Equidistribution and related Ergodic methods in Number Theory

A thesis submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

Thesis Supervisor: Ritabrata Munshi

by
Mohammed Zuhair. M. M
April, 2012

Indian Institute of Science Education and Research Pune
Sai Trinity Building, Pashan, Pune India 411021

This is to certify that this thesis entitled "Equidistribution and related Ergodic methods in Number Theory" submitted towards the partial fulfillment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research Pune, represents work carried out by Mohammed Zuhair. M. M under the supervision of Ritabrata Munshi.

Mohammed Zuhair. M. M

Thesis committee:
Ritabrata Munshi
Baskar Balasubramanyam
Anupam Kumar Singh

A. Raghuram
Coordinator of Mathematics

Acknowledgments

I express my sincere thanks to Ritabrata Munshi for kindly accepting me to guide me through my thesis and making it an exciting learning experience. I thank the Tata Institute of Fundamental Research, Mumbai, for the intellectually stimulating atmosphere, and for their warm hospitality. Many thanks are also due to Prof. S. G. Dani and Dr. Baskar Balasubramanyam for the invaluable help I received from them during the course of this thesis.

It is my pleasure to thank Dr. Anupam Kumar Singh for his advice and encouragement.

Last, but not the least, I thank the Kishore Vaigyanik Protsahan Yojana for financial support.

Abstract

Equidistribution and related Ergodic methods in Number Theory

by Mohammed Zuhair. M. M

There has been a recent surge of interest in distributional problems related to number theory. Equidistribution has been widely recognized as a ubiquitous phenomenon in the subject. Solutions to equidistribution problems often involve techniques from several distinct areas of mathematics, and as such is the meeting ground of number theory, analysis and ergodic theory.

In this thesis, we study several equidistribution problems and the techniques used for their resolution. We also study some ergodic methods relevant to the subject.

In chapter 1, we introduce the notion of equidistribution and proceed to study equidistribution modulo 1 through the Weyl criterion. The Weyl criterion is an important and effective tool in proving equidistribution. We then give a generalized version of the Weyl criterion and use this to look at the distribution of Farey fractions in $[0, 1]$. It is shown that the rate of their equidistribution is intimately related to the distribution of zeros of the Riemann zeta function. The chapter ends with the study of "randomness" in the map $x \mapsto \overline{x} \pmod{p}$.

In chapter 2 we explore the Linnik's problem - a classical problem regarding the distribution of integral solutions to the equation $x^2 + y^2 + z^2 = n$ as $n \to \infty$. We will see that the problem is inextricably linked to bounds for the size of fourier coefficients of modular forms of half integral weight. We also study the Linnik's problem for squares, an easier version of the problem, using the Shimura correspondence.

In chapter 3, we begin by establishing the equidistribution of $(n^2\alpha)$ modulo 1 using ergodic theory. We then study the dynamics of unimodular lattices under the action of the diagonal torus and prove the isolation of periodic orbits. We then connect these results with the Littlewood conjecture and Minkowski's theorem on ideal classes.

Contents

Chapter 1

Equidistribution in Number theory

1.1 Introduction

The notion of equidistribution is of fundamental importance to number theory. Many results in number theory are best described as equidistribution of certain sequences in appropriate spaces. Before illustrating this, we first make the much needed definition.

Definition 1.1. Let $(X, \mathcal{B}, \mu)$ be a Borel probability space, where X is a topological space, $\mathcal{B}$ the σ-algebra of Borel sets and μ a normalized probability measure on $(X, \mathcal{B})$. A sequence of points (x_n) in X is said to be equidistributed with respect to μ if for every open set U, we have

$$\lim_{N\to\infty} \frac{\#\{n \leq N : x_n \in U\}}{N} \longrightarrow \mu(U). \tag{1.1}$$

The notion of equidistribution is a ubiquitous one in number theory. Several examples will be provided, in the course of this dissertation, to illustrate this fact. We first begin with two important examples:

Dirichlet's theorem on arithmetic progressions: Let q be any positive integer. Our space of interest is the set of units modulo q, i.e. $(\mathbb{Z}/q\mathbb{Z})^\times$, with uniform probability measure. The famous prime number number theorem of Dirichlet on arithmetic progressions is equivalent to the equidistribution of the sequence of prime numbers mod q, on $(\mathbb{Z}/q\mathbb{Z})^\times$.

Let L/K be a Galois extension of number fields with Galois group G. To every unramified prime $\mathfrak{p}$ of $\mathcal{O}_K$ we can associate a conjugacy class in G, namely the Frobenius element of $\mathfrak{p}$. Our space of interest X is the set of conjugacy classes in G with measure μ of a class proportional to its size. Now, if we arrange the (unramified) primes in K with increasing order of their norm, the equidistribution of the corresponding Frobenius elements in X with respect to μ amounts to the famous Chebotarev density theorem.

1.2 Equidistribution mod 1 and the Weyl criterion

We say a sequence of real numbers (x_n) is equidistributed modulo 1 if their fractional parts $\{x_n\}$ is equidistributed in $[0,1)$ with respect to the lebesgue measure. Let α be a real number. We are interested in the distribution of $\{n\alpha\}$, the fractional part of $n\alpha$, in $[0,1)$. One finds quickly that there is a dichotomy depending on whether α is in $\mathbb{Q}$ or not. If $\alpha \in \mathbb{Q}$, then $\{n\alpha\}$ repeats periodically. On the other hand if α is irrational, the points $\{n\alpha\}$ are all distinct. More over, by Kronecker's theorem, the points $\{n\alpha\}$ from a dense subset of $[0,1)$. What is more interesting, is that the points $\{n\alpha\}$ equidistribute in $[0,1)$ with respect to the Lebesgue measure, a fact first proved by Weyl. Originally, Weyl was interested in the the distribution modulo 1 of the sequence $x_n = f(n)$, where f is a polynomial with real coefficients. In [26] Weyl introduced a criterion, now bearing his name, which has since become a fundamental tool in establishing equidistribution results.

Before stating Weyl criterion we observe that $[0,1)$ can be identified with the unit circle $\mathbb{T}$ via the natural identification $t \mapsto e^{2\pi it}$, which also identifies the the respective Lebesgue measures. This has its advantages because $\mathbb{T}$ is a compact topological group. From now on we shall use $e(x)$ to denote $\exp(2\pi ix)$. With the above identification, it is clear that a sequence of real numbers (x_n) is equidistributed mod 1 if and only if $e(x_n)$ is equidistributed on $\mathbb{T}$.

Our first lemma, essentially a restatement of the equivalence between measure and integration, is a step towards the Weyl criterion.

Lemma 1.2. Let $(\mathbb{T}, l)$ be the unit circle with the normalized Lebesgue measure on it. A sequence (x_n) is equidistributed in $\mathbb{T}$ with respect to l if and only if, for every continuous $f : \mathbb{T} \to \mathbb{R}$, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} f(x_n) \longrightarrow \int_{\mathbb{T}} f dl. \tag{1.2}$$

Proof. ($\Leftarrow$) Suppose (x_n) is a sequence that is equidistributed in $\mathbb{T}$ and let f be a continuous real valued function on $\mathbb{T}$. Let $I_1 \cup \cdots \cup I_m$ be a finite partition of $\mathbb{T}$ into intervals. Let U and L be the upper and lower Riemann sum with respect to this partition. Now for any interval I_j in the partition we have that

$$\lim_{N\to\infty} \#\{n \leq N : x_n \in I_j\}/N \to l(I_j).$$

It follows that

$$L \leq \lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} f(x_n) \leq U.$$

Since this is true for any partition we conclude that $\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} f(x_n) \to \int_{\mathbb{T}} f dl$.

($\Rightarrow$) Let I be an open set in $\mathbb{T}$. It is enough to consider the case when I is an interval. It is easy to construct a sequence of continuous functions f_m and g_m (using Urysohn's lemma, say) such that $f_m \leq \chi_I \leq g_m$ and $\lim_{m\to\infty} \int_0^1 g_m\, dx = \lim_{m\to\infty} \int_0^1 f_m\, dx = l(I)$. It follows that

$$\frac{1}{N} \sum_{n\leq N} f_m(x_n) \leq \frac{1}{N} \sum_{n\leq N} \chi(x_n) \leq \frac{1}{N} \sum_{n\leq N} g(x_n).$$

Taking the limit $N \to \infty$, we obtain for any m,

$$\int_I f_m dl \leq \lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} \chi_I(x_n) \leq \int_I g_m dl.$$

Now, by our choice f_m and g_m, as $m \to \infty$, we have that

$$\lim_{m\to\infty} \int_I f_m dl = l(I) = \lim_{m\to\infty} \int_I g_m dl.$$

It follows that $\lim_{N\to\infty} \sum_{n\leq N} \chi_I(x_n)/N = l(I)$. But observe that,

$$\frac{1}{N} \sum_{n\leq N} \chi_I(x_n) = \frac{\#\{n \leq N : x_n \in I\}}{N}. \tag{1.3}$$

This completes the proof. □

We remark that for a general space $(X, \mathcal{B}, \mu)$ conditions (1.1) and (1.2) are not equivalent. However under mild assumptions on the space X (normality, locally compactness etc), they are equivalent. Since all spaces of our interest (smooth manifolds, homogeneous spaces for Lie groups etc) satisfy these conditions, we shall take the analog of (1.2) as our definition of equidistribution. More explicitly

Definition 1.3. Let X be a locally compact topological space and μ a Borel probability measure on it. A sequence of points (x_n) in X is said to be equidistributed with respect to μ if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} f(x_n) \longrightarrow \int_X f d\mu$$

for every $f \in C_c(X)$ (recall $C_c(X)$ is the space of compactly supported continuous functions on X).

The above definition is more useful than the former as it is easier to work with the space of continuous functions than with open sets.

In order to state the Weyl criterion in a succinct manner we introduce the little-o notation from analytic number theory. By the notation $f(n) = o(g(n))$ we mean that $\lim_{n\to\infty} f(n)/g(n) = 0$.

Theorem 1.4. (Weyl criterion) A sequence of real numbers (u_n) is equidistributed modulo 1 if and only if for every $h \in \mathbb{Z}, h \neq 0$, we have

$$\sum_{n\leq x} e(hu_n) = o(x).$$

Proof. Suppose (u_n) is equidistributed modulo 1. Observe that, for all $h \neq 0$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} \cos(2\pi h u_n) = \int_0^1 \cos(2\pi hx) dx = 0,$$

where the first equality is due to the previous lemma. A similar statement holds for $\sin(x)$. We conclude that $\sum_{n\leq x} e(hu_n) = o(x)$.

($\Rightarrow$) Conversely, suppose $\sum_{n\leq x} e(hu_n) = o(x)$ for every $h \neq 0$. Let $f : [0,1] \to \mathbb{R}$ be a continuous function. We know from basic Fourier theory that finite linear combination of trigonometric polynomials are dense in $C([0,1])$ in the uniform metric. That is, for any $\epsilon > 0$, there exists a function $p(t) = \sum_{k=-M}^{M} c_k e(kt)$ such that $|f(t) - p(t)| < \epsilon$, for all $t \in [0,1]$. From the given condition, it follows that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} p(u_n) = c_0 = \int_0^1 p(t)dt.$$

Therefore, we have that

$$\begin{aligned}\Big|\frac{1}{N}\sum_{n\leq N} f(u_n) - \int_0^1 f(t)dt\Big| \leq & \Big|\frac{1}{N}\sum_{n\leq N} f(u_n) - \frac{1}{N}\sum_{n\leq N} p(u_n)\Big| + \Big|\int_0^1 (p(t) - f(t))dt\Big| + \\ & \Big|\frac{1}{N}\sum_{n\leq N} p(u_n) - \int_0^1 p(t)dt\Big| \\ \leq \quad & 2\epsilon + \Big|\frac{1}{N}\sum_{n\leq N} p(u_n) - \int_0^1 p(t)dt\Big|\end{aligned}$$

This implies

$$\int_0^1 f(t)dt - 2\epsilon \leq \liminf_{N\to\infty} \frac{1}{N}\sum_{n\leq N} f(u_n) \leq \limsup_{N\to\infty} \frac{1}{N}\sum_{n\leq N} f(u_n) \leq \int_0^1 f(t)dt + 2\epsilon.$$

We deduce that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n\leq N} f(u_n) = \int_0^1 f(t)dt.$$

□

Corollary 1.5. If $\alpha \in \mathbb{R}$ is irrational, then $(n\alpha)$ is equidistributed modulo 1.

Proof. Let h be a non-zero integer. Since α is irrational, it follows that $e(h\alpha) \neq 1$.

$$\Big|\sum_{n\leq x} e(hn\alpha)\Big| = |e(h\alpha)|\Big|\frac{e(xh\alpha) - 1}{e(h\alpha) - 1}\Big| \leq \frac{2}{|e(h\alpha) - 1|}.$$

By Weyl criterion we obtain the required equidistribution result. □

Weyl criterion makes establishing equidistribution results considerably easy. (The above equidistribution result would have been a lot harder to prove directly). We illustrate this with yet another example (based on [10]).

Theorem 1.6. For any increasing sequence of integers $a_1, a_2, \ldots$, the sequence $\{a_n x : n \geq 1\}$ is uniformly distributed mod 1 for almost all $x \in \mathbb{R}$.

Proof. Let $h \in \mathbb{Z} \setminus \{0\}$. We have that

$$\int_0^1 \big|\frac{1}{N}\sum_{n\le N} e(ha_n x)\big|^2 dx = \frac{1}{N^2}\sum_{m,n\le N}\int_0^1 e(hx(a_m - a_n))dx = \frac{1}{N}.$$

Therefore, by replacing N by m^2 and summing over m, we get

$$\int_0^1 \sum_{m\ge 1}\big|\frac{1}{m^2}\sum_{n\le m^2} e(ha_n x)\big|^2 dx = \sum_{m\ge 1}\frac{1}{m^2} = \pi^2/6.$$

If we call the above integrand as f, we have $\int_0^1 f(x)dx < \infty$. Since f is non-negative, we must have $f(x) < \infty$ for almost all x in $[0,1]$. But note that $f(x) = f(x+1)$, so $f(x) < \infty$ for almost all x. That is, we have

$$\sum_{m\ge 1}\big|\frac{1}{m^2}\sum_{n\le m^2} e(ha_n x)\big|^2 < \infty$$

for almost all x, and hence

$$\lim_{m\to\infty}\big|\frac{1}{m^2}\sum_{n\le m^2} e(ha_n x)\big| = 0.$$

Now if $m^2 \le N < (m+1)^2$ then $\sum_{n\le N} e(ha_n x) = \sum_{n\le m^2} e(ha_n x) + O(m)$, hence

$$\frac{1}{N}\sum_{n\le N} e(ha_n x) = \frac{m^2}{N}\frac{1}{m^2}\sum_{n\le m^2} e(ha_n x) + O(\frac{m}{N})$$

and the theorem follows. □

Corollary 1.7. Almost all $x \in \mathbb{R}$ are normal.

Proof. Recall a real number α *is normal to base* $b \in \mathbb{N}$ if the sequence $(b^n\alpha)$ is equidistributed distributed mod 1. (That is, each sequence of digits appears, in the expansion of α to the base b, about as often as in a random sequence). And α is *normal* if it is normal in every base $b \ge 2$.

To see the proof of the corollary, take $a_n = b^n$ in the above theorem, by noting that the exceptional set has measure zero, as it is the countable union of measure zero sets. □

1.2.1 Weyl criterion in greater generality

If one looks carefully at the proof of Weyl criterion for equidistribution mod 1, it will become clear that the crux of the proof has to do with the fact that the linear span of the set of functions $\{e(nx) : n \in \mathbb{Z}\}$ is dense in the space of continuous functions on $\mathbb{T}$. This idea can be used in many other situations and an analogous Weyl criterion can be devised to prove equidistribution.

For concreteness, let us focus on the case of Compact metric spaces. Let X be a compact metric space and let $\mathcal{P}(X)$ be the set Borel probability measures on X. Let $C(X)$ be the space of continuous functions on X endowed with the uniform norm (i.e. $\|f\| = \sup_{x\in X}|f(x)|$).

Definition 1.8. A sequence of measures $\mu_n \in \mathcal{P}(X)$ is said to be equidistributed with respect to

$\mu \in \mathcal{P}(X)$ if they converge to μ in the weak* topology, i.e. for every $f \in C(X)$ we have

$$\mu_n(f) = \int_X f d\mu_n \to \mu(f) = \int_X f d\mu \quad \text{as } n \to \infty. \tag{1.4}$$

When this is the case, we write $\mu_n \to \mu$. (More about measures on compact metric spaces can be read in section 3.1). The Weyl criterion generalizes as follows:

Weyl criterion. Let X be a compact metric space and let $\phi_n \in C(X)$ be a sequence of functions with the property that their linear combination is dense in $C(X)$. Then $\mu_n \to \mu$ if and only if $\mu_n(\phi_m) \to \mu(\phi_m)$ for all $m \in \mathbb{N}$.

Proof. ($\Rightarrow$) Let $f \in C(X)$. Given $\epsilon > 0$ there exists $g = \sum_{k=1}^{N} c_k \phi_k$ such that $\|f - g\| < \epsilon$. (We assume c_k's are non-zero). Write $f - g = h$. Then

$$|\nu(h)| < \epsilon \text{ for any } \nu \in \mathcal{P}(X).$$

Now,

$$\begin{aligned} |\mu_n(f) - \mu(f)| &= |\mu_n(g+h) - \mu(g+h)| = |\mu_n(g) + \mu_n(h) - \mu(g) + \mu(h)| \\ &\leq |\mu_n(g) - \mu(g)| + 2\epsilon = \Big|\sum_{k=1}^{N} c_k \mu_n(\phi_k) - \sum_{k=1}^{N} c_k \mu(\phi_k)\Big| + 2\epsilon \end{aligned}$$

Let $c = \max\{|c_1|, \ldots, |c_N|\}$. Now, for each $k \in \{1, 2, \ldots, N\}$ there exists an $M(k) \in \mathbb{N}$ such that $|\mu_n(\phi_k) - \mu(\phi_k)| < \epsilon/cN$ for all $n > M(k)$. Let $M = \max\{M(1), \ldots, M(N)\}$. Then, for $n > M$ we have

$$\sum_{k=1}^{N} |c_k| \, |\mu_n(\phi_k) - \mu(\phi_k)| < \epsilon$$

and therefore $|\mu_n(f) - \mu(f)| < 3\epsilon$. Thus we have shown $\mu_n(f) \to \mu(f)$ for every $f \in C(X)$. □

Let us look at some examples of spaces for which nice Weyl criterion exists. By the Weyl criterion, as in the above form, there could be plenty of choices for the system of test functions ϕ_n; any set that generates a dense subset of $C(X)$ will do, for example any orthonormal basis for $L^2(X, \mu)$. But very often there is a natural choice. The Stone-Weierstrass theorem provides us with a tool to find such a system of functions. We recall this theorem, which, we shall have the opportunity to invoke later.

Stone-Weierstrass theorem. Let X be a compact metric space and let $\mathcal{A} \subset C(X)$ be a linear subspace with the following properties:

- $\mathcal{A}$ is closed under multiplication.
- $\mathcal{A}$ contains the constant functions.
- $\mathcal{A}$ separates points in X.

Then $\mathcal{A}$ is dense in $C(X)$.

Now, let us look again at the example of $\mathbb{T}$. What is so special about the functions $e(nx)$? $\mathbb{T}$ is a compact abelian Lie group and $e(nx)$ are precisely the characters of irreducible representations

(over $\mathbb{C}$) of $\mathbb{T}$. They form an orthonormal basis for $L^2(\mathbb{T}, l)$, the space of square integrable functions with respect to l.

More generally, let G be a compact Lie group. The famous Peter-Weyl theorem asserts that matrix coefficients of irreducible representations of G, form an orthonormal basis for $L^2(G)$. Similarly for $X = G^{\#}$, the space of conjugacy classes of G, the characters of irreducible representations of G form an orthonormal basis for $L^2(X, dg)$ where dg is the measure derived from the Haar measure on G. Also, the integral of any non-trivial character over X is zero. If (u_n) is a sequence of points on X, the Weyl criterion then reads:

$$\sum_{n \leq x} Tr(\rho(u_n)) = o(x)$$

for all irreducible representations ρ of G.

Although we won't have the occasion to work with general Lie groups, it is important to note that the above choices of test functions provides a potent tool for establishing equidistribution results.

1.3 Equidistribution of rationals and the Riemann hypothesis

Any rational number α can be uniquely represented as a fraction a/b in it lowest terms, that is, a pair of integers a and b with $b > 0$ and the greatest common divisor $(a, b) = 1$. We define the height of α by

$$H(\alpha) = \max\{|a|, b\}.$$

For each $Q > 0$, we can look at the rational numbers in $[0, 1)$ of height at most Q. Let us denote them by $x_1, \ldots, x_N$, where the x_i's are arranged in the increasing order of their magnitude. As we shall soon see, the distribution these numbers is of fundamental importance. But before that we need a lemma.

Lemma 1.9. (Ramanujan sum)

$$\sum_{\substack{a=1 \\ (a,q)=1}}^{q} e\Big(\frac{ah}{q}\Big) = \sum_{\substack{c|h \\ c|q}} c\mu\Big(\frac{q}{c}\Big),$$

where $\mu(n)$ is the Möbius function.

Proof. We have that $\sum_{d|n} \mu(d) = 0$ for $n > 1$. Also the common factors of (a, q) are the common factors of a and q. Therefore the required sum is

$$\sum_{a=1}^{q} \sum_{\substack{d|a \\ d|q}} \mu(d) e\Big(\frac{ah}{q}\Big) = \sum_{d|q} \mu(d) \sum_{b=1}^{q/d} e\Big(\frac{bh}{q/d}\Big),$$

where we have put $a = bd$. The sum over b is zero unless h is a multiple of q/d. Writing $c = q/d$,

we get the sum to be

$$\sum_{\substack{d|q\\ q/d|h}} \mu(d)\frac{q}{d} = \sum_{c|h}\sum_{\substack{d\\ cd=q}} c\mu(d) = \sum_{\substack{c|h\\ c|q}} c\mu\Big(\frac{q}{c}\Big).$$

□

Lemma 1.10. (Weyl sum for the rational numbers) Let Q be a positive integer and let $x_1, \dots, x_N$ be the rational numbers of height at most Q in $[0, 1)$. Then

$$N = \frac{Q^2}{2\zeta(2)} + O(Q\text{loq}Q),$$

and for each integer h

$$\sum_{1}^{N} e(hx_n) = \sum_{d|h} dM(Q/d), \tag{1.5}$$

where $M(x) = \sum_{n\le x}\mu(n)$ is the sum function of the Möbius function.

