Latala. Estimates of moments of tails of gaussian choases. *The Annals of Probability*, 34(6):2315–2331, 2006.

[40] M. Ledoux. Semigroup proofs of the isoperimetric inequality in Euclidean and Gauss space. *Bull. Sci. Math.*, 118:485–510, 1994.

[41] G. Lorentz. *Approximation of functions.* Holt, Rinehart and Winston, 1966.

[42] G. Lusztig. Coxeter orbits and eigenspaces of frobenius. *Inventiones Mathematicae*, 38:101–159, 1976.

[43] B. Mazur. Questions about powers of numbers. *Notices of the AMS*, 47(2):195–202, 2000.

[44] B. Mazur and K. Rubin. Ranks of twists of elliptic curves and Hilbert's 10^{th} problem. *Invent. Math.*, 181:541–575, 2010.

[45] R. Meka and D. Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the 42nd ACM Symposium on Theory Of Computing (STOC)*, 2010.

[46] X. Meng. A mean value theorem on the binary Goldbach problem and its application. *Monatsh. Math.*, 151(4):319–332, 2007.

[47] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. Short version in *Proceedings of 46th Annual IEEE Symposium on the Foundations of Computer Science (FOCS)* Full manuscript available at `http://www.cs.cmu.edu/~odonnell/papers/invariance.pdf`.

[48] F. Nazarov. The Gaussian areas of origin-symmetric quadratic surfaces are uniformly bounded. Unpublished.

[49] Nelson. The free Markov field. *J. Func. Anal.*, 12(2):211–227, 1973.

[50] T. Noda. On the Goldbach problem in algebraic number fields and the positivity of the singular integral. *Kodai Math. J.*, 20(1):8–21, 1997.

[51] R. O'Donnell. Hardness amplification within NP. *J. Computer and System Sciences*, 69:68–94, 2004.

[52] J. P. Pedersen. A function field related to the Ree group. In *Lecture Notes in Mathematics*, volume 1518, pages 122–131. Springer, 1992.

[53] S. Ramanujan. The number of prime factors of a numer. *Quart. J. Math*, 48:76–92, 1917.

[54] H. Riesel and R. C. Vaughan. On sums of primes. *Arkiv för Matematik*, 21(1-2):45–74, 1969.

[55] P. Seymour and T. Zaslavsky. Averaging sets: A generalization of mean values and spherical designs. *Advances in Mathematics*, 52:213–240, 1984.

[56] A. A. Sherstov. Separating AC0 from depth-2 majority circuits. *SIAM J. Computing*, 38:2113–2129, 2009.

[57] Y. Shi. Lower bounds of quantum black-box complexity and degree of approximating polynomials by influence of Boolean variables. *Inf. Process. Lett.*, 75:79–83, 2000.

[58] P. Swinnerton-Dyer. The effect of twisting on the 2-Selmer group. *Mathematical Proceedings of the Cambridge Philosophical Society*, 145(3):513–526, 2008.

[59] G. Szegö. *Orthogonal Polynomials*. American Mathematical Society, 1939.

[60] Z. Tao. On the representation of a large odd integer as the sum of three almost equal primes. *Acta Mathematica Sinica, New Series*, 7(3):259–272, 1991.

[61] M. I. Tuljaganova. The Euler-Goldbach problem for an imaginary quadratic field. *Izv. Akad. Nauk UzSSR Ser. Fiz.-Mat. Nauk*, 7(1):11–17, 1963.