Proof. Put $h = 0$ in the previous lemma. We get,

$$\begin{aligned} N &= \sum_{q=1}^{Q}\sum_{\substack{a=1\\(a,q)=1}} 1 = \sum_{q=1}^{Q}\sum_{cd=q} c\mu(d)\\ &= \sum_{q=1}^{Q}\sum_{d|q}\frac{q}{d}\,\mu(d) = \sum_{d\le Q}\frac{\mu(d)}{d}\sum_{\substack{q=1\\ q\equiv 0(\mathrm{mod}\, d)}}^{Q} q\\ &= \sum_{d\le Q}\frac{\mu(d)}{d}\left[\frac{1}{2}\frac{Q^2}{d} + O(Q)\right]\\ &= \frac{Q^2}{2}\sum_{d=1}^{\infty}\frac{\mu(d)}{d^2} + O\Big[Q^2\sum_{d=Q+1}^{\infty}\frac{|\mu(d)|}{d^2} + Q\sum_{d\le Q}\frac{|\mu(d)|}{d}\Big]. \end{aligned}$$

Now, it is easy to see that,

$$\sum_{d=Q+1}^{\infty}\frac{|\mu(d)|}{d^2} = \int_Q^{\infty}\frac{dx}{x^2} + O(\frac{1}{Q^2}) = \frac{1}{Q} + O(\frac{1}{Q^2}),$$

and

$$\sum_{d\le Q}\frac{|\mu(d)|}{d} \le \sum_{d\le Q}\frac{1}{d} = \log Q + O(1).$$

This completes the proof of the first assertion.

For the second assertion, the sum can be rearranged to give

$$\sum_{q=1}^{Q} \sum_{\substack{a=1 \\ (a,q)=1}}^{q} e\Big(\frac{ah}{q}\Big) = \sum_{q=1}^{Q} \sum_{\substack{d|h \\ d|q}} d\mu\Big(\frac{q}{d}\Big)$$
$$= \sum_{d|h} d \sum_{r \le Q/d} \mu(r) = \sum_{d|h} dM(Q/d).$$

□

We had previously defined the notion of equidistribution of a sequence of points in a space X. Similarly one can define the notion of equidistribution for a sequence of finite subsets of X. Given a finite subset E of X we define $\mu_E \in \mathcal{P}(X)$ as

$$\mu_E = \frac{1}{|E|} \sum_{x \in E} \delta_x$$

where δ_x is the Dirac measure at x.

Definition 1.11. A sequence of finite subsets $E_n \subset X$, is said to be equidistributed with respect to μ if $\mu_{E_n} \to \mu$ as $n \to \infty$.

It is time to introduce yet another notation from analytic number theory. Let f be a complex valued function on $\mathbb{N}$ and g a non-negative real valued function on $\mathbb{N}$. By the notation $f \ll g$, we mean that there is a $c > 0$ such that $|f(n)| \le cg(n)$, for all n large enough.

Theorem 1.12. Let E_Q be the set of rationals of height at most Q in $[0, 1)$. Then, as $Q \to \infty, E_Q$ is equidistributed in $[0, 1)$ with respect to the Lebesgue measure.

Proof. Let $x_1, \ldots, x_N$ correspond to the complete set rationals in $[0, 1)$ of height at most Q. Using the trivial bound $|M(x)| \le x$ in (1.5) we get, for non-zero h,

$$\sum_{1}^{N} e(hx_n) \ll \sum_{d|h} Q \ll d(h)N^{1/2},$$

where $d(h)$ is the number of divisors of h. That is (generalized) Weyl criterion holds true for each non-zero h. The theorem follows. □

We now prove a lemma on Dirichlet series which we shall have multiple occasions to use.

Lemma 1.13. Let $f(s) = \sum_{n \ge 1} a_n n^{-s}$ be a general Dirichlet series. Let $A(x) = \sum_{x \le n} a(n)$. Suppose $A(x) = O(x^{\alpha+\epsilon})$ for every $\epsilon > 0$. Then the Dirichlet series $f(s)$ converges for any s with $\operatorname{Re} s > \alpha$.

Proof. This follows directly from Abel's partial summation formula

$$\sum_{n \le x} \frac{a(n)}{n^s} = \frac{A(x)}{x^s} + s \int_1^x \frac{A(u)}{u^{1+s}} du$$

by noting that if $\mathrm{Re}\, s > \alpha$, as $x \to \infty$, the first term in the right side goes to zero by assumption, and the integral above is (absolutely) convergent. □

We now state an important theorem in the converse direction, which we shall use later.

Wiener-Ikehara theorem. Let f be a non-negative, non-decreasing real valued function on $[1, \infty)$ and suppose that the Mellin transform

$$g(s) := s \int_1^\infty f(x) x^{-(s+1)} dx$$

exists for $\mathrm{Re} s > 1$. Also, suppose that for some constant α, the function

$$g(s) - \frac{\alpha}{s-1}$$

has continuous extension to the closed half-plane $\mathrm{Re}\, s \geq 1$. Then

$$\lim_{x \to \infty} \frac{f(x)}{x} = \alpha.$$

Remark 1.14. Of special interest is the case when $f(x) = A(x)$, the sum function of a sequence of non-negative integers a_n. Suppose that the corresponding Dirichlet series $\phi(s) = \sum_{n \geq 1} a_n n^{-s}$ converges on the half-plane $\mathrm{Re}\, s > 1$ and has analytic continuation except for a simple pole at $s = 1$ with residue α. First, note that $\phi(s)$ is precisely the Mellin transform of $A(x)$. To see this, in the Abel's formula quoted above, take $s = \sigma > 1$. Then, all the terms involved are positive. Since the series converge for any $\sigma > 1$, we must have that $A(x)/x^\sigma$ goes to zero as $x \to \infty$, for any $\sigma > 1$.

Now, we can apply the Wiener-Ikehara theorem to deduce that $\lim_{x \to \infty} A(x)/x = \alpha$. If $\phi(s)$ has analytic continuation with a simple pole at $s = b$ with residue α, we can shift s to $s - b$ and can apply the Wiener-Ikehara theorem to get

$$A(x) \sim \frac{\alpha x^b}{b}.$$

We now make some remarks pertinent to the title of this subsection. From the general theory of Dirichlet series mentioned above we know that

$$M(x) \ll x^{\theta + \epsilon} \tag{1.6}$$

holds for all $\epsilon > 0$ if and only if the Dirichlet series for $1/\zeta(s)$ converges in the half plane $\mathrm{Re}(s) > \theta$, that is when $\zeta(s)$ is non-zero in the region $\mathrm{Re}(s) > \theta$. Now, if (1.6) holds true for some $\theta \geq 1/2$ ($\theta < 1/2$ is ruled out by the existence of zeros of $\zeta(s)$ on the critical line $1/2 + it$), we see using (1.5) that

$$\sum_1^N e(h x_n) \ll \sum_{d|h} d (Q/d)^{\theta + \epsilon} \ll f(h) N^{\theta/2 + \epsilon/2},$$

where $f(h) = \sum_{d|h} d^{1-\theta-\epsilon}$ (a function polynomially bounded in h). Therefore, the extent to which the rationals are equidistributed depends on the truth or falsity of the Riemann hypothesis. Quoting

from [12], "...perhaps this is what the Riemann hypothesis really means".

1.4 The map $x \mapsto \overline{x} \bmod p$

[1]Take a large prime. The function inverse modulo p in the interval $[1, 2, \ldots, p-1]$ has certain "randomness" which is of great importance in analytic number theory (see [11]), and is also exploited in cryptography. We shall explain what this "randomness" is using the notion of effective equidistribution.

Let $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. Through out this section, $\overline{x}$ will mean the inverse of x mod p. Define

$$S(p) = \left\{ \left(\frac{x}{p}, \frac{\overline{x}}{p} \right) : x \in (\mathbb{Z}/p\mathbb{Z})^\times \right\} \subseteq \mathbb{T}^2$$

and let

$$\nu_p = \frac{1}{p-1} \sum_{w \in S(p)} \delta_w$$

be the uniform measure on $S(p)$. Let λ be the normalized Lebesgue measure on $\mathbb{T}^2$. We shall prove that

$$\nu_p \to \lambda \text{ as } p \to \infty. \tag{1.7}$$

The above limit is taken over primes p and the convergence is in the weak* topology.

The qualitative (1.7) is an immediate consequence of the following theorem

Theorem 1.15. Let f be a smooth function on $\mathbb{T}^2$, define $\mathcal{S}(f)$ by

$$\mathcal{S}(f)^2 = \|f\|_2^2 + \|\partial^2 f/\partial x^2\|_2^2 + \|\partial^2 f/\partial y^2\|_2^2.$$

Then there exists a $\kappa > 0$ such that for all $f \in C^\infty(\mathbb{T}^2)$ one has

$$|\nu_p(f) - \int f d\lambda| \ll p^{-\kappa} \mathcal{S}(f). \tag{1.8}$$

We shall prove theorem 1.15 with the exponent $\kappa = 1/4$.

Clearly, (1.8) is an effective version of (1.7), as any qualitative statement that can be deduced from (1.7) can be made quantitative using (1.8). For example from (1.7), we can conclude that the set $S(p)$ intersects , the box $[0.5, 0.51] \times [0.7, 0.71]$ for p sufficiently large, whereas using (1.8) we can compute how large p should be for this to happen.

We say a sequence of measures μ_n is *effectively equidistributed* with respect to an ambient measure λ when an estimate of the type (1.8) holds true.

Kloosterman Sums. Since f is a smooth function on $\mathbb{T}^2$, write f in its Fourier series.

$$f(x, y) = \sum_{\mathbf{n} \in \mathbb{Z}^2} \hat{f}(\mathbf{n}) e_n(x, y), \tag{1.9}$$

[1]Based on a lecture by Farrell Brumely, at the Summer School on Analytic Questions in Arithmetic (AQUA), 2010.

where if $\mathbf{n} = (n_1, n_2)$ we have put

$$e_{\mathbf{n}}(x, y) = e(n_1 x + n_2 y).$$

A direct computation using the above Fourier expansion will show that the Sobalev norm of f as defined above can be written as

$$\mathcal{S}(f)^2 = \sum_{\mathbf{n}} (1 + (2\pi)^2 \|n\|^4) |\hat{f}(\mathbf{n})|^2. \tag{1.10}$$

As f is continuously differentiable, the series (1.9) is absolutely convergent. We may therefore interchange the order of summation to obtain

$$\nu_p(f) = \sum_{\mathbf{n}} \hat{f}(\mathbf{n}) \hat{\nu}_p(\mathbf{n}), \tag{1.11}$$

where $\hat{\nu}_p(\mathbf{n}) = \nu_p(e_{\mathbf{n}})$, is the Fourier coefficient of ν_p at the harmonic $e_{\mathbf{n}}$. More explicitly we have,

$$\hat{\nu}_p(\mathbf{n}) = \int e_{\mathbf{n}}(w) d\nu_p(w) = (p-1)^{-1} \sum_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} e\Big(\frac{n_1 x + n_2 \overline{x}}{p}\Big).$$

The last sum is called the *classical Kloosterman sum.*

Note that $\hat{f}(0) = \int_{\mathbb{T}^2} f d\lambda$ and also that if $\mathbf{n} \in p\mathbb{Z}^2$ then $\hat{\nu}_p(\mathbf{n}) = 1$. Therefore the series (1.11) may be rewritten as

$$\nu_p(f) - \int_{\mathbb{T}^2} f d\lambda = \sum_{\mathbf{n} \notin p\mathbb{Z}^2} \hat{f}(\mathbf{n}) \hat{\nu}_p(\mathbf{n}) + E(p), \tag{1.12}$$

where

$$\begin{aligned} E(p) = \sum_{\mathbf{n} \in p\mathbb{Z}^2 \setminus \{0\}} \hat{f}(\mathbf{n}) &= \sum_{\mathbf{n} \in p\mathbb{Z}^2 \setminus \{0\}} \frac{1}{\|n\|^2} \|n\|^2 \hat{f}(\mathbf{n}) \\ &\ll p^{-2} \Big(\sum_{\mathbf{n} \in \mathbb{Z}^2} \|n\|^4 |\hat{f}(\mathbf{n})|^2 \Big)^{1/2} \leq p^{-2} \mathcal{S}(f). \end{aligned}$$

We have used the Cauchy-Schwartz inequality and (1.10) to arrive at the above result.

Now, as we shall see shortly , and as first demonstrated by Kloosterman in [15], if $\mathbf{n} \notin p\mathbb{Z}^2$ then

$$\hat{\nu}_p(\mathbf{n}) \ll p^{-1/4} \tag{1.13}$$

uniformly in $\mathbf{n}$. Inserting this in (1.12), we get

$$|\nu_p(f) - \int_{\mathbb{T}^2} f d\lambda| \ll p^{-1/4} \sum_{\mathbf{n} \in \mathbb{Z}^2} |\hat{f}(\mathbf{n})| + p^{-2} \mathcal{S}(f).$$

But $\sum_{\mathbf{n}\in\mathbb{Z}^2} |\hat{f}(\mathbf{n})| \leq \mathcal{S}(f)$ using the Cauchy-Swartz trick as above. Hence we have

$$|\nu_p(f) - \int_{\mathbb{T}^2} f d\lambda| \ll p^{-1/4}\mathcal{S}(f).$$

Therefore theorem 1.15 follows from Kloosterman's estimate (1.13).

Proof of Kloosterman bound (1.13)

Let $\mathbf{n} = (n_1, n_2)$, be such that $\mathbf{n} \notin p\mathbb{Z}^2$. Put $K_p(\mathbf{n}) = (p-1)\hat{\nu}_p(\mathbf{n})$. If a is relatively prime to p, then as x varies through residues modulo p, ax also varies through residues modulo p. Therefore we have,

$$K_p(n_1, n_2) = \sum_{x=1}^{p-1} e(\frac{n_1 ax + n_2\overline{ax}}{p}) = K_p(an_1, \overline{a}n_2).$$

Now, if $p|n_2$ then $p \nmid n_1$, in which case $K_p(\mathbf{n}) = 0$. Therefore we may assume $p \nmid n_1, n_2$, so that the sum

$$M_4(p) = \sum_{r=0}^{p-1}\sum_{s=0}^{p-1} |K_p(r,s)|^4$$

contains $p-1$ copies of $|K_p(n_1, n_2)|^4$. Hence,

$$(p-1)|K_p(n_1, n_2)|^4 \leq M_4(p). \tag{1.14}$$

Now, we may expand $|K_p(r,s)|^4$ as

$$\sum_{m_1=1}^{p-1}\sum_{m_2=1}^{p-1}\sum_{m_3=1}^{p-1}\sum_{m_4=1}^{p-1} e(\frac{rA+sB}{p}),$$

where $A = m_1 + m_2 - m_3 - m_4,\ B = \overline{m}_1 + \overline{m}_2 - \overline{m}_3 - \overline{m}_4$.

On rearranging the order of summation, it follows that

$$M_4(p) = \sum_{m_i}\sum_{r,s} e(\frac{rA+sB}{p}) = \sum_{m_i}\{\sum_r e(\frac{rA}{p})\}\{\sum_s e(\frac{sB}{p})\}.$$

The innermost sums are easy to evaluate. In fact we have

$$\sum_{r=0}^{p-1} e(\frac{rA}{p}) = \begin{cases} 0, p \nmid A, \\ p, p \mid A, \end{cases}$$

and similarly for the sum over s. We therefore conclude that

$$M_4(p) = p^2.\#\{(m_1, m_2, m_3, m_4) : p \mid A, B\}.$$

Let $m_1 + m_2 = m_3 + m_4 \pmod{p}$ and $\overline{m}_1 + \overline{m}_2 = \overline{m}_3 + \overline{m}_4 \pmod{p}$. Multiplying the two equations we get $m_1\overline{m}_2 + \overline{m}_1 m_2 = m_3\overline{m}_4 + \overline{m}_3 m_4 \pmod{p}$. Put $\alpha = m_1\overline{m}_2$ and $\beta = m_3\overline{m}_4$. We have $\alpha + \overline{\alpha} = \beta + \overline{\beta} = c \pmod{p}$ (say), i.e. α and β both satisfy the quadratic equation $x^2 - cx + 1 = 0 \pmod{p}$. We conclude that $\alpha = \beta$ or $\alpha = \overline{\beta}$. Consider the case $\alpha = \beta$. Then the

equations $m_1 + m_2 = m_3 + m_4 \mod p$ and $m_1\overline{m}_2 = m_3\overline{m}_4 (\mod p)$ holds true. Multiplying both, we get, $m_1(m_1\overline{m}_2 + 1) = m_3(m_3\overline{m}_4 + 1) (\mod p)$. This implies that either $m_1 = m_3 (\mod p)$ or $(m_1\overline{m}_2 + 1) = (m_3\overline{m}_4 + 1) = 0 (\mod p)$. The latter is equivalent to $m_1 + m_2 = m_3 + m_4 = 0 \mod p$. The case $\alpha = \overline{\beta}$ can be treated similarly.

We have just shown that if $p \mid A, B$ then either m_3, m_4 is a permutation of m_1, m_2 or $m_1 + m_2 = m_3 + m_4 = 0 (\mod p)$. Thus there are at most $3(p-1)^2$ available sets of values for (m_1, m_2, m_3, m_4). It follows that

$$M_4(p) \leq 3p^2(p-1)^2 < 3p^3(p-1).$$

We now deduce from (1.14) that

$$|K_p(\mathbf{n})| < 3^{1/4}p^{3/4} \quad (\mathbf{n} \notin p\mathbb{Z}^2).$$

The Kloosterman bound (1.13) follows.

Chapter 2

Linnik's Problem

Let $\alpha = (x_1, x_2, x_3) \in \mathbb{Z}^3$ with $|\alpha|^2 = x_1^2 + x_2^2 + x_3^2$. For $n \in \mathbb{Z}^+$ the set

$$V_n = \{x = \alpha/|\alpha| : \alpha \in \mathbb{Z}^3; |\alpha|^2 = n\}$$

lies on the unit sphere S^2. As was first observed by Legendre, V_n is non-empty if and only if $n \neq 4^a(8b+7)$, for a, b integers, a non-negative. In [18] Linnik asked if the set V_n gets equidistributed on S^2 with respect to the Lebesgue measure $d\sigma$ as $n \to \infty$, subject to the condition that $n \equiv 1, 2, 3, 5, 6 (\text{mod } 8)$. He was able to prove this using an "ergodic method", under an additional hypothesis on n that $\left(\frac{n}{p}\right) = 1$, for a small fixed prime p. Much later (in 1987) Iwaniec made a breakthrough in the estimation of Fourier coefficients of modular forms of half-integral weight in [13], which allowed this condition to be removed.

We shall see soon that modular forms enters the picture via the Weyl criterion.

What is a good choice for the system of test functions to apply Weyl criterion? One should be looking for an orthonormal basis for $L^2(S^2)$. We shall see that there is indeed a 'natural' choice for such a basis. We begin by noting the following; if we let $e(\theta) = (x + iy)/|x + iy$, then for $m > 0$

$$e(m\theta) = \Big(\frac{x + iy}{|x + iy|}\Big)^m$$

and

$$e(-m\theta) = \Big(\frac{x - iy}{|x - iy|}\Big)^m.$$

Now $(x+iy)^m$ and $(x-iy)^m$ are homogeneous harmonic polynomials on $\mathbb{R}^2$. This example generalizes nicely to $\mathbb{R}^3$ (in fact to $\mathbb{R}^n$). Let $\mathcal{H} \subset C(S^2)$ be the subspace of finite sum of homogeneous harmonic polynomials in $\mathbb{R}^3$ restricted to S^2. Clearly, $\mathcal{H}$ is multiplicatively closed and contains the constant functions. Further, it can be shown that they separate points in S^2. Therefore, by the Stone-Weierstrass theorem $\mathcal{H}$ is dense in $C(S^2)$. Also, since they are eigen functions of the spherical laplacian, their integral over S^2 with respect to the Lebesgue measure σ is zero.

Let P be a homogeneous harmonic polynomial of degree $l \geq 1$. From the preceding discussion, in order to show a sequence of measures μ_n equidistribute to σ, it is sufficient to show that $\mu_n(P) \to 0$

as $n \to \infty$ for every homogeneous harmonic polynomial P. Linnik's problem asks if $\mu_{V_n} \to \sigma$ as $n \to \infty$ through admissible values of n. By Weyl criterion it is sufficient to show

$$\frac{1}{\#V_n} \sum_{x \in V_n} P(x) \to 0 \text{ as } n \to \infty$$

through admissible values. Equivalently, we require

$$\sum_{\substack{\alpha \in \mathbb{Z}^3 \\ |\alpha|^2 = n}} P\Big(\frac{\alpha}{|\alpha|}\Big) = o(r_3(n)) \tag{2.1}$$

where $r_3(n) = \#\{\alpha \in \mathbb{Z}^3 : |\alpha|^2 = n\} = |V_n|$.

The above bound is established by noting that the left side of (2.1) is essentially the fourier coefficient of a modular form. Define

$$\theta_P(z) = \sum_{\alpha \in \mathbb{Z}^3} P(\alpha) e(|\alpha|^z) = \sum_{n=1}^{\infty} r(n, P) e(nz).$$

Shimura in [23] proves the following theorem (in a vastly more general form).

Theorem 2.1. The function $\theta_P(z)$ is a holomorphic cusp form of weight $3/2 + l$ for $\Gamma_0(4)$. Also $\theta_P(z) = 0$, for l odd.

Proof. See [23]. □

Note that

$$r(n, P) = \sum_{|\alpha|^2 = n} P(\alpha).$$

Since P is homogeneous of degree l, we have $P(\alpha/|\alpha|) = |\alpha|^{-l} P(\alpha)$, so that

$$r(n, P) = n^{l/2} \sum_{|\alpha|^2 = n} P\Big(\frac{\alpha}{|\alpha|}\Big). \tag{2.2}$$

To prove (2.1), we will also need bounds on $r_3(n)(= |V_n|)$. It was Gauss (in [9]) who first discovered some remarkable algebraic structure in the set V_d, when d is square free. In modern language, he had proved that the ideal class group of the quadratic order $\mathbb{Z}[\sqrt{-d}]$ acts transitively on the quotient $SO_3(\mathbb{Z})\backslash V_d$. (See [7] for the details, and for an exposition of Linnik's original approach in modern language). From this, Gauss was able to prove

$$r_3(n) = \frac{24 h(d)}{w(d)} \Big(1 - \Big(\frac{d}{2}\Big)\Big), \tag{2.3}$$

where $d = \text{disc}(\mathbb{Q}(\sqrt{-n}))$, $h(d)$ the class number of $\mathbb{Q}(\sqrt{-n})$ and $w(d)$ the number of roots of unity in this field. $\big(\frac{d}{2}\big)$, of course, is the quadratic symbol. Now, it is a famous (alas non-effective!) theorem of Siegel that $h(d) \gg_\epsilon |d|^{1/2-\epsilon}$. It follows that

$$r_3(n) \gg_\epsilon n^{1/2-\epsilon}. \tag{2.4}$$

Suppose we had a theorem of the following form: If $f(z) = \sum a_n e(nz)$ is a cusp form of half integral weight k for $\Gamma_0(N)$ (where $4 \mid N$), then there exists a $\delta > 0$ such that for each n we have

$$|a_n| \ll n^{k/2-1/4-\delta}. \tag{2.5}$$

Then with $k = 3/2 + l$, combining (2.2), (2.4) and (2.5), we would have

$$\frac{1}{r_3(n)} \sum_{|\alpha|^2=n} P(\alpha/|\alpha|) = O_\epsilon(n^{-\delta+\epsilon}) \tag{2.6}$$

and Linnik's conjecture would follow. In fact, Iwaniec [13] had proven precisely an estimate of type (2.6) with $\delta = 1/28$. We shall settle for a weaker estimate with $\delta = 1/222$ (as given in [14]). We shall see later that the bound (2.5), as stated, is false for general n. (One has to make an assumption like n is square; but we shall ignore it for the time being and return to it later.)

The bound (2.5), may at first, seem out of the blue. But we shall, in due course, explain the naturality in (asking for) it.

At this stage, rather than embarking on proving Iwaniec's bound, we shall consider an ε-modification of the Linnik's problem which will (hopefully) shed some light on the original problem. Our presentation is based on [4].

2.1 Rational points on the sphere

Let $\mathcal{R}$ be the set of rational points on the unit sphere, i.e. $\mathcal{R} = \mathbb{Q}^3 \cap S^2$. Define the height $h(x)$ of a rational point $x \in \mathcal{R}$ as the least common denominator of its coordinates in reduced form. We shall show that the rational points of height $\leq T$ become equidistributed on S^2 with respect to σ as $T \to \infty$.

For a function ρ on S^2, define

$$A(T, \rho) = \sum_{\substack{x \in \mathcal{R} \\ h(x) \leq T}} \rho(x).$$

Thus $A(T, 1)$ is the number of rational points on on S^2 of height $\leq T$.

Theorem 2.2. As $T \to \infty$, $A(T, 1) \sim \frac{3}{2\kappa} T^2$, where $\kappa = 1/1^2 - 1/3^2 + 1/5^2 - 1/7^2 + \cdots \simeq 0.9159$ is the Catalan's constant. For any continuous function $\rho : S^2 \to \mathbb{C}$ we have

$$A(T, \rho)/A(T, 1) \to \int_{S^2} \rho \, d\sigma \quad \text{as } T \to \infty.$$

Proof. As before we may restrict our attention to homogeneous harmonic polynomials. Let P be a such a polynomial of degree $l \geq 0$ (if $l = 0$ then P is just the constant function 1). Define

$$a(n, P) = \sum_{\substack{x \in \mathcal{R} \\ h(x)=n}} P(x)$$

and consider the Dirichlet series

$$\phi(s,P)=\sum_{n\geq 1}a(n,P)n^{-s}.$$

We shall soon see that this Dirichlet series has analytic continuation and a functional equation, owing to the fact that $a(n,P)$'s are related to $r(n,P)$'s, which are Fourier coefficients of modular forms. We now derive this relation. Put $b(n,P)=r(n,P)n^{-l/2}$, i.e. the sum in (2.2). Consider the set of integral vectors

$$V=\{(x_1,x_2,x_3,y)\in\mathbb{Z}^4 : x_1^2+x_2^2+x_3^2=y^2, y>0 \text{ and } gcd(x_1,x_2,x_3,y)=1\}$$

and observe that the map

$$(x_1,x_2,x_3,y)\to(x_1/y,x_2/y,x_3/y)$$

gives a bijection from V onto $\mathcal{R}$, where y is the height of the image of (x_1,x_2,x_3,y). It follows that

$$b(n^2,P)=\sum_{d|n}a(d,P)$$

which on Möbius inversion gives

$$a(n,P)=\sum_{d|n}b(d^2,P)\mu(n/d).$$

This is equivalent to the following identity of Dirichlet series

$$\phi(s,P)=\zeta(s)^{-1}\sum_{n\geq 1}b(n^2,P)n^{-s}. \tag{2.7}$$

For the first part of the theorem, put $P=1$, and note that $b(n^2,1)=r_3(n^2)$. By a classical result of Hurwitz we have the following identity

$$\sum_{n\geq 1}b(n^2,1)n^{-s}=6(1-2^{1-s})\frac{\zeta(s)\zeta(s-1)}{L(s,\chi_{-4})}$$

where $\chi_{-4}(p)=\left(\frac{-4}{p}\right)$ is the Kronecker symbol. This gives

$$\phi(s,1)=6(1-2^{1-s})\frac{\zeta(s-1)}{L(s,\chi_{-4})}, \tag{2.8}$$

which is holomorphic for $\mathrm{Re}(s)>1$, except for a simple pole at $s=2$ with residue $3/\kappa$, where $\kappa=L(2,\chi_{-4})^{-1}$. By applying the Wiener-Ikehara theorem (see remark 1.14) we derive the asymptotic relation

$$A(T,1)\sim\frac{3}{2\kappa}T^2 \quad \text{as } T\to\infty.$$

To finish the proof of the theorem, by Weyl criterion, we need to show that

$$T^{-2}A(T,P) = T^{-2}\sum_{n\leq T} a(n,P) \to 0 \tag{2.9}$$

as $T \to \infty$ for any P with degree $l > 0$. Furthermore, we may assume that l is even (as for odd l, $P(-x) = -P(x)$ and $A(T,P) = 0$). The analog of Hurwitz's result (2.8) for a general P is given by the Shimura's correspondence. In [23], Shimura introduced a family of correspondence between modular forms of half-integral weight and modular forms of even integral weight. We state his theorem and specialize it to our case.

Shimura Map. Let t be a positive square-free integer, and suppose $f(z) = \sum_{n=1}^{\infty} a(n)e(nz) \in S_{k+1/2}(\Gamma_0(4N),\psi)$, where k is a positive integer. If the numbers $A(n)$ are defined by

$$\sum_{n\geq 1} A(n)n^{-s} = L(s-k+1,\psi\chi_4^k\chi_t)\sum_{n\geq 1} a(tn^2)n^{-s}, \tag{2.10}$$

where $\chi_t = \left(\frac{t}{\cdot}\right)$ is the usual Kronecker symbol modulo t, then $F(z) = \sum_{n=1}^{\infty} A(n)e(nz) \in M_{2k}(2N,\psi^2)$. Moreover, if $k > 1$ then F(z) is a cusp form.

We shall take $t = 1$, $f(z) = \theta_P(z) = \sum_{n\geq 1} r(n,P)e(nz)$. In this case $N = 1, \psi = 1$ and $k = l+1$. Since l is even, k is odd and $\chi_4^k = \chi_4$.

$A(n)$'s are defined by

$$\sum_{n\geq 1} A(n)n^{-s} = L(s-l,\chi_4)\sum_{n\geq 1} r(n^2,P)n^{-s}.$$

Substituting $b(n^2,P)n^l$ for $r(n^2,P)$ and using (2.7), we get

$$\sum_{n\geq 1} A(n)n^{-s} = L(s-l,\chi_4)\phi(s-l,P)\zeta(s-l). \tag{2.11}$$

Shimura's theorem says, $F(z) := \sum_{n\geq 1} A(n)e(nz)$ is a cusp form of weight $2l+2$ for $\Gamma_0(2)$. If we define the (normalized) L-function associated F by

$$L(s,F) = \sum_{n\geq 1} A(n)n^{-l-\frac{1}{2}}n^{-s}, \tag{2.12}$$

it is clear from (2.12) that

$$\phi(s,P) = \frac{L(s-1/2,F)}{\zeta(s)L(s,\chi_4)}. \tag{2.13}$$

Now, by the famous Deligne bound (originally a conjecture by Ramanujan),

$$A(n) \ll_\epsilon n^{l+\frac{1}{2}+\epsilon}$$

for any $\epsilon > 0$. It follows from lemma 1.13 that the Dirichlet series (2.12) converges absolutely for $\mathrm{Re}(s) > 1$, hence that for $\phi(s,P)$ in (2.13) converges absolutely for $\mathrm{Re}(s) > 3/2$. A consideration

using Abel's summation formula, similar to the one in remark 1.14, will imply that

$$\sum_{n\leq T} a(n,P) \ll_\epsilon T^{\frac{3}{2}+\epsilon}$$

for any $\epsilon > 0$. This completes the proof, as we have proved (2.9). □

2.2 Modular forms of half integral weight

Jacobi's theta series

$$\theta(z) = \sum_{n\in\mathbb{Z}} e(n^2 z),$$

has the following remarkable transformation property for any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(4)$:

$$\theta(\gamma z) = \left(\frac{c}{d}\right) \varepsilon_d^{-1} (cz+d)^{1/2} \theta(z),$$

where $\left(\frac{c}{d}\right)$ is the extended Legendre symbol (see [16]) and ε_d is 1 or i depending on whether d is 1 or 3 mod 4. (As usual $\sqrt{z}$ denotes the branch which is positive on $\mathbb{R}^+$). Define

$$j(\gamma, z) := \left(\frac{c}{d}\right) \varepsilon_d^{-1} (cz+d)^{1/2}.$$

Definition 2.3. Let $4|N$ and k be a $1/2$ integer. A modular form $f(z)$ of weight k for $\Gamma_0(N)$ is a holomorphic function on $\mathbb{H}$ satisfying

(i) $f(\gamma z) = (j(\gamma, z))^{2k} f(z) \quad$ for $\gamma \in \Gamma_0(N)$

(ii) $f(z)$ is holomorphic at each cusp.

(See [16, Chapter 4] for the precise meaning of (ii).)

We are interested in finding 'good' bounds for Fourier coefficients of cusp forms of half integral weight. Firstly, we derive what is called as the 'trivial' bound.

Proposition 2.4. Let $f \in S_k(\Gamma)$ and with Fourier coefficients a_n, then

$$a_n = O(n^{k/2}).$$

Proof. Since $\mathrm{Im}(\gamma z) = \mathrm{Im}(z)/|(cz+d)|^2$, $F(z) = |f(z)|y^{k/2}$ is Γ invariant. Also, since f vanishes at cusps, $F(z)$ is bounded on $\mathbb{H}$, i.e. $|F(z)| < M$ (this condition on F is equivalent to f being a cusp form, see [21]). Now, the coefficient a_n may be expressed as

$$a_n = e^{2\pi n y} \int_0^1 e(-nx) f(x+iy) dx.$$

Hence $|a_n| \leq M e^{2\pi n y} y^{-k/2}$. Put $y = 1/n$ to get the required result. □

It is an important result in the theory of modular forms that, for any congruence subgroup Γ, $M_k(\Gamma)$-modular forms of weight k, form a finite dimensional vector space over $\mathbb{C}$. Furthermore,

$S_k(\Gamma)$-the space of cusp forms of weight k, is a finite dimensional Hilbert space. For $f, g \in S_k(\Gamma)$, the inner product, called the Peterson inner product, is defined as

$$\langle f, g \rangle = \int_{\Gamma \backslash \mathbb{H}} f(z)\overline{g(z)} y^k \frac{dxdy}{y^2}.$$

For any congruence subgroup Γ there is a canonical way to generate cusp forms of weight k. For given cusp p of Γ, the idea is to average out the factor of automorphy over coset representatives of Γ_p (the stabilizer of p in Γ). This is called the Poincare series at p. We shall be only considering the case $p = \infty$. More precisely, the m-th Poincare series of weight k at ∞ is defined by

$$P_m(z, k) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} (j(\gamma, z))^{-2k} e(m\gamma z).$$

If we assume $k > 2$, then it can be shown that the above series converges absolutely. The importance of Poincare series lies in the fact that $P_m(z, k) \in S_k(\Gamma)$ (which we assume) and that

Proposition 2.5. The $P_m(z, k), m \geq 1$ span the space of cusp forms $S_k(\Gamma)$.

Proof. Let $f \in S_k(\Gamma)$ then

$$\begin{aligned} \langle P_m, f \rangle &= \int_{\Gamma \backslash \mathbb{H}} P_m(z)\overline{f(z)} y^k \frac{dxdy}{y^2} \\ &= \int_{\Gamma \backslash \mathbb{H}} \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} j(\gamma, z))^{-2k} e(m\gamma z)\overline{f(z)} y^k \frac{dxdy}{y^2} \\ &= \int_0^\infty \int_0^1 e(mz)\overline{f(z)} y^k \frac{dxdy}{y^2} \\ &= \frac{\overline{a}_m}{(4\pi m)^{k-1}} \Gamma(k-1), \end{aligned} \tag{2.14}$$

where $f(z) = \sum_{n=1}^\infty a_n e(nz)$ and Γ is the gamma function. It follows that if $\langle P_m, F \rangle = 0$ for all m, then $a_m = 0$ for all m, i.e. $f \equiv 0$. Hence P_m's span $S_k(\Gamma)$. □

From now on, we shall exclusively focus on Poincare series, as bounds on fourier coefficients of P_m's will imply similar bounds for general cusp forms. A computation will show that if we write

$$P_m(z, k) = \sum_{n \geq 1} \hat{P}_m(n) e(nz)$$

then

$$\hat{P}_m(n) = 2\left(\frac{n}{m}\right)^{(k-1)/2} \left\{ \delta_{m,n} + 2\pi i^{-k} \sum_{\substack{c \equiv 0 (N) \\ c > 0}} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right) \frac{K(m, n, c)}{c} \right\} \tag{2.15}$$

where

$$J_{k-1}(z) = \sum_{l \geq 0} \frac{(-1)^l}{l!\Gamma(l+k)} \left(\frac{z}{2}\right)^{k-1+2l} \tag{2.16}$$

is the J-Bessel function and

$$K(m,n,c) = \sum_{\substack{d \bmod c \\ (d,c)=1}} \left(\frac{c}{d}\right)^{2k} \varepsilon_d^{-2k} e\left(\frac{m\overline{d}+nd}{c}\right) \tag{2.17}$$

is a Kloosterman sum closely related to the one we have met in section 1.2.

Let $f_1, \ldots, f_R$ be an orthonormal basis for $S_k(N)$ (with respect to the Petersson inner product). If we write $f_j = \sum_{n\geq 1} \hat{f}_j(n)e(nz)$, then (2.14) gives

$$\begin{aligned} P_m(z,k) &= \sum_{j=1}^{R} \langle P_m, f_j\rangle f_j \\ &= \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{j=1}^{R} \overline{\hat{f}}_j(m) f_j(z). \end{aligned}$$

Hence

$$\hat{P}_m(n) = \sum_{j=1}^{R} \langle P_m, f_j\rangle \hat{f}_j(n) = \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{j=1}^{R} \overline{\hat{f}}_j(m)\hat{f}_j(n).$$

Substituting for $\hat{P}_m(n)$ from (2.15) and setting $m = n$ we get

$$\hat{P}_n(n) = \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{j=1}^{R} |\hat{f}_j(n)|^2 = 1 + 2\pi i^{-k} \sum_{c\equiv 0(N)} J_{k-1}\left(\frac{4\pi n}{c}\right) \frac{K(n,n,c)}{c}. \tag{2.18}$$

The above equation is called the *Petersson Formula* , and it will form the backbone for our estimates of $\hat{f}_j(n)$'s. In order to do so, we must evaluate $K(m,n,c)$ in a form in which we can see cancelation in the sum in (2.18).

2.3 Salié sums

In the case of our interest (i.e. for Linnik's problem) $k = 3/2+l$, where l is even, so $\varepsilon_d^{-2k} = \varepsilon_d^{-3} = \varepsilon_d$ (recall ε_d takes 1 or i as its value). Substituting this in (2.17) gives

$$K(m,n,c) = \sum_{\substack{d \bmod c \\ (d,c)=1}} \varepsilon_d \left(\frac{c}{d}\right)\ e\left(\frac{m\overline{d}+nd}{c}\right).$$

An interesting thing about this sum is that it can be evaluated in a simpler form. After using Chinese remainder theorem and quadratic reciprocity we are led to the following

Lemma 2.6. If $c = qr$ with $(q,r) = 1$ and $4|r$ then

$$K_k(m,n,c) = K_{k-q+1}(m\overline{q}, n\overline{q}, r)S(m\overline{r}, n\overline{r}, q),$$

where for q odd $S(m,n,q)$ is the Salié sum defined as

$$S(m,n,q) = \sum_{x \bmod q} \left(\frac{x}{q}\right) e\left(\frac{mx+n\overline{x}}{q}\right).$$

Proof. Expand the RHS. Set $d = xr\overline{r} + yq\overline{q}$ with x, y ranging over the residue classes modulo q and r respectively. So $d = x \,(\mathrm{mod}\, q), d = y \,(\mathrm{mod}\, r)$. Also $\overline{d} = \overline{xr}r + \overline{yq}q \,(\mathrm{mod}\, c)$ (as can be directly verified by multiplying d and $\overline{d}$ and using CRT). By the quadratic reciprocity we have

$$\left(\frac{c}{d}\right) = (-1)^{\frac{y-1}{2}\frac{q-1}{2}} \left(\frac{r}{y}\right)\left(\frac{x}{q}\right).$$

We also have $\varepsilon_d = \varepsilon_y$ and $(-1)^{(y-1)/2(q-1)/2} = \varepsilon_y^{q-1}$. Combining all this,we get the LHS. □

We now prove some properties satisfied by $S(m,n,q)$.

Lemma 2.7. Let $(m,q) = 1 = (n,q)$.

$$S(m,n,q) = \left(\frac{m}{q}\right) S(1,mn,q) \tag{i}$$

$$S(1,n^2,q) = \varepsilon_q \sqrt{q} \sum_{x^2 \equiv 1 (q)} e\left(\frac{2xn}{q}\right) \tag{ii}$$

Proof. (i) Since $(m,q) = 1$, do a change of variable by setting $y = mx$, (hence $x = \overline{m}y, \overline{x} = m\overline{y}$). We get

$$S(m,n,q) = \sum_{y(q)} \left(\frac{\overline{m}y}{q}\right) e\left(\frac{y+mn\overline{y}}{q}\right) = \left(\frac{m}{q}\right) S(1,mn,q)$$

as $\left(\frac{m}{q}\right) = \left(\frac{\overline{m}}{q}\right)$.

(ii) The proof is via Gauss sum. Let

$$G(a,b;q) = \sum_{x\,(q)} e\Big(\frac{ax^2+bx}{q}\Big).$$

It is a classical evaluation that

$$G(a,0;q) = \varepsilon_q \sqrt{q} \Big(\frac{a}{q}\Big).$$

Set $A = \sum_{x^2 \equiv n^2 (q)} e(2x/q)$. Since $(n,q) = 1$, it follows that we need to show

$$S(n^2,1;q) = \varepsilon_q \sqrt{q}\, A.$$

Using the trivial fact that $\sum_{a(q)} e(ay/q) = q$ or 0 depending on $y \equiv 0(q)$ or not, we get that

$$\begin{aligned} A &= \frac{1}{q}\sum_{x\,(q)} e\Big(\frac{2x}{q}\Big)\sum_{a\,(q)} e\Big(\frac{a(x^2-n^2)}{q}\Big) \\ &= \frac{1}{q}\sum_{a\,(q)} G(a,2;q)e\Big(\frac{-an^2}{q}\Big) \end{aligned}$$

Claim : $G(a,b;q) = 0$ if $(a,q) \nmid b$.
Proof : Let $d = (a,q)$. Write $a = a'd$, $q = q'd$ and $x = x_1 + q'x_2$. Then the sum over $x(\mathrm{mod}\, q)$ splits as a double sum over $x_1(\mathrm{mod}\, q')$ and $x_2(\mathrm{mod}\, d)$. Clearly, we have $x^2 \equiv x_1^2(\mathrm{mod}\, q')$. With this observation, we write

$$e\Big(\frac{ax^2+bx}{q}\Big) = e\Big(\frac{a'x^2}{q'}\Big)e\Big(\frac{b(x_1+q'x_2)}{q}\Big) = e\Big(\frac{a'x_1^2}{q'}\Big)e\Big(\frac{bx_1}{q}\Big)e\Big(\frac{bx_2}{d}\Big).$$

Now the sum $\sum_{x_2(q)} e\big(\frac{bx_2}{d}\big)$ is zero, unless $d \mid b$. $\diamond$

In the present case, since q is odd and $b = 2$, the above sum then is

$$A = \frac{1}{q}\sum_{(a,q)=1} G(a,2;q)e\Big(\frac{-an^2}{q}\Big).$$

But if $(a,q) = 1$ then $ax^2 + 2x = a((x+\overline{a})^2 - \overline{a}^2) = a(x+\overline{a})^2 - \overline{a}$. Therefore,

$$\begin{aligned} G(a,2;q) &= e\Big(\frac{-\overline{a}}{q}\Big)G(a,0;q) \\ &= e\Big(\frac{-\overline{a}}{q}\Big)\Big(\frac{a}{q}\Big)\varepsilon_q\sqrt{q}, \end{aligned}$$

so

$$A = \frac{\varepsilon_q}{\sqrt{q}}\sum_{a\,(q)}\Big(\frac{a}{q}\Big)e\Big(\frac{-an^2-\overline{a}}{q}\Big).$$

From the definition of A, note that $A = \overline{A}$ (as $\overline{A}$ is obtained by changing x to $-x$ and a to $-a$), which when combined with the fact that $\overline{\varepsilon_q} = 1/\varepsilon_q$ gives that

$$\varepsilon_q\sqrt{q}A = \sum_{a\,(q)}\Big(\frac{a}{q}\Big)e\Big(\frac{an^2+\overline{a}}{q}\Big) = S(n^2,1,q)$$

$\square$

A word on notation. We shall use $\overline{x}$ to denote inverse of x with respect to some modulus which shall be implicit, or sometimes clear from the context. If $\overline{x}$ appears in an expression with a denominator y, then generally $\overline{x}$ is the inverse with respect to y.

Corollary 2.8. Let q be odd and $(a,q)=1$, we have

$$S(n,n,q)=\varepsilon_q\sqrt{q}\Big(\frac{n}{q}\Big)\sum_{\substack{ab=q\\(a,b)=1}} e\Big(2n\Big(\frac{\overline{a}}{b}-\frac{\overline{b}}{a}\Big)\Big). \tag{2.19}$$

Proof. Since $S(n,n,q)=\big(\frac{n}{q}\big)S(n^2,1,q)$, it is enough, by lemma 2.7(ii) to show that

$$\sum_{x^2=1(q)} e\Big(\frac{2xn}{q}\Big)=\sum_{\substack{ab=q\\(a,b)=1}} e\Big(2n\Big(\frac{\overline{a}}{b}-\frac{\overline{b}}{a}\Big)\Big).$$

Consider $y=a\overline{a}-b\overline{b}$. $y \bmod b = 1$ and $y \bmod a = -1$. Hence $y^2=1 \bmod (ab)$(by CRT). Conversely it can be shown that every solution to $y^2=1(q)$ can be written in the form $a\overline{a}-b\overline{b}$ for some $ab=q,(a,b)=1$. □

Corollary 2.9. If q is odd and $(n,q)=1$, the Salié sums $S(n,n,q)$ satisfy the bound

$$|S(n,n,q)|\le\tau(q)q^{1/2} \tag{2.20}$$

where τ is the divisor function.

Proof. Clear from (2.19). □

For any k, the Kloosterman sums (2.17) also satisfies the bound analogous to (2.20). This is a consequence of the deep work of A. Weil on Riemann hypothesis for curves over finite fields.

Weil Bound on Kloosterman sums. The Kloosterman sums $K(m,n,c)$ satisfy the bound

$$K(m,n,c)\le(m,n,c)^{1/2}\tau(c)c^{1/2}. \tag{2.21}$$

In particular if $(n,c)=1$ we have

$$K(n,n,c)\ll c^{1/2+\epsilon}$$

So we have

$$\sum_{c>0}c^{-\sigma}|K(n,n,c)|\ll n^{\epsilon}$$

for any $\epsilon>0$, if $\sigma>3/2$. Also the J-Bessel function satisfies the following bound for $x>0$ [25]

$$J_{k-1}(x)\ll \min\big\{x^{k-1},\frac{1}{\sqrt{x}}\big\}\le x^{\nu},\ \text{ if }\ -1/2\le\nu\le k-1.$$

Choosing $\nu=1/2+\delta$ with δ arbitrarily small, we get by putting these bounds in (2.18) that

$$\hat{P}_n(n)\ll n^{1/2+\delta}\sum_{c>0}\frac{|K(n,n,c)|}{c^{3/2+\delta}}\ll n^{1/2+\delta+\epsilon}.$$

Since ϵ and δ are arbitrary, we conclude that $\hat{P}_n(n)\ll n^{1/2+\epsilon}$ for any $\epsilon>0$. (We are abusing the notation by using the same ϵ everywhere). It follows readily from (2.18) that $\hat{f}_j(n)\ll n^{k/2-1/4+\epsilon}$

for any $\epsilon > 0$, for $j = 1, \ldots, R$ (recall $f_1, \ldots, f_R$ is an orthonormal basis for $S_k(N)$). Since the Fourier coefficient of any $f \in S_k(N)$ is a linear combination of Fourier coefficient of f_j's we have

Proposition 2.10. Let k be half an odd integer and let $f \in S_k(N), (4 \mid N)$, then the n-th Fourier coefficient of f satisfies

$$a_n = O(n^{k/2-1/4+\epsilon}). \tag{2.22}$$

We now make some important remarks. The bound (2.22) just falls short of (2.5), the bound required to solve the Linnik's problem. However, we note that (2.22) is essentially the best possible bound for a general n. To see this, consider the theta series $\theta(z, \psi) = \sum_{m \in \mathbb{Z}} \psi(m) e(m^2 z)$, where ψ is a character (mod 4) with $\psi(-1) = -1$. It is a cusp form of weight $3/2$ for $\Gamma_0(8)$. Clearly, $|a_{m^2}| = m$ and $k/2 - 1/4 = 1/2$. Hence a bound of the type (2.5) is simply false in general. So how do we proceed to settle Linnik's problem now? The best way out is to assume that n's are square free. This is not a major restriction as we have already solved the Linnik's problem for squares (this is essentially the content of section 2.1). The case for a general n can be handled similarly (as we did for squares) using the Shimura correspondence. Suppose a_n's are Fourier coefficients of a modular form of half integral weight k. Then from Shimura correspondence (2.10) we get (upon inversion) that

$$a(tn^2) = a(t) \sum_{d \mid n} \chi(d) \mu(d) d^{k-3/2} A\left(\frac{n}{d}\right),$$

where $A(n)$'s are Fourier coefficients of a modular form of even integral weight $2k-1$ (and χ some character). By the Deligne bound, alluded to previously, we have

$$A(n) \ll n^{k-1+\epsilon}.$$

Hence

$$\begin{aligned} |a(tn^2)| &\ll |a(t)| \sum_{d \mid n} d^{k-3/2} \left|\frac{n}{d}\right|^{k-1+\epsilon} \\ &= |a(t)| n^{k-1+\epsilon} \sum_{d \mid n} d^{-1/2} \ll |a(t)| n^{k-1+\epsilon}. \end{aligned}$$

So if a_t's satisfied a bound of the form

$$a(t) \ll t^{k/2-1/4-\delta+\epsilon},$$

for all square free t (recall t's in the Shimura map are square free) for some $0 < \delta \leq 1/4$, we would have

$$|a(tn^2)| \ll (tn^2)^{k/2-1/4-\delta+\epsilon}.$$

With this insight, we shall restrict ourselves to the case of a_m's where m is square free.
Notice that in proving proposition 2.10 we did not use any cancellation that might occur in the sum in (2.18). As we shall see, improvement in (2.22) for square free n is obtained by exploiting this cancellation.

2.4 Iwaniec's bound

Theorem 2.11. Let $k \geq 5/2$, $4 \mid N$ and $f \in S_k(N)$. Then for n square free the n-th Fourier coefficient of f satisfies the bound

$$a(n) \ll n^{\frac{k}{2}-\frac{1}{4}-\frac{1}{222}} \tau(n) log(n), \tag{2.23}$$

where the implied constant depends on f

We shall break the proof into several small steps. We first outline the strategy. We know that if $f \in S_k(N)$, then $f \in S_k(qN)$ for any q. Also the Petersson norm of f with respect to $\Gamma_0(qN)$ (denoted by $\|f\|_q^2$) is related to $\|f\|^2$ by the relation

$$\|f\|_q^2 = [\Gamma_0(N) : \Gamma_0(qN)]\|f\| \tag{2.24}$$

where the index is determined by

$$[\Gamma_0(N) : \Gamma_0(qN)] = \frac{qN \prod_{p|qN}(1+p^{-1})}{N \prod_{p|N}(1+p^{-1})}. \tag{2.25}$$

If a_n is the n-th Fourier coefficient of f, then the n-th Fourier coefficient of normalized form $f_q = f/\|f\|_q$ (normalized with respect to $\Gamma_0(qN)$) is given by

$$\hat{f}(n) = a(n)/\|f\|_q \tag{2.26}$$

Take an orthonormal basis for $S_k(qN)$ which contains f_q. Using the positivity of Petersson formula (2.18) and substituting for $\hat{f}(n)$ from (2.26) we get

$$\frac{\Gamma(k-1)\,|a(n)|^2}{(4\pi n)^{k-1}\|f\|^2[\Gamma_0(N) : \Gamma_0(qN)]} \leq 1 + 2\pi i^{-k} \sum_{c \equiv 0(qN)} c^{-1} K(n,n,c) J_{k-1}\left(\frac{4\pi n}{c}\right) \tag{2.27}$$

The main idea in the proof of Iwaniec is to sum the above inequalities by varying q through primes in some interval of the form $(P, 2P)$, exploiting cancellations that occur in the right side due to change in the arguments of the Kloosterman sums.

We now prove some lemmas which shall be used in the course of the proof. It is time to introduce yet another notation from analytic number theory. For two real valued functions f and g, by the notation $f(x) \sim g(x)$ we mean $\lim_{x\to\infty} f(x)/g(x) = 1$

Lemma 2.12. Let τ be the divisor function.

$$\sum_{n \leq x} \tau(n) \log n \sim x(\log x)^2$$

$$\sum_{n \geq x} \frac{\tau(n)\log n}{n^2} \sim \frac{3(\log x)^2}{x}$$

Proof. Recall that the sum of divisor function $T(x) = \sum_{n \leq x} \tau(n)$ satisfies the bound $T(x) = x \log x + O(x)$. Also, recall Abel's partial summation formula

$$\sum_{y < n \leq x} a_n f(n) = A(x)f(x) - A(y)f(y) - \int_y^x A(t)f'(t)dt$$

where A is the sum function of a_n's and f a continuously differentiable function on $[0, 1]$. Applying this formula with $f(t) = \log t$ we get

$$\begin{aligned}\sum_{n \leq x} \tau(n) \log n &= T(x) \log(x) - \int_1^x \frac{T(u)}{u} du \\ &= (x \log x + O(x)) \log x + \int_1^x (\log u + O(1)) du \\ &= x(\log x)^2 + O(\log x)\end{aligned}$$

For the second statement, we use the above result and put $f(t) = t^{-2}$ in Abel's formula. We get

$$\sum_{n \geq x} \frac{\tau(n) \log n}{n^2} = \frac{(\log x)^2}{x} + 2 \int_x^\infty \frac{(\log u)^2}{u^2} du + O\left(\frac{\log x}{x^2} + \int_x^\infty \frac{\log u}{u^2} du\right).$$

After evaluating the integrals we get

$$\sum_{n \geq x} \frac{\tau(n) \log n}{n^2} = \frac{3(\log x)^2}{x} + O\left(\frac{\log x}{x}.\right)$$

□

Lemma 2.13. Let $F : \mathbb{Z} \to \mathbb{C}$ be a periodic function with period m. Let $\hat{F}$ be its fourier transform. Then

$$\sum_{u \leq X} F(u) \ll \frac{X|\hat{F}(0)|}{m} + \|\hat{F}\|_\infty \log m$$

Proof.

$$\begin{aligned}\sum_{u \leq X} F(u) &= \sum_{u \leq X} \frac{1}{m} \sum_{r \pmod m} \hat{F}(r) e\left(\frac{ru}{m}\right) \\ &= \frac{X}{m}\hat{F}(0) + \sum_{\substack{r \pmod m \\ r \neq 0}} \frac{1}{m}\hat{F}(r) \sum_{u \leq X} e\left(\frac{ru}{m}\right) \\ &\ll \frac{X}{m}|\hat{F}(0)| + \frac{1}{m} \sum_{\substack{r \pmod m \\ r \neq 0}} \frac{|\hat{F}(r)|}{|1 - e(r/m)|} \\ &\ll \frac{X}{m}|\hat{F}(0)| + \|\hat{F}\|_\infty \sum_{1 \leq r \leq m/2} \frac{1}{r} \\ &\ll \frac{X|\hat{F}(0)|}{m} + \|\hat{F}\|_\infty \log m\end{aligned}$$

□

Corollary 2.14.

$$\sum_{u\leq X} e\left(\frac{\nu\overline{u}}{m}\right) \ll \frac{X\tau(m)(m,\nu)}{m} + m^{1/2}(m,\nu)^{1/2}\tau(m)\log m. \tag{2.28}$$

Proof. Take $F(u) = e\left(\frac{\nu\overline{u}}{m}\right)$. It is a periodic function with period $m/(m,\nu)$. Its Fourier transform is

$$\hat{F}(r) = \sum_{u(\mathrm{mod}\, m)} e\left(\frac{\nu\overline{u} - ru}{m}\right),$$

the classical Kloosterman sum. The corollary follows by using the Weil bound (2.21). □

The above technique is called completing the sum and is often employed in analytic number theory. Note that since the bound is linear in X, the estimate is also valid for intervals.

Lemma 2.15. (**Polya-Vinogradov inequality.**) Let χ be a primitive character modulo D, with D square free and let $r < D$. Then, for any a, M, N with $M < N$

$$\sum_{\substack{l\equiv a\,(\mathrm{mod}\, r)\\ M\leq n\leq N}} \chi(l) \ll D^{1/2}\log D \tag{2.29}$$

Proof. Define the Gauss sum associated to χ by

$$g(\chi) = \sum_{m=1}^{D} \chi(m) e\left(\frac{m}{D}\right).$$

If $(n, D) = 1$, then

$$\begin{aligned}\chi(n)g(\overline{\chi}) &= \sum_{m=1}^{D} \overline{\chi}(m)\chi(n)e\left(\frac{m}{D}\right)\\ &= \sum_{h=1}^{D} \chi(h)e\left(\frac{nh}{D}\right),\end{aligned} \tag{2.30}$$

where we have put $m \equiv nh \pmod{D}$. It is easy to show that if $(n, D) > 1$, then the right side of (2.30) vanishes (by a simple change of variable, as in lemma 2.17.(ii) , and using the primitivity of χ). Hence (2.30) holds true for any n. From this, we get

$$|\chi(n)|^2|g(\chi)|^2 = \sum_{h_1=1}^{D}\sum_{h_2=1}^{D} \overline{\chi}(h_1)\chi(h_2)e\left(\frac{n(h_1-h_2)}{D}\right).$$

Summing over n over a complete set of residues (mod D), the left side gives $\phi(D)|g(\chi)|^2$. While on

the right side, the sum over the exponentials is zero unless $h_1 \equiv h_2$. So we get

$$\phi(D)|g(\chi)|^2 = D \sum_{h\,(D)} |\chi(h)|^2 = D\phi(D).$$

We deduce that

$$|g(\chi)| = D^{1/2}. \tag{2.31}$$

Let $(r, D) = d$. Write $D = dq$. Since $r < D$, we have $d < D$, so $q > 1$. Also, since D is square free, we have $(d, q) = 1$. Hence $\chi = \chi_1\chi_2$, where χ_1 and χ_2 are primitive characters modulo d and q respectively. Note that χ_2 is non-trivial as $q > 1$. Also, $d|r$, so the congruence $l \equiv a \pmod r$ implies $l \equiv a \pmod d$. Therefore, the factor $\chi_1(a)$ comes out of the sum in (2.29) and what is left is a sum over $\chi_2(l)$. Further, $(q, r) = 1$, so we could have as well started by assuming $(r, D) = 1$, which we do now.

Put $l = a + kr$. The condition $M \le l \le N$ gives $\left\lceil \frac{M-a}{r} \right\rceil \le k \le \left\lfloor \frac{N-a}{r} \right\rfloor$. We denote these new bounds by M' and N' respectively. Then, expressing $\chi(n)$ using (2.30), we get

$$\begin{aligned}\sum_{\substack{l \equiv a \pmod r \\ M \le n \le N}} \chi(l) &= \sum_{M' \le k \le N'} \chi(a + kr) \\ &= \frac{1}{g(\overline{\chi})} \sum_{h=1}^{D} \chi(h) e\left(\frac{ah}{D}\right) \sum_{M' \le k \le N'} e\left(\frac{hkr}{D}\right) \end{aligned} \tag{2.32}$$

Now the inner sum is a GP which satisfies the trivial bound

$$\Big| \sum_{M' \le k \le N'} e\left(\frac{hkr}{D}\right) \Big| \le \frac{2}{\left|1 - e\left(\frac{hr}{D}\right)\right|}.$$

Note that, as r is relatively prime to D, hr will be a multiple of D only if $h = D$, in which case there is no contribution to the sum (2. 32) as $\chi(h) = 0$. Hence taking absolute values in (2.32) and using (2.31), we get

$$D^{1/2} \Big| \sum_{\substack{l \equiv a \pmod r \\ M \le n \le N}} \chi(l) \Big| \le \sum_{h=1}^{D-1} \frac{2}{\left|1 - e\left(\frac{hr}{D}\right)\right|} = \sum_{h=1}^{D-1} \frac{2}{\left|1 - e\left(\frac{h}{D}\right)\right|} = \sum_{h=1}^{D-1} \frac{1}{|\sin(\pi h/D)|}. \tag{2.33}$$

The first equality above is due to the fact that hr and h runs through the same set of residues (mod D), as $(r, D) = 1$. We now estimate the last sum. For any convex function $f(x)$ we have that

$$f(x) \le \frac{1}{\delta} \int_{x-\delta/2}^{x+\delta/2} f(t)dt.$$

Taking $f(x) = (\sin \pi x)^{-1}, \delta = 1/D$ we see that

$$\sum_{h=1}^{D-1} \frac{1}{\sin(\pi h/D)} \le D \sum_{h=1}^{D-1} \int_{\frac{h}{D} - \frac{1}{2D}}^{\frac{h}{D} + \frac{1}{2D}} \frac{dt}{\sin \pi t} = D \int_{\frac{1}{2D}}^{1 - \frac{1}{2D}} \frac{dt}{\sin \pi t} = 2D \int_{\frac{1}{2D}}^{\frac{1}{2}} \frac{dt}{\sin \pi t}.$$

Now $\sin \pi t > 2t$ for $0 < t < 1/2$, so that

$$2D \int_{\frac{1}{2D}}^{\frac{1}{2}} \frac{dt}{\sin \pi t} < 2D \int_{\frac{1}{2D}}^{\frac{1}{2}} \frac{dt}{2t} = D \log D.$$

Substitute this back in (2.33) to complete the proof. □

We now get on with the proof of Iwaniec's bound. There will be four independent parameters C, D, P and R whose values will be chosen at the end. We shall be taking q to a prime, in which case the index in (2.25) is just p or $p+1$. We average the inequality (2.27) over prime $p \nmid n$ in the interval $P < p < 2P$, each such p being weighted by $\log p$. We shall choose P later subject to $N(\log n)^2 < P \leq n^{1/6}$. The resulting left side is estimated with help of the following

Proposition 2.16.

$$\sum_{\substack{P<p<2P \\ p \nmid n}} \frac{\log p}{p+1} \sim \log 2$$

Proof. We may omit the restriction $p \nmid n$ without affecting the result. Define a function θ on $\mathbb{N}$ by, $\theta(m) = \log m$ if m is a prime and 0 otherwise. We are interested in the asymptotics of

$$\sum_{P<k<2P} \frac{\theta(k)}{k+1}$$

By the prime number theorem we have that

$$\sum_{P<k<2P} \theta(k) \sim P.$$

A simple partial summation using the above gives the required result. □

Therefore, the resulting left side is asymptotically $\mu n^{1-k}|a(n)|^2$, where μ is a positive constant depending only on f. On the right hand side of our basic inequality we obtain a sum of Kloosterman sums to moduli $c \equiv 0 (\mathrm{mod}\, N)$ weighted by

$$\omega(c) = \sum_{\substack{P<p<2P \\ p \nmid n, p|c}} \log p.$$

We clearly have $\omega(c) \leq \log c$ if $c > 0$. The contribution from the constant term 1 on the right side is $\omega(0) \sim P$ (by the prime number theorem). Therefore we have

$$n^{1-k}|a(n)|^2 \ll P + |S| \tag{2.34}$$

where S is the weighted sum of Kloosterman sums,

$$S = \sum_{c \equiv 0 (\mathrm{mod}\, N)} \omega(c) c^{-1} K(n, n, c) J_{k-1}\left(\frac{4\pi n}{c}\right)$$

The Bessel function satisfies the bound

$$J_{k-1}\left(\frac{4\pi n}{c}\right) \ll \min\left\{\left(\frac{c}{n}\right)^{1/2}, \left(\frac{n}{c}\right)^{3/2}\right\}.$$

This along with Weil bound (2.21) implies that terms with $c \leq C$ contribute to S at most

$$S_1 \ll n^{-1/2} \sum_{c \leq C} (c,n)^{1/2} \tau(c) \log c \ll C n^{-1/2} (\tau(n) \log n)^2, \tag{2.35}$$

where the last inequality is a consequence of

Proposition 2.17. If $C = n^{\alpha}$, then

$$S_1^* := \sum_{c \leq C} (c,n)^{1/2} \tau(c) \log c \ll C(\tau(n) \log n)^2. \tag{2.36}$$

Proof. The required sum is nothing but

$$\sum_{c \leq C} \sum_{\substack{d|n \\ (c,n)=d}} d^{1/2} \tau(c) \log c = \sum_{d|n} d^{1/2} \sum_{\substack{c \leq C \\ (c,n)=d}} \tau(c) \log c.$$

But

$$\sum_{d|n} d^{1/2} \sum_{\substack{c \leq C \\ (c,n)=d}} \tau(c) \log c \leq \sum_{d|n} d^{1/2} \sum_{\substack{c \leq C \\ d|c}} \tau(c) \log c$$

Writing $c = kd$, and using $\tau(kd) \leq \tau(k)\tau(d)$, the last sum is

$$\sum_{d|n} d^{1/2} \tau(d) \log d \sum_{k \leq C/d} \tau(k) \log k$$

Estimating the inner sum using lemma 2.12 we get that

$$S_1^* \leq C(\log C)^2 \sum_{d|n} \frac{\tau(d)(\log d)^3}{d^{1/2}} \ll C(\tau(n) \log n)^2.$$

For the last bound we have used the fact that C is a power of n and that $\tau(d)(\log d)^3 \ll d^{1/4}$(say), and hence the sum over d is $\ll \tau(n) \ll (\tau(n))^2$. □

Similarly, using lemma 2.12, terms with $c \geq D$ contribute to S at most

$$S_3 \ll n^{3/2} \sum_{c \geq D} (c,n)^{3/2} \tau(c) c^{-2} \log c \ll D^{-1} n^{3/2} (\tau(n) \log n)^2. \tag{2.37}$$

We shall later choose $C = n^{\alpha} < n < D = n^{\beta}$ so that the above bound will be valid. In fact we shall

choose α and β quite close to 1. We are then left with estimating the central term

$$S_2 = \sum_{\substack{C<c<D \\ c\equiv 0 (\mathrm{mod}\, N)}} \omega(c) c^{-1} K(n,n,c) J_{k-1}\left(\frac{4\pi n}{c}\right).$$

We set $c = qr$ where q is the largest factor of c coprime with nN, so r has all its prime factors in nN and is divisible by N. Therefore $\omega(c) = \omega(q)$ and S_2 splits into

$$S_2 = \sum_{\substack{r|(nN)^\infty \\ r\equiv 0 (\mathrm{mod}\, N)}} r^{-1} T_r \tag{2.38}$$

where

$$T_r = \sum_{\substack{C<qr<D \\ (q,nN)=1}} \omega(q) q^{-1} K(n,n,qr) J_{k-1}\left(\frac{4\pi n}{qr}\right).$$

We first estimate T_r as follows:

$$T_r \ll (n,r)^{1/2} r^{1/2} \tau(r) \sum_q \tau(q) q^{-1/2} \min\left\{\left(\frac{qr}{n}\right)^{1/2}, \left(\frac{n}{qr}\right)^{3/2}\right\} \log q,$$

where we have used the fact that $(q,r) = 1$ and hence $(n,qr) = (n,r), \tau(qr) = \tau(q)\tau(r)$.

Proposition 2.18.

$$T_r^* := \sum_q \tau(q) q^{-1/2} \min\left\{\left(\frac{qr}{n}\right)^{1/2}, \left(\frac{n}{qr}\right)^{3/2}\right\} \log q \ll r^{-1/2} n^{1/2} (\log n)^2. \tag{2.39}$$

Proof.

$$T_r^* = r^{1/2} n^{-1/2} \sum_{q<n/r} \tau(q) \log q + n^{3/2} r^{-3/2} \sum_{q\geq n/r} q^{-2} \tau(q) \log q.$$

Applying lemma 2.12 we get the required result (recall that $qr < D$ hence $r < D$; therefore we can replace $\log r$ terms with $\log n$ terms). □

We conclude that

$$T_r \ll \tau(r)(n,r)^{1/2} n^{1/2} (\log n)^2.$$

We shall use this bound only for r sufficiently large, say $r > R$. Hence the contribution to S_2 of terms with $r > R$ is bounded by

$$\sum_{r>R} r^{-1} T_r \ll R^{-1/2} n^{1/2} (\log n)^2 \sum_r \tau(r)(n,r)^{1/2} r^{1/2}.$$

Recall that all primes factors of r are in nN. Hence the complete sum over r above is given by a finite product over primes in nN which is estimated by $O(\tau(n)^2)$. So, we get

$$\sum_{r>R} r^{-1} T_r \ll R^{-1/2} n^{1/2} (\tau(n) \log n)^2. \tag{2.40}$$

It remains estimate T_r for $r \leq R$. By lemma 2.6 the Kloosterman sum in T_r factors into

$$K(n,n,c) = K(n\overline{q}, n\overline{q}, r)S(n\overline{r}, n\overline{r}, q)$$

where $q\overline{q} = 1 (\mathrm{mod}\ r), r\overline{r} = 1 (\mathrm{mod}\, q)$ and $S(n\overline{r}, n\overline{r}, q)$ is the Salié sum. Since the Kloosterman sum $K(n\overline{q}, n\overline{q}, r)$ depends only on $q (\mathrm{mod}\, r)$, we can split the sum over q in T_r into residue classes mod r to obtain

$$T_r \leq \sum_{s (\mathrm{mod}\, r)}^{*} |K(n\overline{s}, n\overline{s}, r)T_{rs}| \tag{2.41}$$

where

$$T_{rs} = \sum_{\substack{C<qr<D \\ q \equiv s (\mathrm{mod}\, r), (q, nN)=1}} \omega(q) q^{-1} S(n\overline{r}, n\overline{r}, q) J_{k-1}\left(\frac{4\pi n}{qr}\right).$$

The Salié sum was evaluated in Corollary 2.8; we have

$$S(n\overline{r}, n\overline{r}, q) = \varepsilon_q q^{1/2} \left(\frac{nr}{q}\right) \sum_{\substack{ab=q \\ (a,b)=1}} e\left(2n\overline{r}\left(\frac{\overline{a}}{b} - \frac{\overline{b}}{a}\right)\right).$$

If $(a,b) = 1$, then we clearly have $a\overline{a} + b\overline{b} = 1 (\mathrm{mod}\, ab)$ (by CRT). From this we get the following 'reciprocity formula'

$$\frac{\overline{a}}{b} + \frac{\overline{b}}{a} \equiv \frac{1}{ab} \pmod 1$$

From now on we shall assume $(a,b) = 1$ without explicit mention. Using the above reciprocity formula we write

$$\sum_{ab=q} e\left(2n\overline{r}\left(\frac{\overline{a}}{b} - \frac{\overline{b}}{a}\right)\right) = 2\,\mathrm{Re} \sum_{\substack{ab=q \\ a<b}} e\left(2n\frac{\overline{b}}{ar} + 2n\frac{\overline{b}\overline{r}}{a} - \frac{2n}{abr}\right).$$

Recall $r \equiv s (\mathrm{mod}\, r)$ and $N \mid r$ and $4 \mid N$, so $\left(\frac{r}{q}\right) = \left(\frac{r}{s}\right)$. Using this, and inserting the above expression into T_{rs} we get

$$T_{rs} = \varepsilon_s \left(\frac{r}{s}\right) \left(\frac{2r}{n}\right)^{1/2} \mathrm{Re} \sum_{\substack{C<abr<D \\ a<b,\ ab \equiv s (\mathrm{mod}\, r)}} \omega(ab) \left(\frac{n}{ab}\right) e\left(2n\frac{\overline{b}}{ar} + 2n\frac{\overline{b}\overline{r}}{a}\right) j\left(\frac{2n}{abr}\right)$$

where

$$j(x) = x^{1/2} e(-x) J_{k-1}(2\pi x).$$

$j(x)$ satisfies the estimates $j(x) \ll x^{1/2}$, $j(x) \ll 1$ and $j'(x) \ll 1$. Using these estimates we remove $j(2n/abr)$ using partial summation in b to obtain

$$T_{rs} \ll C^{-1}(nr)^{1/2} \sum_{\substack{a<A \\ (a,nr)=1}} \Big| \sum_{\substack{a<b<B \\ ab \equiv s (\mathrm{mod}\, r)}} \omega(ab) \left(\frac{n}{b}\right) e\left(2n(1 + r\overline{r})\frac{\overline{b}}{ar}\right)\Big|$$

where $A = (D/r)^{1/2}$ and some B which depends on a such that $C < arB < D$. Since $\omega(ab) = \omega(a) + \omega(b)$, the innermost sum splits into $V_a + \omega(a)V'(a)$, where

$$V_a = \sum_{\substack{a<b<B \\ ab\equiv s(\mathrm{mod}\, r)}} \omega(b) \left(\frac{n}{b}\right) e\left(2n(1+r\overline{r})\frac{\overline{b}}{ar}\right),$$

and $V'(a)$ is given by the same sum without the weight $\omega(b)$. Thus

$$T_{rs} \ll C^{-1}(nr)^{1/2} \sum_{\substack{a<A \\ (a,nr)=1}} (|V_a| + \omega(a)|V'_a|).$$

We now turn to estimating V_a, the most non-trivial part of the proof. First, observe that if we put $b = pl$, with $P < p < 2P$, (if there is no such prime then $\omega(b) = 0$ by definition) the sum over b splits into sum over p and l as follows

$$V_a = \sum_{P<p<2P} (\log p) \left(\frac{n}{p}\right) \sum_{\substack{a/p<l<B/p \\ apl\equiv s(\mathrm{mod}\, r)}} \left(\frac{n}{l}\right) e\left(2n(1+r\overline{r})\frac{\overline{pl}}{ar}\right). \tag{2.42}$$

Suppose $a \leq P$. We now make the following observation: The character $\left(\frac{n}{l}\right)$ is non-trivial on any arithmetic progression to modulus ar because n is square free and larger than ar. Let us look at the sum over l more closely. l is varying over some interval subject to the condition $l \equiv \overline{ap}s(\mathrm{mod}\, r)$. If we further impose the condition that $l \equiv \lambda(\mathrm{mod}\, a)$, by Chinese remainder theorem (as $(a, r) = 1$) we can split the sum over l into residue classes $(\mathrm{mod}\, ar)$ as follows

$$\sum_{\lambda(\mathrm{mod}\, a)} \sum_{\substack{a/p<l<B/p \\ l\equiv\overline{ap}s(\mathrm{mod}\, r) \\ l\equiv\lambda(\mathrm{mod}\, a)}} \left(\frac{n}{l}\right) e\left(2n(1+r\overline{r})\frac{\overline{pl}}{ar}\right) = \sum_{\lambda(\mathrm{mod}\, a)} e\left(2n(1+r\overline{r})\frac{\overline{pl}}{ar}\right) \sum_{\substack{a/p<l<B/p \\ l\equiv\lambda'(\mathrm{mod}\, ar)}} \left(\frac{n}{l}\right).$$

Put $L = B/P$, then $l < B/p < L$. We now apply the Polya-Vinogradov estimate for character sums (lemma 2.15)

$$\sum_{\substack{l<L \\ l\equiv\lambda(\mathrm{mod}\, ar)}} \left(\frac{n}{l}\right) \ll n^{1/2} \log n$$

on the inner sum and the trivial estimate on the outer sum to get that in (2.42) $\sum_l \ll an^{1/2} \log n$. Substituting this back in (2.42) and performing the sum over P we get.

$$V_a \ll aPn^{1/2} \log n. \tag{2.43}$$

If $P < a < A$ we interchange the sums in V_a to get

$$V_a = \sum_{l<L} \left(\frac{n}{l}\right) \sum_{\substack{P_1<p<P_2 \\ apl\equiv s(\mathrm{mod}\, r)}} \log p \left(\frac{n}{p}\right) \left(2n(1+r\overline{r})\frac{\overline{pl}}{ar}\right),$$

where $P_1 = \max(P, a/l)$ and $P_2 = \min(2P, B/l)$. We now use Cauchy-Schwarz inequality in the outer sum on l (in the form $|\sum_{i=1}^n x_i|^2 \le n\sum_{i=1}^n |x_i|^2$) to get

$$V_a^2 \le L\sum_{l<L}\Big|\sum_{\substack{P_1<p<P_2\\ apl\equiv s(\mathrm{mod}\, r)}} \log p\left(\frac{n}{p}\right)\left(2n(1+r\overline{r})\frac{\overline{pl}}{ar}\right)\Big|^2.$$

Squaring out and changing the order of summation back gives

$$V_a^2 \ll L(\log P)^2 \sum_{\substack{P<p_1\le p_2<2P\\ p_1\equiv p_2(\mathrm{mod}\, r)}} \Big|\sum_{\substack{a/p_2<l<B/p_1\\ ap_1l\equiv s(\mathrm{mod}\, r)}} e\left(2n(1+r\overline{r})(p_1-p_2)\frac{\overline{p_1p_2l}}{ar}\right)\Big|. \tag{2.44}$$

We have used the trivial bound $\log p < \log P$ to take it out of the summation. As for the term in the exponential, originally we have $(\overline{p_2} - \overline{p_1})$. But $(\overline{p_2} - \overline{p_1}) \equiv (p_1 - p_2)\overline{p_1p_2} \,(\mathrm{mod}\, a)$ (recall $p|b$ and hence $p \nmid a$, thus we can talk of $\overline{p_1}$ and $\overline{p_2}$). Since $p_1 \equiv p_2 \,(\mathrm{mod}\, r)$ we also have $(\overline{p_2} - \overline{p_1}) \equiv (p_1 - p_2)\overline{p_1p_2} \,(\mathrm{mod}\, r)$, and hence $(\overline{p_2} - \overline{p_1}) \equiv (p_1 - p_2)\overline{p_1p_2} \,(\mathrm{mod}\, ar)$ which justifies the replacement. The contribution of the terms with $p_1 = p_2$ to the sum is $O(L^2(\log P)^2\, P/\log P) = O(L^2 P \log P)$

Suppose $p_1 \ne p_2$. We now take a closer look at the inner sum in (2.44). Firstly, the congruence condition in the sum that $l \equiv \overline{ap}s(\mathrm{mod}\, r)$ imposes no restriction on $l(\mathrm{mod}\, a)$ as a and r are relatively prime. Also, $r \mid (p_1 - p_2)$ and $r\overline{r} = 1(\mathrm{mod}\, a)$ (recall it is 1 $(\mathrm{mod}\, q)$ and $q = ab$), so the argument in the exponential is essentially $4n\frac{p_1-p_2}{r}\frac{\overline{p_1p_2}\overline{l}}{a}$. Therefore the sum over l is precisely an incomplete sum of the form (2.28). Since $(r, a) = 1 = (n, a)$, we have $(\frac{4n(p_1-p_2)}{r}, a) = (p_1 - p_2, a)$. Applying corollary 2.14 we deduce that the sum over l in (2.44) satisfies

$$\sum_l \ll (p_1 - p_2, a)a^{-1}L + (p_1 - p_2, a)^{1/2}a^{1/2}\tau(a)\log a.$$

Summing over $p_1 \ne p_2$, we get

$$V_a^2 \ll \left(L^2P + a^{-1}L^2P^2 + a^{1/2}P^2\right)(\tau(a)\log n)^2.$$

Since $a > P$ the middle term can be ignored. After substituting $L = B/P$ and taking square we obtain

$$V_a \ll \left(BP^{-1/2} + a^{1/4}B^{1/2}P^{1/2}\right)\tau(a)\log n. \tag{2.45}$$

The same estimates hold for V_a' by similar arguments. We now use these estimates to bound T_{rs}. By (2.43), terms with $a \le P$ contribute to T_{rs} at most

$$C^{-1}(nr)^{1/2}Pn^{1/2}\log n\sum_{a\le P} a \ll C^{-1}(nr)^{1/2}P^3n^{1/2}(\log n)^2. \tag{2.46}$$

And, by using (2.45) terms with $P < a < A$ contribute to T_{rs} at most

$$C^{-1}(nr)^{1/2}\sum_{P<a<A}\left(BP^{-1/2} + a^{1/4}B^{1/2}P^{1/2}\right)\tau(a)\log n.$$

Using the fact that $B < D/ra$ and that $A = (D/r)^{1/2}$, the above sum can be estimated using Abel's formula as we did in lemma 2.12. Combining the result with (2.46) we get

$$T_{rs} \ll C^{-1}(nr)^{1/2}\left(n^{1/2}P^3 + DP^{-1/2} + D^{7/8}P^{1/2}\right)(\log n)^2.$$

We may drop the term $n^{1/2}P^3$ in comparison to other terms, a step which can be justified after we have chosen P and D. Inserting the rest into (2.41) and summing over $s(\mathrm{mod}\, r)$ using the trivial estimate $|K(n\overline{s}, n\overline{s}, r| \leq r$ we deduce that

$$T_r \ll r^{5/2}C^{-1}\left(DP^{-1/2} + D^{7/8}P^{1/2}\right)n^{1/2}(\log n)^2.$$

Next summing over $r \leq R$, we get

$$\sum_{r\leq R} r^{-1}T_r \ll R^2C^{-1}\left(DP^{-1/2} + D^{7/8}P^{1/2}\right)n^{1/2}(\log n)^2\sum_r r^{-1/2}.$$

The sum over r can be estimated to be $O(\tau(n))^2$. We deduce

$$\sum_{r\leq R} r^{-1}T_r \ll R^2C^{-1}\left(DP^{-1/2} + D^{7/8}P^{1/2}\right)n^{1/2}(\tau(n)\log n)^2. \tag{2.47}$$

Combining the bounds (2.47),(2.40),(2.37) and (2.35) we get that

$$S \ll \left[Cn^{-\frac{1}{2}} + D^{-1}n^{\frac{3}{2}} + R^{-\frac{1}{2}}n^{\frac{1}{2}} + R^2C^{-1}\left(DP^{-\frac{1}{2}} + D^{\frac{7}{8}}P^{\frac{1}{2}}\right)n^{\frac{1}{2}}\right](\tau(n)\log n)^2.$$

We choose $C = n^{\frac{110}{111}}$, $D = n^{\frac{112}{111}}$, $P = n^{\frac{14}{111}}$, $R = n^{\frac{2}{111}}$ to obtain the much sought out bound

$$S \ll n^{\frac{1}{2}-\frac{1}{111}}(\tau(n)\log n)^2. \tag{2.48}$$

Finally, substituting (2.48) into (2.34), we deduce theorem 2.

Remark 2.19. As mentioned before, Iwaniec in [13] proves the bound

$$a(n) \ll n^{\frac{k}{2}-\frac{1}{4}-\frac{1}{28}+\epsilon}$$

for all $\epsilon > 0$, for square free n. It is believed that the analogue of Ramanujan's conjecture holds for the half-integral modular forms as well, i.e.

$$a(n) \ll n^{\frac{k}{2}-\frac{1}{2}+\epsilon},$$

for square free n. But this still remains a conjecture.

Chapter 3

Ergodic methods in number theory

Ergodic methods have been successfully employed to solve a variety of problems in number theory, especially problems involving diophantine approximations and equidistribution. Just as the study of diophantine equations had opened up new vistas of mathematics over the course of centuries, the study of diophantine inequalities, a subject still in its infancy, holds a similar promise.

As a case in point, we give the example of the Oppenheim conjecture. Let Q be an indefinite quadratic form on $\mathbb{R}^n$ which is not a multiple of a rational form, for example $x^2 + y^2 - \sqrt{2}z^2$. Oppenheim conjectured in 1929 that if $n \geq 3$, the set of vales taken by Q on $\mathbb{Z}^n$ is dense in $\mathbb{R}$. In other words

$$Q(\mathrm{x}) < \epsilon \text{ is solvable, with x } \in \mathbb{Z}^n \text{ for any } \epsilon > 0.$$

The conjecture was solved for large values of n using techniques from analytic number theory. However, the case $n = 3$ remained open for a long time (it was known that if the conjecture is true for a particular n then it is true for any larger n, hence the case $n = 3$ was of prime interest). The conjecture was finally solved in 1987 by G. Margulis using deep techniques from the theory of unipotent flows. More specifically, he considered the action of $H = SO(Q)$ on $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$ and showed that relatively compact orbits of H are necessarily compact. The Oppenheim conjecture then followed as a consequence of a remarkable observation made by M. S. Raghunathan.

In this chapter we shall try to explore some aspects ergodic methods which have number theoretic relevance, especially the dynamics of lattices under the action of the diagonal torus.

3.1 Measure rigidity and equidistribution

We saw in chapter 1 that the sequence $(n\alpha)$ is equidistributed modulo 1 when α is irrational. We had also remarked that Weyl had proved a similar result for $(p(n))$ where p is a polynomial with at least one irrational coefficient. The purpose of this section is to prove the equidistribution of $(n^2\alpha)$ mod 1, using ergodic theory. Such a proof was first given by Furstenberg, but we shall follow [8] and [17] in our exposition.

3.1.1 Measures on Compact Metric spaces

In this section we will recall some results from measure theory and functional analysis and use these to establish certain results needed in the proof of equidistribution on $(n^2\alpha)$. Through out this section and the next, X will denote a compact metric space and $\mathcal{B}$ the Borel σ-algebra on X. Let μ be a measure on $(X, \mathcal{B})$. For a measurable function $f : X \to \mathbb{R}$, we shall use $\int f d\mu$ and $\mu(f)$ interchangeably to denote the integral of f over whole of X.

A map $T : X \to X$ is said to be *measurable* if $T^{-1}A \in \mathcal{B}$ for all $A \in \mathcal{B}$, and is *measure preserving* if it is measurable and if $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{B}$. If T is measure preserving we say μ is T-invariant (to put the emphasis on μ when T is fixed). Let L^1_μ denote the space of (equivalence classes) of measurable functions $f : X \to \mathbb{R}$ with $\int |f| d\mu < \infty$.

Proposition 3.1. A measure μ on X is T invariant if and only if

$$\int f d\mu = \int f \circ T d\mu \tag{3.1}$$

for all $f \in L^1_\mu$.

Proof. If (3.1) holds, then for any $A \in \mathcal{B}$, we have

$$\mu(A) = \int \chi_A d\mu = \int \chi_A \circ T d\mu = \int \chi_{T^{-1}A} d\mu = \mu(T^{-1}A).$$

Conversely suppose μ is T-invariant, then (3.1) holds for functions of the form χ_A with $A \in \mathcal{B}$, and hence for any simple function (a finite linear combination of characteristic functions). Let f be a non-negative real valued function in L^1_μ. Choose a sequence of functions simple functions (f_n) increasing to f. Then $(f_n \circ T)$ increases to $f \circ T$. By the dominated convergence theorem we have

$$\int f \circ T d\mu = \lim_{n\to\infty} \int f_n \circ T d\mu = \lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

□

From now on, we shall restrict our attention to Borel probability measures on X. As in section 1.2, let $\mathcal{P}(X)$ be the space of Borel probability measures on X and let $\mathcal{P}^T(X)$ denote the space of T-invariant measures in $\mathcal{P}(X)$. As usual, $C(X)$ denotes the space of continuous functions on X. Recall the following important result:

Riesz representation theorem. There is a one-to-one correspondence between Borel probability measures on X and positive linear functionals $\Lambda : C(X) \to \mathbb{R}$, with $\Lambda(1) = 1$. More specifically, given such a Λ, there exists a unique $\mu \in \mathcal{P}(X)$ such that

$$\Lambda(f) = \int f d\mu.$$

(A functional $\Lambda : C(X) \to \mathbb{R}$ is positive if $f \geq 0 \Rightarrow \Lambda(f) \geq 0$.)

An immediate corollary is that if $\mu_1, \mu_2 \in \mathcal{P}(X)$, then

$$\mu_1 = \mu_2 \iff \int f d\mu_1 = \int f d\mu_2, \text{ for all } f \in C(X). \tag{3.2}$$

Recall that the weak* topology on $\mathcal{P}(X)$ is the smallest topology on $\mathcal{P}(X)$ making each of the maps $\mu \mapsto \int f d\mu$ continuous for every $f \in C(X)$

Corollary 3.2. $\mathcal{P}(X)$ endowed with the weak* topology is compact.

Proof. The Riesz representation theorem enables us to identify $\mathcal{P}(X)$ as a subspace of $C^*(X)$ endowed the weak* topology. Under this identification $\mathcal{P}(X)$ is mapped into the unit sphere in $C^*(X)$. (as $\|\Lambda\| = |\Lambda(1)| = |\int d\mu| = 1$). The Banach-Alaoglu theorem says that the closed unit ball in $C^*(X)$ is weak* compact. Being a closed subset of this, $\mathcal{P}(X)$ is compact. □

Given any continuous map $T : X \to X$, it induces a map $T_* : \mathcal{P}(X) \to \mathcal{P}(X)$ defined by

$$T_*(\mu)(A) = \mu(T^{-1}A)$$

for any $A \in \mathcal{B}$.

Proposition 3.3. Let $f \geq 0$ be a measurable map and $\mu \in \mathcal{P}(X)$. Then

$$\int f dT_*\mu = \int f \circ T d\mu. \tag{3.3}$$

Proof. Clearly, (3.3) holds true for functions of the form χ_A for any $A \in \mathcal{B}$, and hence for simple functions. We now proceed as in proposition 3.1 by taking a sequence of simple functions (f_n) increasing to f. We then have

$$T_*(\mu)(f) = \lim_{n\to\infty} T_*(\mu)(f_n) = \lim_{n\to\infty} \mu(f_n \circ T) = \mu(f \circ T).$$

□

In particular, (3.3) holds for any continuous f. As a consequence we get that $T_* : \mathcal{P}(X) \to \mathcal{P}(X)$ is a continuous map. To see this, suppose $\mu_n \to \mu$. Then, for $f \in C(X)$, $T_*(\mu_n)(f) = \mu_n(f \circ T) \to \mu(f \circ T) = T_*(\mu)(f)$, so $T_*(\mu_n) \to T_*(\mu)$. Using these observations, we strengthen proposition 3.1.

Lemma 3.4. Let μ be measure in $\mathcal{P}(X)$. Then $\mu \in \mathcal{P}^T(X)$ if and only if $\int f d\mu = \int f \circ T d\mu$ for all $f \in C(X)$.

Proof. ($\Rightarrow$) Indeed, by the previous proposition, $T_*(\mu)(f) = \mu(f \circ T) = \mu(f)$, for all $f \in C(X)$. We conclude from (3.2) that $T_*(\mu) = \mu$, i.e. $\mu \in \mathcal{P}^T(X)$. Of course, the other implication follows directly from proposition 3.1. □

The next theorem demonstrates that for continuous maps on X, we can always find invariant measures.

Theorem 3.5. Let $T : X \to X$ be a continuous map of a compact metric space and let (ν_n) be any sequence in $\mathcal{P}(X)$. Then any weak* limit point of the sequence (μ_n) defined by $\mu_n = \frac{1}{n}\sum_{k=0}^{n-1} T_*^k(\nu_n)$ is a member of $\mathcal{P}^T(X)$.

Proof. First note that the sequence of measures (μ_n) will have a limit point in $\mathcal{P}(X)$ as $\mathcal{P}(X)$ is weak* compact by corollary 3.2. Let μ be a limit point and let $\mu_{n_j} \to \mu$. For a continuous function f, let $\|f\|_\infty$ denote the supremum of f on X, which is finite as X is compact. By the definition of $T_*\mu_n$ as in (3.2), we get

$$\begin{aligned}\Big|\int f\circ T d\mu_{n_j} - \int f d\mu_{n_j}\Big| &= \frac{1}{n_j}\Big|\int \sum_{k=0}^{n_j-1}(f\circ T^{k+1} - f\circ T^k)d\nu_{n_j}\Big| \\ &= \frac{1}{n_j}\Big|\int (f\circ T^{n_j+1} - f)d\nu_{n_j}\Big| \\ &\le \frac{2}{n_j}\|f\|_\infty \longrightarrow 0\end{aligned}$$

as $j \to \infty$. It follows that $\int f\circ T d\mu = \int f d\mu$ for all $f \in C(X)$, so $\mu \in \mathcal{P}^T(X)$ by lemma 3.4. □

3.1.2 Ergodicity and Unique ergodicity

Definition 3.6. A measure preserving transformation $T : X \to X$ on a probability space $(X, \mathcal{B}, \mu)$ is said to be *ergodic* if for any $B \in \mathcal{B}$

$$T^{-1}(B) = B \implies \mu(B) = 0 \text{ or } \mu(B) = 1.$$

In words, a transformation is ergodic if there is no non-trivial way to divide the space into smaller T-invariant ones.

Proposition 3.7. Let $T : X \to X$ be a measure preserving transformation on $(X, \mathcal{B}, \mu)$. T is ergodic if and only if, for any measurable function $f : X \to \mathbb{R}$

$$f\circ T = f \text{ a.e} \Longrightarrow f \text{ is a constant a.e.}$$

Proof. ($\Rightarrow$) Let $B \in \mathcal{B}$ be T-invariant, i.e. $\chi_B \circ T = \chi_{T^{-1}(B)} = \chi_B$. We take $f = \chi_B$ and conclude that χ_B is constant almost everywhere. It follows that $\mu(B) = 0$ or 1.

($\Leftarrow$) Conversely suppose T is ergodic. Let $f : X \to \mathbb{R}$ be such that $f \circ T = f$ almost everywhere. We can redefine f on a measure zero set such that the new function, which we still call f, is invariant under T. Define

$$B_t = \{t \in \mathbb{R} : f(x) > t\}.$$

Since f is T-invariant, we have that $T^{-1}(B_t) = B_t$. Thus, for any t, $\mu(B_t) = 0$ or 1. Now, if f were not constant almost everywhere, there would be some t_0 such that $0 < \mu(B_{t_0}) < 1$, which contradicts our previous statement. □

Example 3.8. The circle rotation $R_\alpha : \mathbb{T} \to \mathbb{T}$ given by $z \mapsto ze(\alpha)$ is ergodic with respect to the Lebesgue measure l if and only if α is irrational.

Proof. It is easy to construct non-trivial subsets invariant under R_α when α is rational. So, suppose α is irrational. Then it follows from Dirichlet's approximation theorem that $\mathbb{Z}\alpha$, considered in $\mathbb{T}$, is dense in $\mathbb{T}$. Now, let $B \subseteq \mathbb{T}$ be invariant under R_α. By Lusin's theorem, for any $\epsilon > 0$ we can choose a function $f \in C(\mathbb{T})$ such that $\|f - \chi_B\|_1 < \epsilon$. By invariance of B we have

$$\|f \circ R_\alpha^n - f\| < 2\epsilon$$

for all n. Since f is continuous it follows that

$$\|f \circ R_t - f\| < 2\epsilon \tag{3.4}$$

for all $t \in \mathbb{R}$. Thus, since l is rotation invariant, we have

$$\begin{aligned}\|f - \int f(t)dt\|_1 &= \int |\int (f(x) - f(x+t))dt|dx \\ &\leq \int\int \|f(x) - f(x+t)|dxdt \leq 2\epsilon\end{aligned}$$

by Fubini's theorem and (3.4). We deduce that

$$\|\chi_B - \mu(B)\|_1 \leq \|\chi_B - f\|_1 + \|f - \int f(t)dt\|_1 + \|\int f(t)dt - \mu(B)\|_1 \leq 4\epsilon.$$

Since this holds true for any $\epsilon > 0$, we conclude that χ_B is equal to $\mu(B)$ all most everywhere. But this can only be true if $\mu(B) = 0$ or $\mu(B) = 1$. Therefore R_α is ergodic.

We now give a second proof using Fourier series. Let $\sum_{n\in\mathbb{Z}} a_n e(nx)$ be the Fourier expansion of χ_B. Since χ_B is invariant under R_α, this expansion is invariant under the change of variable $x \to x + \alpha$. From the uniqueness of Fourier series we get $a_n = a_n e(n\alpha)$. Since α is irrational this can only be true if $a_n = 0$ for $n \neq 0$, i.e. χ_B is a constant almost everywhere. We conclude as in the first proof that R_α is ergodic. □

We now state a fundamental result in ergodic theory.

Birkhoff's ergodic theorem. Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T : X \to X$ be an ergodic measure preserving transformation. Then for any $f \in L_1(X, \mu)$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int f d\mu \quad \text{for almost every } x \in X.$$

We shall not prove this result as it is not entirely in line with the spirit of our results; a proof could be found in any standard book in ergodic theory or dynamical systems. Instead we shall deduce some notable corollaries from it. Before that, we remark that $\int f d\mu$ is the space average of f, and $(\lim_{n\to\infty} \sum_{k=0}^{n-1} f(T^k x))/n$ can be thought of the time average of f along the T-trajectory of x. The Birkhoff ergodic theorem then says that if T is ergodic, then the time average of f along the trajectory of almost any point in X is a constant equal to the space average of f. In fact, this interpretation is what originally prompted Boltzmann to introduce, albeit in an implicit way, the

notion of ergodicity through his 'Ergodic hypothesis' on thermodynamical systems.

Corollary 3.9. Let $(X, \mathcal{B}, \mu)$ be a Borel probability space. If $T : X \to X$ is an ergodic measure preserving transformation, then for almost every $x \in X$ the sequence of points $(T^n x)$ is equidistributed with respect to μ.

Proof. Let U be an open set. Note that

$$\frac{\#\{0 \leq k < n \,:\, T^k x \in U\}}{n} = \frac{1}{n} \sum_{k=0}^{n-1} \chi_U(T^k x).$$

We apply the Birkhoff ergodic theorem to $f = \chi_U$, which is in $L_1(X, \mu)$ to get

$$\frac{1}{n} \sum_{k=0}^{n-1} \chi_U(T^k x) = \int_X \chi_U d\mu = \mu(U).$$

Therefore (1.1) holds true for any open set U and the corollary follows. □

Definition 3.10. Let $T : X \to X$ be a continuous μ-invariant transformation. We say a point $x \in X$ is *generic* (with respect to T and μ) if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int f d\mu,$$

for any $f \in C(X)$.

Remark 3.11. Birkhoff's ergodic theorem says that generic points with respect to an ergodic transformation (T, μ) have full measure. Also, notice that by (3.2), if $x \in X$ is generic with respect to some $\mu \in \mathcal{P}^T(X)$, then it cannot be generic with respect to any other measure in $\mathcal{P}^T(X)$.

Definition 3.12. Let X be a compact metric space. A transformation $T : X \to X$ is said to be *uniquely ergodic* if there is exactly one T-invariant probability measure on X.

Remark 3.13. We now make an observation. Let X and T be as in the above definition and let μ be the unique T-invariant measure. Since X is compact, for any $f \in C(X)$, the limit $\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(T^k x)$ exists for any $x \in X$. Therefore we can define a measure ν_x on X as follows

$$\nu_x(f) = \int f d\nu_x = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x).$$

By construction $\nu_x(f \circ T) = \nu_x(f)$, i.e. ν_x is T-invariant. So $\nu_x = \mu$ for all x. In other words if T is uniquely ergodic, then statement of the Birkhoff ergodic theorem is true for all $x \in X$. It follows, as in corollary (3.3), that $(T^k x)$ is equidistributed for all $x \in X$.

Theorem 3.14. For a continuous map $T : X \to X$ on a compact metric space the following are equivalent.

(1) T is uniquely ergodic.
(2) For every $f \in C(X)$,

$$S_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \to C_f, \tag{3.5}$$

where C_f is a constant independent of x.
(3) The convergence (3.5) holds for every f in a dense subset of $C(X)$.

Proof. (1) $\Rightarrow$ (2). Let μ be the unique invariant measure for T. We apply theorem 3.5 to the constant sequence (δ_x). Since $\mathcal{P}(X)$ is compact and since μ is the only possible limit point, we deduce that

$$\frac{1}{n}\sum_{k=0}^{n-1} \delta_{T^k x} \longrightarrow \mu$$

in the weak* topology, so for any $f \in C(X)$ we have

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \longrightarrow \int f d\mu.$$

(2) $\Rightarrow$ (1). Let $\mu \in \mathcal{P}^T(X)$. (3.5), by the dominated convergence theorem, implies that

$$\int f d\mu = \int \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) d\mu = C_f$$

for all $f \in C(X)$. It follows that C_f is the integral of f with respect to any measure in $\mathcal{P}^T(X)$. Hence by (3.2), $\mathcal{P}^T(X)$ can contain only a single measure.

(3) $\Rightarrow$ (1). Suppose $\mu, \nu \in \mathcal{P}^T(x)$. Then, as in the the previous step we get

$$\int f d\mu = C_f = \int f d\nu$$

for any f in a dense subset of $C(X)$. In other words the functionals $\mu \mapsto \int f d\mu$ and $\nu \mapsto \int f d\nu$ agrees on a dense subset of $C(X)$. From continuity, it follows that they agree on the whole of $C(X)$, hence $\mu = \nu$.

(2) $\Rightarrow$ (3) is trivial. □

Example 3.15. Let $\alpha \in \mathbb{R}$ be irrational. Then the circle rotation $R_\alpha : \mathbb{T} \to \mathbb{T}$ is uniquely ergodic. The unique invariant measure is the Lebesgue measure l.

Proof. We shall prove this using property (3) of theorem (3.14). We already know that l is invariant R_α. Let $f(t) = e(ht)$ for some $h \in \mathbb{Z}$. We have

$$\frac{1}{n}\sum_{k=0}^{n-1} f(R_\alpha^k(t)) = \frac{1}{n}\sum_{k=0}^{n-1} e(h(t+k\alpha)) = \begin{cases} 1 & \text{if } h = 0 \\ \frac{1}{n} e(ht)\dfrac{e(nh\alpha)-1}{e(h\alpha)-1} & \text{if } h \neq 0 \end{cases}$$

The above equation clearly shows that

$$\frac{1}{n}\sum_{k=0}^{n-1} f(R_\alpha^k(t)) \to \int f dl = \begin{cases} 1 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0. \end{cases}$$

By linearity, the above convergence holds true for any trigonometric polynomial, which are dense in $C(X)$ is be the Stone-Weierstrass theorem. □

As a corollary, we obtain yet another proof of the equidistribution of $(n\alpha)$ mod 1, for α irrational.

3.1.3 Equidistribution of $(n^2\alpha)$ mod 1

We now turn towards proving the equidistribution of $(n^2\alpha)$. Let $\alpha \in \mathbb{R}$ be irrational. We first define a map T on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ using α which we shall show to be uniquely ergodic with λ as the unique invariant measure . Define $T : \mathbb{T}^2 \to \mathbb{T}^2$ by

$$T(x, y) = (x + \alpha, y + 2x + \alpha).$$

A simple induction argument will give us that

$$T^n(x, y) = (x + n\alpha, y + 2n\alpha + n^2\alpha).$$

Now, if $(\mathbb{T}^2, T)$ were uniquely ergodic it would follow from remark 3.13 that all T orbits would equidistribute in $\mathbb{T}^2$ with respect to λ. In particular the orbit of (0, 0) would be equidistributed. Since $T^n(0, 0) = (n\alpha, n^2\alpha)$, this would imply that $(n^2\alpha)$ is equidistributed modulo 1.

As a first step, we will prove that T is ergodic with respect to the Lebesgue measure λ on $\mathbb{T}^2$. Clearly, T preserves the Lebesgue measure as it is a unimodular transformation $\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$ followed by a translation.

Proposition 3.16. Lebesgue measure, λ, on $\mathbb{T}^2$ is ergodic under T

Proof. Let $f \in L^2(\lambda)$ be T-invariant. We expand f to a Fourier series

$$f(x, y) = \sum_{m,n} \hat{f}_{m,n} e(mx + ny).$$

By T-invariance, we have

$$\hat{f}_{m,n} = \hat{f}_{m+2n,n} e((m + n)\alpha). \tag{3.6}$$

In particular $|\hat{f}_{m,n}| = |\hat{f}_{m+2n,n}|$ $(= |\hat{f}_{m+2kn,n}|$ for any $k)$. By the Riemann-Lebesgue lemma, $\hat{f}_{m,n} \to \infty$ as $(m, n) \to \infty$. Hence $\hat{f}_{m,n} = 0$ if $n \neq 0$ (by taking $k \to \infty$). On the other hand if $n = 0$, (3.6) gives $\hat{f}_{m,0} = e(m\alpha)\hat{f}_{m,0}$, from which we conclude $\hat{f}_{m,0} = 0$ as α is irrational. It follows that f is constant almost everywhere and that T is ergodic. □

We now turn towards the proof of unique ergodicity.

Theorem 3.17. Let $g : \mathbb{T} \to \mathbb{T}$ be a continuous function, and $T_g : \mathbb{T}^2 \to \mathbb{T}^2$ be the map

$$T_g(x, y) = (x + \alpha, y + g(x))$$

with α irrational. If the Lebesgue measure λ is T_g-ergodic then it is the only T_g-invariant probability measure, i.e. $(\mathbb{T}^2, T_g)$ is uniquely ergodic.

Proof. Let l denote the Lebesgue measure on $\mathbb{T}$. Assume that T_g is ergodic with respect to $\lambda = l \times l$. Let

$$E = \{(x, y) \,|\, (x, y) \text{ is generic w.r.t } \lambda\}.$$

As T_g is ergodic with respect to λ, we have that $\lambda(E) = 1$ (see remark 3.11). We claim that E is invariant under the map $(x, y) \mapsto f(x, y + a)$. To see this, notice that $(x, y) \in E$ means

$$\frac{1}{k} \sum_{k=0}^{n-1} f(T_g^k(x, y)) \longrightarrow \int f d\lambda$$

for all $f \in C(\mathbb{T}^2)$. Define f_a to be the map $(x, y) \mapsto (x, y + a)$. It follows that

$$\begin{aligned} \frac{1}{n} \sum_{k=0}^{n-1} f(T_g^k(x, y + a)) = \frac{1}{n} \sum_{n=0}^{n-1} f_a(T_g^k(x, y)) \\ \longrightarrow \int f_a d\lambda = \int f d\lambda, \end{aligned}$$

since l is invariant under rotations. So $(x, y+a) \in E$ also. This means that $E = E_1 \times \mathbb{T}$ for some set $E_1 \subseteq \mathbb{T}$ with $\mu(E_1) = 1$. Now, suppose ν is a T_g-invariant ergodic measure on $\mathbb{T}^2$. Write $\pi : \mathbb{T}^2 \to \mathbb{T}$ for the projection $(x, y) \to x$. Then $\pi_*\nu$ is R_α invariant, so by unique ergodicity of R_α, $\pi_*\nu = l$. In particular, $\nu(E) = \nu(E_1 \times \mathbb{T}) = l(E_1) = 1$. By ergodicity of ν, ν-almost every point in $\mathbb{T}^2$ is generic with respect to ν. Thus there must be a point $(x, y) \in E$ generic with respect to ν. But we have already seen (remark 3.11) that a point in X cannot be generic with respect to more than one invariant probability measure. We conclude that $\nu = \lambda$. □

Corollary 3.18. If α is irrational then the sequence $(n^2\alpha)$ is equidistributed modulo one.

Proof. Take $g(x) = 2x+\alpha$, we have already proved that $T_g = T$ is ergodic for the Lebesgue measure λ. From the last theorem we conclude that T_g uniquely ergodic, therefore every T orbit equidistributes with respect to λ. The orbit of $(0, 0)$ is $(n\alpha, n^2\alpha)$. Projecting to the second coordinate gives the required result. □

Remark 3.19. The word measure rigidity has no precise meaning. It used to refer to the dearth of invariant measures under specific situations as in theorem 3.17. Sometimes it provides a quick route to equidistribution.

3.2 Dynamics of Lattices and the Littlewood conjecture

For number theoretic applications, one of the most important and interesting space to study is the space of lattices. By a *lattice* in $\mathbb{R}^n$ we mean a discrete subgroup of $\mathbb{R}^n$ whose $\mathbb{R}$-span is the whole of $\mathbb{R}^n$. If L is a lattice in $\mathbb{R}^n$, fixing a basis for L gives us an element A of $GL_n(\mathbb{R})$, and changing the basis takes A to AM where $M \in GL_n(\mathbb{Z})$. Therefore, the space of lattices in $\mathbb{R}^n$ is naturally identified with the quotient $GL_n(\mathbb{R})/GL_n(\mathbb{Z})$. A lattice L is said be unimodular if $vol(\mathbb{R}^n/L) = 1$, which is the same as saying $|A| = 1$ where A is any matrix in $GL_N(\mathbb{R})$ representing L. So the space of oriented unimodular lattices is naturally identified with $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$. We shall denote this space by X_n. Two lattices L_1 and L_2 are said to be homothetic (or similar) if one is obtained by scaling the other, i.e. $L_2 = \lambda L_1$ where λ is a non-zero real number. And the space of lattices up to homothety, denoted by Y_n, is identified with $PGL_n(\mathbb{R})/PGL_n(\mathbb{Z})$. The space Y_n is essentially the same as X_n (as can be seen by scaling a lattice to have covolume one), except that in Y_n we do not keep track of the orientation. Almost anything that can be said in the context of X_n can also be said about Y_n and vice versa. So, if necessary, we shall switch back and forth between these spaces.

From now on, we shall refer $SL_n(R)$ by G and $SL_n(\mathbb{Z})$ by Γ. The topology on X_n can be described as follows: a sequence of lattices L_i in X_n converges to L if and only if each L_i has a basis $(b_1^{(i)}, \ldots, b_n^{(i)})$ converging, as $i \to \infty$, to $(b_1, \ldots, b_n)$ - a basis for L. For $n \geq 2$, the space X_n is not compact, as can be seen by considering the sequence of lattices

$$L_i = \text{span}_{\mathbb{Z}}(i^{-1}, i^{-1}, \ldots, i^{n-1}).$$

What makes the space X_n amenable to ergodic methods is the fact that there is a natural *probability measure* on X_n which is invariant under the action of G. (See [1]). We denote it by μ. Further the precise way in which X_n fails to be compact is described by the Mahler's compactness criterion. Before we get to this, let us introduce some notations. For $x \in \mathbb{R}^n$ let $|x|$ denote the usual Euclidean norm of x. Let $L \in \mathbb{R}^n$ be a unimodular lattice. We define $|L|$ to be

$$|L| = \min\{|x| \,:\, x \in L, x \neq 0\}.$$

Theorem 3.20. (Mahler) Let $r > 0$. Define $\Omega_r = \{L \in X_n : |L| \geq r\}$. Then Ω_r is compact.

Proof. Let $L \in \Omega_r$. Choose $v_1, v_2, \ldots, v_n$ in the following way. v_1 is the shortest non-zero vector in L (if there are multiple such vectors choose any). Having chosen $(v_1, \ldots, v_i)$, v_{i+1} is the shortest vector in L linearly independent from the previous ones. Thus

$$0 < r \leq |v_1| \leq \ldots \leq |v_n|. \tag{3.7}$$

It is easy to show that the $\mathbb{Z}$ span of the set $\{v_1, \ldots, v_n\}$ has bounded index in L (a bound independent of L). If we show that $|v_k| \leq f_k(r)$, for $k = 1, \ldots, n$, then the theorem would follow. For then, v_i's would vary in a compact set and span a lattice of bounded covolume (because of (3.7)) which is contained in L with bounded index.

We prove this by induction on k. By applying Minkowski's convex body theorem with the closed

ball of radius r, we see that we take $f_1(r) = cr$, where c is a constant(depending only on n). For the inductive case, let V be the subspace spanned by $(v_1, \ldots, v_k)$. This subspace determines a k-torus

$$Y = V/(V \cap L) \subset X = \mathbb{R}^n/L.$$

Clearly, Y contains an embedded k-ball of radius $r/2$, so the volume of Y is bounded below by in terms of r. Also, by our hypothesis, Y has diameter at most $f_k(r)$.

Suppose $|v_{k+1}| \gg \text{diam}(Y)$. Then X would contain an embedded product of Y with a large $(n-k)$ ball. This will lead to a contradiction as X has volume 1. □

Corollary 3.21. (Mahler's compactness criterion) Let $E \subset X_n$. Then $\overline{E}$ is compact if and only if $\inf\{|L| : L \in E\} > 0$.

Proof. From the previous theorem we have that $\bigcup_r \Omega_r$, $r > 0$ is an increasing union of compact sets in X_n which cover the whole of X_n (as r decreases to zero). So $\overline{E}$ is compact if and only it is contained in some Ω_r for some $r > 0$. This is precisely a restatement of what we want to prove. □

There is a natural left action of G on X_n given by $[X] \mapsto [MX]$. Now, let H be a closed subgroup of G. Clearly μ is invariant for the action of H on X_n. As with the Oppenheim conjecture mentioned in the introduction, very often, we are interested in knowing the properties of H-orbits (closedness, cocompactness, compactness etc). Let A be the diagonal torus in G. That is

$$A = \{\text{diag}\,(t_1, \ldots, t_n),\ t_i > 0,\ t_1 t_2 \cdots t_n = 1\}.$$

In this section we shall be chiefly concerned with the action of A on X_n. We now define the notion of periodic orbits for A.

Definition 3.22. An A-orbit $Ax\Gamma$ is said to be *periodic* if it is compact.

More generally, for a closed subgroup H of G, an H-orbit is defined to be periodic if it carries a finite H-invariant measure. When H is abelian this definition coincides with the above one.

We shall give a complete characterization of periodic A-orbits in this section.

3.2.1 Lattices arising from number fields

Let K be a number field of degree n over $\mathbb{Q}$.

Definition 3.23. A *lattice* in K is the set of all integral linear combination of n $\mathbb{Q}$-linearly independent elements of K. In other words it is a $\mathbb{Z}$-submodule of K of rank n.

Let M be a lattice and let $\mu_1, \ldots, \mu_n$ be a $\mathbb{Z}$-basis for M. We define the *discriminant* of M, denoted by $\text{disc}(M)$, to be the discriminant of $\mu_1, \ldots, \mu_n$. (Recall discriminant of an n-tuple $\alpha_1, \ldots, \alpha_n$ in K is the determinant of the matrix $(\text{Tr}(\alpha_i \alpha_j))$.) Different bases for M are obtained by a unimodular transformation (i.e. an element of $GL_N(\mathbb{Z})$) and they have the same discriminant, so there is no ambiguity involved in the definition. Note that the discriminant of M is non-zero as $\mu_1, \ldots, \mu_n$ are linearly independent over $\mathbb{Q}$.

Two lattices M_1 and M_2 are said to be *similar* if there exists a non-zero $\alpha \in K$ such that $M_1 = \alpha M_2$. $\alpha \in K$ is called a *coefficient* of M if $\alpha M \subseteq M$, i.e. for any $\xi \in M, \alpha\xi \in M$. It is easily verified that the set of all coefficients of M form a ring. It is denoted by $\mathcal{O}_M$ and called the *coefficient ring* ring of M. We now prove that $\mathcal{O}_M$ is itself a lattice. We only need to prove that it has full rank. Let $\mu_1, \ldots, \mu_n$ be a basis for M. It is also a $\mathbb{Q}$-basis for K. Let $\alpha \in K$ be non-zero. Write

$$\alpha\mu_i = \sum_j a_{ij}\mu_j$$

where a_{ij}'s are in $\mathbb{Q}$. Let c be the least common denominator of all the a_{ij}. Then for each i, $c\alpha\mu_i \in M$, and hence $c\alpha \in \mathcal{O}_M$. So if we start with a basis $\alpha_1, \ldots, \alpha_n$ for K, then we can find an integer c such that each of $c\alpha_i$ is in $\mathcal{O}_M$. It follows that $\mathcal{O}_M$ has full rank and hence a lattice.

For any $\gamma \neq 0$ the condition $\alpha M \subseteq M$ is equivalent to the condition $\alpha\gamma M \subseteq \gamma M$. This implies that similar lattices have same coefficient ring, i.e. $\mathcal{O}_M = \mathcal{O}_{\gamma M}$.

Notice that any element α of $\mathcal{O}_M$ is actually an algebraic integer (as the condition $\alpha M \subseteq M$ is equivalent to saying that α satisfies a monic polynomial with coefficients in $\mathbb{Z}$). Hence $\mathcal{O}_M$ is actually an order in the ring of integers, $\mathcal{O}_K$, of K. If $M \subset \mathcal{O}_M$, then M is actually an ideal in $\mathcal{O}_M$. On the other hand given any lattice M it is easy to find (using the least common denominator trick) a non-zero integer b such that $bM \subset \mathcal{O}_M$. We summarize these results.

Theorem 3.24. The coefficient ring of any lattice in a number field K is an order in this field. Coefficient rings of similar lattices coincide. Any lattice is similar to an ideal contained in its coefficient ring.

Let $\mathcal{O}$ be an order in $\mathcal{O}_K$. Then $\mathcal{O}$ is the coefficient ring for some lattice in K (for example one could take the lattice to be $\mathcal{O}$ itself). We are only interested in lattices up to similarity. We now state a basic and fundamental result in algebraic number theory, proof of which could be found in [2].

Theorem 3.25. Let $\mathcal{O}$ be any order in a number field K. Then there only finite many equivalence class of similar modules with $\mathcal{O}$ as their coefficient ring.

From now on, we shall assume that K is a totally real field of degree n. (Recall: a totally real field is a field all whose embeddings into $\mathbb{C}$ is actually into $\mathbb{R}$). Let $\sigma_1, \ldots, \sigma_n$ be the distinct embeddings K into $\mathbb{R}$. Recall the *geometric embedding* of $\theta : K \hookrightarrow \mathbb{R}^n$ given by

$$\theta(x) = (\sigma_1 x, \ldots, \sigma_n x). \tag{3.8}$$

Proposition 3.26. Let M be a lattice in K. Then $\theta(M)$ is a lattice in $\mathbb{R}^n$.

Proof. Let $\alpha_1, \ldots, \alpha_n$ be a $\mathbb{Z}$-basis for M. It is enough to show that $\theta(\alpha_1)$, $\ldots$, $\theta(\alpha_n)$ are linearly independent over $\mathbb{R}$. Consider the matrix, B, corresponding these n vectors, namely $b_{ij} = \sigma_i\alpha_j$. $(\det B)^2$ is nothing but the discriminant of M and hence non-zero. The proposition follows. □

The *logarithmic embedding* of K into $\mathbb{R}^n$ given by

$$Log(x) = (\log|\sigma_1 x|, \ldots, \log|\sigma_n x|). \tag{3.9}$$

If η is a unit in $\mathcal{O}_K$, then $Log(\eta)$ lies in the $n-1$ dimensional subspace

$$\mathfrak{a} = \{(x_1, \ldots, x_n) : x_1 + \ldots + x_n = 0\}.$$

The converse is true as well, as $Log(x)$ being in $\mathfrak{a}$ is equivalent to $|N_{\mathbb{Q}}^K(x)| = 1$.

Let $\mathcal{O}$ be an order in K and let $\mathcal{O}^\times$ be the group of units in $\mathcal{O}$. We look at the restriction of the map Log on $\mathcal{O}^\times$. The Dirichlet unit theorem [2] states that image of $\mathcal{O}^\times$ under Log is a lattice in $\mathfrak{a}$. In particular, the group of units, as an abelian group, has rank $n-1$. We define the *regulator* of $\mathcal{O}$ to be the volume of the quotient $\mathfrak{a}/(Log\,\mathcal{O}^\times)$. (As can easily be seen, this definition of regulator is equivalent to the classical one). A unit η in $\mathcal{O}$ is said to be totally positive if $\sigma_i(\eta) > 0$ for all i. Clearly, totally positive units in $\mathcal{O}$, denoted by $\mathcal{O}_+^\times$, form a group and the Dirichlet unit theorem also holds for $\mathcal{O}_+^\times$. That is $Log(\mathcal{O}_+^\times)$ is a lattice in $\mathfrak{a}$. Our interest in totally positive units is partly due to the fact that if η is such a unit then $\mathrm{diag}(\theta(\eta)) \in A$, the diagonal group.

Let M be a lattice in K with coefficient ring $\mathcal{O}_M$. If $\gamma \in \mathcal{O}_M^\times$, then $\gamma M \subseteq M$ and $\gamma^{-1}M \subseteq M$ hence $\gamma M = M$. So the units in $\mathcal{O}$ are precisely the elements $\alpha \in K$ which stabilize M, i.e. $\alpha M = M$. Given a lattice M in K, we shall use the notation L_M to denote the unimodular lattice in $\mathbb{R}^n$ defined by

$$L_M = \alpha\theta(M), \quad \text{where } \alpha = |\det \theta(M)|^{-1/n}. \tag{3.10}$$

Lemma 3.27. The A-orbit of L_M is compact in X_n.

Proof. Let $\mathcal{O}$ be the coefficient ring of M. We have from the previous discussion that M is stabilized by each element in $\mathcal{O}_+^\times$. Therefore, L_M is stabilized by elements of A of the form $\mathrm{diag}(\theta(\gamma))$ where $\gamma \in \mathcal{O}_+^\times$. Let A_{L_M} denote the stabilizer of L_M under the action of A. There is a natural map $\log : A \to \mathfrak{a}$ (the inverse of $\exp : \mathfrak{a} \to A$) which takes A_{L_M} to $Log(\mathcal{O}_+^\times)$. But, by Dirichlet's unit theorem, $\mathfrak{a}/Log(\mathcal{O}_+^\times)$ is compact (isomorphic to $\mathbb{T}^{n-1}$). We conclude that A/A_{L_M} is compact. As $AL_M \simeq A/A_{L_M}$ (the orbit stabilizer theorem), we deduce that AL_M is compact. □

We say that a lattice L in $\mathbb{R}^n$ *arises from a number field* if $AL = AL_M$, where M is a lattice in a totally real number field $K/\mathbb{Q}$ of degree n. Observe that if $M_1 = \gamma M_2$ with $\gamma \in K\backslash\{0\}$, then L_{M_1} and L_{M_2} are in the same A-orbit. More specifically, we have

$$L_{M_1} = \beta\,\mathrm{diag}(\sigma_1\gamma, \ldots, \sigma_n\gamma)\,L_{M_2}, \quad \beta = |N_{\mathbb{Q}}^K(\gamma)|^{-1/n}. \tag{3.11}$$

Let $\alpha_1, \ldots, \alpha_n$ be a basis for M. Define the *norm form*, F, on M by

$$F(x_1, \ldots, x_n) = \mathrm{N}_{\mathbb{Q}}^K(x_1\alpha_1 + \ldots + x_n\alpha_n). \tag{3.12}$$

It is a easy to see that F is a homogeneous polynomial in $x_1, \ldots, x_n$ of degree n with integral coefficients. Choosing a different basis for M will only change $F(x_1, \ldots, x_n)$ to $F(x_1', \ldots, x_n')$, where $(x_1, \ldots, x_n)$ is related to $(x_1', \ldots, x_n')$ by a unimodular transformation. So up to this equivalence, F is uniquely determined by M. We use the notation (M, F) to denote the norm form associated to M. With a little more work it can be shown that F is irreducible over $\mathbb{Q}$. Conversely any homogenous

polynomial in $x_1, \ldots, x_n$ of degree n irreducible over $\mathbb{Q}$ which splits over $\mathbb{R}$ is of the form (3.12) (see [2]).

More generally, given any lattice $L \subset \mathbb{R}^n$, we can associate a form to it as follows. Choose a basis $v_1, \ldots, v_n$ for L. Let L_i be the linear form defined as

$$L_i(x_1, \ldots, x_n) = (\sum_{j=i}^{n} x_j v_j)_i,$$

where the notation $(y)_i$ means the i-th coordinate of y. We define the *product form* associated to L, denoted as (L, F), by

$$F(x_1, \ldots, x_n) = \prod_i L_i(x_1, \ldots, x_n). \tag{3.13}$$

It is easily verified that if L is a lattice arising from a number field, then the product form agrees with the norm form.

The next theorem is rather interesting in that it shows every A-periodic orbit comes from a number field.

Theorem 3.28 ([19])**.** For any unimodular lattice $L \subset \mathbb{R}^n$ the following conditions are equivalent.

1. AL is periodic.
2. L arises from a number field.
3. The pair (L, F) is equivalent to $(\mathbb{Z}^n, \alpha f)$ where $\alpha \in \mathbb{R}$ and f is an integral form that is irreducible over $\mathbb{Q}$.

Proof. $(2 \Rightarrow 1)$. Suppose $A.L$ is compact and let A_L denote the stabilizer of L in A. Then $L \otimes \mathbb{Q}$ is a module over the commutative algebra $R = \mathbb{Q}[A_L] \subset M_n(\mathbb{R})$. The matrices in A have only real eigen values and so are semi-simple. By the structure theorem for semi-simple rings we get that R is the direct sum of m totally real fields and therefore the rank of $\mathcal{O}_R^\times$ is $n - m$. Since $A.L$ is compact $A_L \cong \mathbb{Z}^{n-1}$. (Otherwise A_L will have infinite index in A and $A.L \cong A/A_L$ cannot be compact). Now, since $A_L \subset \mathcal{O}_R^\times$ we conclude that $m = 1$ and R itself is a totally real field. Thus $L \otimes \mathbb{Q}$ is a one dimensional vector space over R, so the lattice L itself is obtained from a full module $M \subset R$ by the process described above. We have already seen from lemma 3.27 that $(1 \Rightarrow 2)$.

(2) and (3) are equivalent by our preceding discussion. □

This theorem allows us to associate two important invariants to periodic A-orbits.

Definition 3.29. Let $x \in X_n$ be such that Ax is periodic. Then by the theorem above, $Ax = AL_M$ where M is a lattice in a totally real number field of degree n. We define the *discriminant* of Ax to be $\mathrm{disc}(Ax) = \mathrm{disc}(\mathcal{O}_M)$, and we define the *volume* of Ax to be $\mathrm{vol}(Ax) = \mathrm{reg}(\mathcal{O}_M)$.

Remark 3.30. We see that periodic A-orbits in X_n come naturally in packets. Each packet corresponds to an order $\mathcal{O}$ in a totlally real field K of degree n and consists of orbits of the form AL_M where M is a lattice in K, up to similarity, with $\mathcal{O}$ as its coefficient ring.

3.2.2 Discreteness of periodic orbits.

The main purpose of this section is to prove that if $n \geq 3$, periodic orbits of A of discriminant less than D cannot be too close in X_n. We shall also explain the duality between lattices and linear forms and use this to put the Littlewood conjecture (or rather a generalization of it due to Cassels and Swinnerton-Dyer) in the perspective of the A action on X_n. Our treatment is based on [3], [20] and [22].

Let $N : \mathbb{R}^n \to \mathbb{R}$ denote *norm* function defined by $N(x) = \prod_1^n x_i$. For a unimodular lattice $L \subset \mathbb{R}^n$ we define norm of the lattice by

$$N(L) = \inf\{|N(w)| \,:\, w \in L, w \neq 0\}.$$

Clearly, the function N is constant on A-orbits.

Proposition 3.31. The function $N : X_n \to \mathbb{R}$ is semicontinuous, i.e. if $L_n \to L$ in X_n, then $\limsup N(L_n) \leq N(L)$.

Proof. Define the *star body* of radius $\epsilon > 0$ by $S_\epsilon = \{x \in \mathbb{R}^n : |N(x)| < \epsilon\}$. In terms of star bodies, $N(L) = \inf\{\epsilon : S_\epsilon \cap L \neq \{0\}\}$. Since $L_n \to L$, given any $\epsilon > 0$, all but finitely many L_n's will intersect $S_{N(L)+\epsilon}$. In other words $N(L_n) \leq N(L) + \epsilon$, for all n large enough. The proposition follows. □

Proposition 3.32. Let $L \subset \mathbb{R}^n$ be a unimodular lattice. Then $\overline{AL}$ is compact if and only if $N(L) > 0$.

Proof. By AM-GM inequality we have that $|x| \geq \sqrt{n}N(x)^{1/n}$ for any $x \in \mathbb{R}^n$; also, if $N(x) \neq 0$, the equality holds for some $y \in Ax$. So, if $N(L) > r > 0$ we would also have that $|L| > \sqrt{n}r^{1/n} = r'$. But $N(L) = N(aL)$ for any $a \in A$, so $|aL| > r'$ for any a. By Mahler's criterion we conclude that $\overline{AL}$ is compact.

Conversely suppose $\overline{AL}$ is compact. Then there exists $r > 0$ such that $|aL| > r$ for all $a \in A$. If $x \neq 0$ in L we can find an a such that

$$N(x) = N(ax) = |ax|^n/n^{n/2} \geq r^n/n^{n/2} > 0.$$

So $N(L) > 0$. □

Remark 3.33. $\overline{AL}$ being compact is equivalent to saying that AL is bounded. So the proposition may be restated as AL is bounded if and only if $N(L) > 0$.

We now turn towards computing the norm of lattices arising from number fields. Observe that if $\theta(\alpha) = x$ then $N(x) = N_{\mathbb{Q}}^K(\alpha)$. Using (3.10), we find that if M is a lattice in K then $N(L_M) = \alpha^n N(\theta(M))$. But $\alpha^n = |\det \theta(M)| = |\mathrm{disc}(M)|^{-1/2}$. And $N(\theta(M)) = \inf\{N_{\mathbb{Q}}^K(\beta) : \beta \in M, \beta \neq 0\}$. Since similar lattices are related as in (3.11), $N(L_M)$ depends only on the similarity class of M. Assume then that $M = I$, an ideal in its coefficient ring $\mathcal{O}_I$. We have that $\mathrm{disc}(I) = \mathrm{disc}(\mathcal{O}_I)N(I)^2$, where $N(I)$ is the familiar multiplicative function $[\mathcal{O}_I : I]$ on ideals. Set $N^*(I) = \min\{|N_{\mathbb{Q}}^K(\beta)| :$

$\beta \in I, \beta \neq 0\}$. We then see that

$$N(L_I) = \frac{N^*(I)}{N(I)\sqrt{|\text{disc}(\mathcal{O}_I)|}}. \tag{3.14}$$

Suppose $\beta \in I$ be an element such that $|N^K_{\mathbb{Q}}(\beta)| = N^*(I)$. But $|N^K_{\mathbb{Q}}(\beta)| = N(\beta) = N(I)[I : (\alpha)]$. We deduce that

$$N(L_I) \geq \frac{1}{\sqrt{|\text{disc}(\mathcal{O}_I)|}}. \tag{3.15}$$

We some times use $N(L)$ to denote the set of values $N(l)$ taken by $l \in L$. The meaning will be clear from the context.

We now turn towards the main result in this section: isolation of periodic orbits. First, we construct the necessary set up for the proof. For each pair $1 \leq i \neq j \leq n$, we define the *root* $\alpha_{ij} : \mathfrak{a} \to \mathbb{R}$ to be the linear functional

$$\mathbf{t} \mapsto t_i - t_j.$$

The set of roots will be denoted by Φ.

Recall the logarithmic embedding (3.9) $Log : \mathcal{O}_K^\times \to \mathfrak{a}$. For any order $\mathcal{O}$ in K, let us denote the image of $\mathcal{O}^\times$ under Log by $\Omega_{\mathcal{O}}$.

Lemma 3.34. ([22]) Let $n \geq 3$. Then, for any root $\alpha_{ij} \in \Phi$, the set $\{\alpha_{ij}(a) : a \in \Omega_{\mathcal{O}}\}$ is dense in reals.

Proof. Since $\Omega_{\mathcal{O}}$ has finite index in Ω_K, it is enough to justify why $\alpha(\Omega_K)$ is dense in $\mathbb{R}$. As Ω_K is a lattice is $\mathfrak{a}$, this is equivalent $\Omega_K \cap \ker(\alpha_{ij})$ not being a lattice in $\ker(\alpha_{ij})$. If this is indeed the case then the field $L = \{\theta \in K : \sigma_i(\theta) = \sigma_j(\theta)\}$ is a subfield of K with a group of units containing a copy of $\mathbb{Z}^{n-2}$. Since L has degree at most $n/2$, by Dirichlet's unit theorem, the degree of the group of units in L is at most $n/2 - 1$. This means that $n/2 - 1 \geq n - 2$, or equivalently $n \leq 2$, a contradiction. □

Corollary 3.35. Let $\mathcal{O}$ be an order in a totally real field K of degree $n \geq 3$. Then for any $i \neq j$, the set

$$\left\{ \frac{\sigma_i(\eta)}{\sigma_j(\eta)} : \eta \in \mathcal{O}_+^\times \right\}$$

is a dense subset of $\mathbb{R}_+$.

Proof. $Log(\mathcal{O}_+^\times)$ has finite index in Ω_K, so $\alpha_{ij}(\mathcal{O}_+^\times)$ is dense in $\mathbb{R}$. Applying $\exp : \mathbb{R} \to R_+$ and using the definition of Log map (3.9) gives the desired result. □

An element a in A is said to be *regular* if $\alpha(a) \neq 0$ for any root $\alpha \in \Phi$; equivalently all entries of a are distinct. Let $b \in A$ be regular. We define the *stable horospherical subgroup* corresponding to b to be

$$U^-(b) = \{g \in G : b^n g b^{-n} \to e \text{ as } n \to \infty\},$$

and the *unstable horospherical subgroup* to be

$$U^+(b) = \{g \in G : b^{-n} g b^n \to e \text{ as } n \to \infty\}.$$

The following lemma explains their usefulness.

Lemma 3.36. Let $b \in A$ be regular. Any element $g \in A$ which is close enough to e has a unique decomposition $g = au^+u^-$, where $a \in A, u^+ \in U^+(b), u^- \in U^-(b)$.

Proof. (Sketch). Let $b = \mathrm{diag}(b_1, \ldots, b_n)$. The ij-th entry in bgb^{-1} is $g_{ij}b_i/b_j$. As there is no change to diagonal entries on conjugation, we assume from now on that $i \neq j$. If g_{ij} is non-zero, repeated conjugation will take the corresponding entry to 0 or ∞ depending on whether $b_i/b_j < 1$ or $b_i/b_j > 1$. This is equivalent to the condition $\alpha_{ij}(b) < 0$ or $\alpha_{ij}(b) > 0$ respectively. ($\alpha_{ij}(b) = 0$ is ruled out as b is regular). In particular, u belongs to $U^-(b)$ if and only of all its diagonal entries are 1 and $u_{ij} = 0$ for any ij such that $\alpha_{ij}(b) > 0$. We then see that $U^-(b)$ is of rank $\binom{n}{2}$, half the number of roots. Similar statement holds true for $U^+(b)$ (in fact, $U^+(b)$ is the transpose of $U^-(b)$).

Consider the map from $A \times U^+(b) \times U^-(b) \longrightarrow G$, given by

$$(a, u^+, u^-) \mapsto au^+u^-.$$

It can be shown, by using the aforementioned facts, that the Jacobian of the above map is non-singular at e. By the inverse function theorem, there exists neighborhoods $W^0(b), W^+(b), W^-(b)$ of the identity elements in the groups A, $U^+(b)$ and $U^-(b)$ respectively, such that the above map is a diffeomorphism onto its image. □

From now on, for simplicity, we shall assume $n = 3$. We shall later explain how the results generalize to any $n \geq 3$.

Theorem 3.37. (**Isolation of periodic orbits**). Let $L_0 \in X_3$ be periodic (i.e. AL_0 is compact). Suppose $L \in X_n$ be another lattice such that $\overline{AL}$ intersects AL_0. Then either $AL_0 = AL$ or $N(L)$ is dense in $\mathbb{R}$.

Proof. Suppose $AL_0 \neq AL$, then $L_0 \in \overline{AL} \backslash AL$. We will show that $N(L)$ is dense in $\mathbb{R}$. By semicontinuity of N, it is sufficient to show that $N(L')$ is dense for some $L' \in \overline{AL}$. Note that both AL_0 and $\overline{AL}$ are A-invariant.

Let $V \subseteq G = SL_3(\mathbb{R})$ denote the set of g such that $gAL_0 \cap \overline{AL} \neq \emptyset$. As AL_0 is compact, V is closed. For a, b in A we have

$$agbAL_0 \cap \overline{AL} = a(gAL_0 \cap \overline{AL}),$$

so V is invariant under the action of A from both left and right.

Since L_0 is periodic, from theorem 3.24 , it is of the form aL_I for some ideal I in a totally real order $\mathcal{O}$. Further, since $\overline{AL}$ is invariant under A, we may as well assume $a = 1$. Let $A_0 = \mathrm{stab}_A(L_I)$. Elements in A_0 are precisely of the form $\mathrm{diag}(\sigma_1\eta, \sigma_2\eta, \sigma_3\eta)$ where $\eta \in \mathcal{O}_+^\times$, the group of totally positive units in $\mathcal{O}$. Fix a regular element $b \in A_0$.

Now, choose a sequence of lattices $L_n \to L_0$ from the orbit AL. Write $L_n = g_nL_0$ where $g_n \in G$ and $g_n \to e$. From the previous lemma we have a decomposition $g_n = a_nu_n^+u_n^-$, where $a_n \in A, u_n^+ \in U^+(b), u_n^- \in U^-(b)$, and $a_n, u_n^+, u_n^- \to e$. Replacing g_n by $a_n^{-1}g_n$, if necessary, we may further assume that $a_n = 1$. Since $g_nL_0 \in AL$, we must have that $g_n \neq e$ for any n (as we have

assumed $AL_0 \neq AL$). This condition is equivalent to saying that the pairs (u_n^+, u_n^-) are non-trivial. We can then write

$$g_n = (I + v_n^+)(I + v_n^-),$$

where v_n^+ and v_n^- are matrices with diagonal entries zero. Further, we also have that diagonal entries of $v_n^+ v_n^-$ is zero (owing to the complementarity in structure of the groups $U^+(b)$ and $U^-(b)$). We deduce that

$$g_n = I + v_n,$$

where v_n is a non-zero matrix with diagonal entries zero. Since $g_n \to e$, we must have $v_n \to 0$. Let $\|v_n\|$ denote the maximum of the absolute values of the entries of v_n. As n varies from 1 to ∞, there must be some entry ij which achieves $\|v_n\|$ infinitely often. We pass to this subsequence and abusing notation, we denote it again by v_n. For concreteness, assume that the maximum is achieved at $(v_n)_{12}$, i.e., $\|v_n\| = |(v_n)_{12}|$. Further assume, passing to a subsequence if necessary, that all $(v_n)_{12}$ are of the same sign, say positive.

Conjugating by $a = \mathrm{diag}(a_1, a_2, a_3)$ sends $(v_n)_{ij}$ to $(v_n)_{ij} a_i/a_j$. In particular, conjugating by $\mathrm{diag}(a, 1/a, 1)$ multiplies $(v_n)_{ij}$ by the ij-th entry of the matrix

$$\begin{pmatrix} 1 & a^2 & a \\ a^{-2} & 1 & a^{-1} \\ a^{-1} & a & 1 \end{pmatrix}.$$

Given $t > 0$, take a_n such that

$$a_n^2 (v_n)_{12} = t.$$

Since $\|v_n\| \to 0$ we have that $a_n \to \infty$. Further, as $(v_n)_{12}$ is the largest entry, we conclude that the matrix obtained by conjugating g_n with $\mathrm{diag}(a_n, 1/a_n, 1)$ converges to

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

As V is closed we deduce that V contains the semigroup $U = \{u_t : t > 0\}$. Note that U is normalized by A, i.e. $aUa^{-1} = U$ for all $a \in A$.

Since $u_t \in V$, we have that $u_t a(L_0) \in \overline{AL}$ for some $a \in A$. But U is normalized by A, and thus $u_{t_0}(L_0) \in \overline{AL}$ for some $t_0 > 0$. In order to complete the proof, it suffices to show that $N(u_{t_0} L_0)$ is dense in $\mathbb{R}$. Let $x \in L_0$.

$$N(u_{t_0} x) = (x_1 + t x_2) x_2 x_3 = N(x)(1 + t x_2/x_1).$$

Let $c = \mathrm{diag}(c_1, c_2, c_3) \in A_0$, where A_0 is the stabilizer of L_0 in A. Then $cx \in L_0$ and $N(cx) = N(x)$. We infer that

$$N(u_{t_0} cx) = N(x)\big(1 + t \frac{c_2}{c_1} \frac{x_2}{x_1}\big).$$

From corollary 3.35 we have that $\{c_2/c_1 : c \in A_0\}$ is a dense subset of $\mathbb{R}_+$. Choosing yet another x

with a different sign for x_2/x_1, we conclude that $N(u_{t_0}L_0)$ is dense in $\mathbb{R}$.

By semicontinuity of N, it follows that $N(L)$ is dense in $\mathbb{R}$. □

Remark 3.38. The above theorem is not true for $n = 2$. That is, we can construct bounded A orbits in X_2 which spiral in or oscillate between periodic orbits. The reason for the failure of the above theorem is that corollary 3.35 is not valid for $n = 2$. In fact, for any order $\mathcal{O}$ is a real quadratic field, the set $\{\sigma_1\eta/\sigma_2\eta : \eta \in \mathcal{O}^\times\}$ is a discrete subset of $\mathbb{R}$.

We can strengthen theorem (3.37), using techniques similar to the ones used in its proof, as follows:

Theorem 3.39. (**Discreteness of periodic orbits** [3]) Let $n \geq 3$. Suppose $AL_0 \in X_n$ is periodic. Let $(L_k) \in X_n \backslash AL_0$ be a sequence of lattices such that $L_k \to L_0$. Then given any $\epsilon > 0$, we can find, for all $k \geq M(\epsilon)$, $l_k \in L_k$ such that $0 < |N(l_k)| < \epsilon$. In particular, as $k \to \infty$, $N(L_k) \to 0$.

We now describe how the above theorem can be interpreted as the "discreteness of periodic orbits". Let AL_0 be periodic, and let L_{I_k} be a sequence of lattices arising from totally real cubic number fields which converge to L_0. In the light of (3.15) and the theorem above, we conclude that $|\mathrm{disc}(\mathcal{O}_{I_k})| \to \infty$. In other words, qualitatively speaking, periodic orbits whose discriminants are bounded from above cannot be too close in X_n.

We remark that in [3] only $n = 3$ is treated, but the general case ($n \geq 3$) is similar. The following is a conjecture first stated in [3], for $n = 3$, which was later generalized by Margulis.

Conjecture 3.40. (Margulis) Let $L \in X_n$, $n \geq 3$. If $N(L) > 0$, then L arises from a number field. Equivalently, bounded A-orbits in X_n are periodic.

We now state the Littlewood conjecture, an outstanding open problem concerning simultaneous Diophantine approximation of real numbers.

Conjecture 3.41. (Littlewood) For any $\alpha, \beta \in \mathbb{R}$ we have

$$\liminf_{n\to\infty} n \, \|n\alpha\| \, \|n\beta\| = 0,$$

where $\|x\| = \min_{n\in\mathbb{Z}} |x - n|$

Theorem 3.42. Margulis's conjecture implies the Littlewood conjecture.

Proof. Suppose $(\alpha, \beta) \in \mathbb{R}^2$ be a counterexample to the Littlewood conjecture, i.e.

$$|n(n\alpha - a)(n\beta - b)| \geq \delta > 0, \tag{3.16}$$

for all integers $n \neq 0$, a, b. Consider the unimodular lattice $L_0 \in \mathbb{R}^3$ generated by

$$\{e_1, e_2, e_3\} = \{(1, 0, 0), (0, 1, 0), (\alpha, \beta, 1)\}.$$

Let $M_0 = \mathbb{Z}e_1 \cup \mathbb{Z}e_2$. From (3.16) we see that the norm is bounded away from zero on $L_0 - M_0$ and vanishes exactly on M_0. In particular $N(L_0)$ is not dense.

We say a lattice $L \in \mathbb{R}^3$ is *admissible* for a region $V \in \mathbb{R}^3$ if $L \cap V = \{0\}$. From (3.16), we see that L_0 is admissible for the region

$$|xyz| < \delta, \ \ \max(|x|, |y|) < 1.$$

Let $a_n = \mathrm{diag}(n, n, n^{-2})$. Then $L_n = a_n L_0$ is admissible for the region

$$|xyz| < \delta, \ \ \max(|x|, |y|) < n.$$

Note that all these lattices L_n are admissible for $B_{\delta'}$, the open ball of radius δ', for some $\delta' > 0$. (One could take $\delta' = 3\delta^{2/3}$, assuming $\delta < 1$). By Mahler's compactness criterion, the sequence (L_n) lies in a compact subset of X_3. Let L_∞ be a limit point of this sequence. Clearly, L_∞ is admissible for

$$|xyz| < \delta.$$

In other words, $N(L_\infty) \geq \delta$. By Margulis's conjecture, $L_\infty \in T$ for some periodic orbit T. But $L_\infty \in \overline{AL_0}$ and AL_0 is unbounded (as $N(L_0) = 0$). This contradicts theorem 3.37 as $N(L_0)$ is not dense □

We now state yet another conjecture by Margulis. In the next section we shall explain some recent progress made towards it and its implications for strengthening the Minkowski bound on class numbers.

Conjecture 3.43. For any compact set $\Omega \subset X_n$, $n \geq 3$, there are only finitely many periodic A-orbits contained in Ω.

Theorem 3.44. Margulis's conjecture implies conjecture 3.43.

Proof. Suppose there were infinitely many lattices $L_n \in \Omega$ such that AL_n is periodic. From corollary 3.21 we have that $\Omega \subset \Omega_\delta$ for some $\delta > 0$. Then $N(L_n) > r > 0$, for some $r = r(\delta)$ for all n. Let L_0 be a limit point of L_n's. Clearly, $N(L_0) > r$. If Margulis's conjecture is true, then AL_0 is periodic. But now there is subsequence of lattices, whose norms are bounded away from zero, converging to L_0. This is in flat contradiction with theorem 3.39, and we conclude that if conjecture 3.43 is false then Margulis's conjecture is false. □

Remark 3.45. Margulis's conjecture says that if $n \geq 3$, A-orbits in X_n are either compact or unbounded. Using theorem 3.28 one can reformulate the conjecture as follows:
Let $F(x_1, \ldots, x_n)$ be a product of linear forms in n-variables over $\mathbb{R}$ with $n \geq 3$. If F is not proportional to a homogeneous polynomial with integer coefficients, then

$$\inf_{0 \neq x \in \mathbb{Z}^n} |F(x)| = 0.$$

In fact, for $n = 3$, it is in this form that Cassels and Swinnerton-Dyer states this conjecture in [3].

3.3 Application: Strengthening Minkowski's theorem

In this section we shall describe some recent progress made towards conjecture 3.43 in [6]. Following [6], we then explain an interesting application of this work in strengthening Minkowski's theorem regarding ideal classes, which we now recall.

Theorem 3.46. (Minkowski) Let K be a number field of degree n with maximal order $\mathcal{O}_K$. Then any ideal class in $\mathcal{O}_K$ possesses a representative $J \subset \mathcal{O}_K$ of norm $N(J) = O(\sqrt{\operatorname{disc} K})$, where the O-constant depends only on n.

Note that finiteness of the ideal-class group follows from the above theorem.

We shall be working with the group $PGL_n(\mathbb{R})$. As mentioned before, the proofs could easily be translated back to $SL_n(\mathbb{R})$. Throughout this section, we shall denote $PGL_n(\mathbb{R})$ by G, $PGL_n(\mathbb{Z})$ by Γ and $PGL_n(\mathbb{R})/PGL_n(\mathbb{Z})$ by Y_n. For $x \in \mathbb{R}^n$, let $\|x\|_\infty$ denote the sup norm of x. We continue to denote the group of diagonal matrices by A. For $\delta > 0$, let Ω'_δ denote the set of homothety classes of lattices $\Lambda \subset \mathbb{R}^n$ containing no vectors v with $\|v\|_\infty^n < \delta \mathrm{covol}(\Lambda)$, i.e.,

$$\Omega'_\delta = \{g\Gamma \in Y_n : \|gx\|_\infty^n \geq \delta \det(g) \text{ for every } x \in \mathbb{Z}\}.$$

This set is compact by Mahler's compactness criterion. Also, $\Omega'_a \subset \Omega'_b$ if $a > b$.

Let K be a totally real number field, with $\mathcal{O}_K$ as its ring of integers, and let $[J]$ be an ideal class in $\mathcal{O}_K$. We denote the regulator of $\mathcal{O}_K$ by R_K. Define

$$m([J], K) = \min_{J' \in [J], J' \subset \mathcal{O}_K} N(J')$$
$$m(K) = \max_{[J]} m([J], K),$$

where in the later definition the maximum is taken over all ideal classes in $\mathcal{O}_K$. As before, let $\theta : K \hookrightarrow \mathbb{R}^n$ be the geometric embedding of K into $\mathbb{R}^n$. We shall now see that $m([J], K)$ is intimately related to how far the A-orbit of the homothety class of the lattice $\theta(J^{-1})$ penetrates the cusp of Y_n.

Lemma 3.47. Let J be a fractional ideal of K- a totally real number field of degree n. Let Y be the periodic A-orbit corresponding to the ideal class $[J^{-1}]$, i.e., $Y = A.\theta(J^{-1})$. Then the following are equivalent:

(1) $m([J], K) < \delta \operatorname{disc}(K)^{1/2}$;

(2) Y is not contained in Ω'_δ.

Proof. It is clear from the definition of Ω'_δ that if $\Lambda \subset \mathbb{R}^n$ is a lattice, then $A\Lambda \subset \Omega'_\delta$ if and only if, for all non-zero $x \in \Lambda$, we have $\prod_i |x_i| = |N(x)| \geq \delta \mathrm{covol}(\Lambda)$.

We apply this statement to the lattice $\Lambda = \theta(J^{-1})$. Recall that on this lattice the norm $N(\theta x)$ agrees with field norm $N_{\mathbb{Q}}^K(x)$ and thus, that N is a multiplicative function on the fractional ideals of K. Hence, the covolume Λ is $N(J)^{-1}(\operatorname{disc} K)^{1/2}$, where $N(J)$ is the norm of the ideal J. With this, we see that (2) is equivalent to the following

$$\text{There exists } x \in J^{-1} \text{ with } |N_{\mathbb{Q}}^K(x)| < \delta(\mathrm{disc} K)^{1/2}. \tag{3.17}$$

Consider the map $x \mapsto xJ$ form J^{-1} to ideal classes $I \subset \mathcal{O}_K$ equivalent to J. This map is surjective. We conclude that (3.17) is equivalent to condition (1). □

Next, we describe a consequence of conjecture 3.43.

Conjecture 3.48. Let $n \geq 3$ be fixed. Then any ideal class in a totaly real number field of degree n has a representative of norm $o(\sqrt{\operatorname{disc} K})$.

Theorem 3.49. Conjecture 3.43 implies conjecture 3.48.

Proof. Suppose conjecture 3.48 was false. Then for some $\delta > 0$ there would an infinite sequence of totally real fields K_i and ideals $J_i \subset \mathcal{O}_{K_i}$ with $m([J_i], K_i) \geq \delta\sqrt{\operatorname{disc} K_i}$. By lemma 3.47, this gives us an infinite sequence of periodic A-orbits all contained inside the compact set Ω'_δ, in contradiction to conjecture 3.43. □

Finally we describe the theorem of Einsiedler, Lindenstrauss, Michel and Venkatesh mentioned in the beginning of this section.

Theorem 3.50. ([6]) For any fixed compact set $\Omega \subset Y_n, n \geq 3$, and for any $\epsilon > 0$, the total volume of all periodic A-orbits contained in Ω of discriminant $\leq D$ is at most $O_\epsilon(D^\epsilon)$.

We shall not go into the proof of this theorem. We merely mention that the proof involves deriving a relationship between total volume of a collection of A-periodic orbits Y_i and the entropy of any weak limit of A-invariant probability measures, $\mu_i = \mu_{Y_i}$, supported on Y_i. The idea goes back to Linnik and the authors call it Linnik's principle.

Let $h_\delta(K)$ be the number of ideal classes in $\mathcal{O}_K$ with $m([J], K) > \delta \operatorname{disc}(K)^{1/2}$. Substituting conjecture 3.43 with theorem 3.50, we get the following unconditional result towards conjecture 3.48.

Theorem 3.51. Let $n \geq 3$, and let K denote a totally real number field of degree d. For all $\varepsilon, \delta > 0$ we have

$$\sum_{\operatorname{disc} K < X} R_K h_\delta(K) \ll_{\varepsilon,\delta} X^\varepsilon. \tag{3.18}$$

In particular, conjecture 3.48 is true for almost all totally real fields. That is, the number of fields K with discriminant $\leq X$ for which $m(K) \geq \delta \operatorname{disc}(K)^{1/2}$ is $O_\varepsilon(X^\varepsilon)$, for any $\varepsilon, \delta > 0$.

Proof. Suppose that theorem 3.51 were false. Then we have constants $C, \varepsilon, \delta > 0$ and a sequence of integers $D_i \to \infty$ such that

$$\sum_{\operatorname{disc}(K) < D_i} R_K h_\delta(K) > C D_i^\varepsilon, \quad \text{for all } i, \tag{3.19}$$

the summation being over totally real fields of fixed degree n. For every totally real field for which $m(K) \geq \delta\sqrt{\operatorname{disc} K}$, let $[J_{K,j}]$ $j = 1, \ldots, h_\delta(K)$, be the ideal classes of K with $m([J_{K,j}], K) \geq \delta\sqrt{\operatorname{disc} K}$.

Let $Y_{K,j}$ be the periodic A-orbit corresponding to $\theta(J_{K,j}^{-1})$. The volume of $Y_{K,j}$ is proportional to R_K and the discriminant is proportional to $\operatorname{disc}(K)$. By lemma 3.47 we have that $Y_{K,j} \subset \Omega'_\delta$.

The assumption (3.19) implies that the collection of periodic A-orbits

$$C_i = \{Y_{K,j} : \mathrm{disc}(K) < D_i, 1 \leq j \leq h_\delta(K)\} \subset \Omega'_\delta$$

have discriminant $\leq D_i$ and total volume $\gg D_i^\varepsilon$. Clearly, this contradicts theorem 3.50. $\square$

Bibliography

[1] M. B. Bekka and M. Mayer. *Ergodic theory and topological dynamics of group actions on homogeneous spaces*. London Mathematical Society Lecture Notes Series-269. Cambridge Univ. Press, 2000.

[2] Z. I. Borevich and I. R. Shaferevich. *Number theory*. Academic Press, 1966.

[3] J. W. S. Cassels and H. P. F. Swinnnerton Dyer. *On the product of three homogeneous linear forms and indefinite ternary quadratic forms*. Phil. Tran. Roy. Soc. London, Ser.A. 248: 73-96, 1955.

[4] W. Duke. *Rational points on the sphere*. Ramanujan J. No.1-3, 235-239, 2003.

[5] W. Duke. *An Introduction to Linnik Problems*. Equidistribution in Number theory, An Introduction, NATO Science Series, Springer, 2007.

[6] M. Einsiedler, E. Lindenstruass, P. Michelle, A. Venkatesh. *Distribution of periodic torus orbits on homogeneous spaces*. Duke Math. J. Vol. 148, No. 1, 2009, 119-174.

[7] J. Ellenberg, P. Michelle, A. Venkatesh. *Linnik's ergodic method and the distribution of integer points on the sphere*. Preprint.

[8] M. Einsiedler and T. Ward. *Ergodic theory with a view towards number theory*. Graduate Texts in Mathematics-259. Springer, 2011.

[9] C. F. Gauss. *Disquisitiones Arithmeticae*. Translated by A. Clarke. Springer-Verlag, New York, 1986.

[10] A. Granville and Z. Rudnick. *Uniform Distribution*. Equidistribution in Number theory, An Introduction, NATO Science Series, Springer, 2007.

[11] D.R. Heath-Brown. *Arithmetic applications of Kloosteman sums*. Nieuw Arch. Wiskd. (2000), no. 4, 380-384.

[12] M. N. Huxley. *Area, Lattice Points and Exponential Sums*, LMS Monographs. New Series, 13. Oxford University Press, 1996.

[13] H. Iwaniec. *Fourier coefficients of modular forms of half-integral weight*, Invent. Math. 1987, 385-401.

[14] H. Iwaniec. *Topics in automorphis forms* Graduate Studies in Mathematics, AMS, 1997.

[15] H. Kloosterman. *On the representation of numbers in the form* $ax^2 + by^2 + cz^2 + dt^2$. Acta. Math. 46 (1926), 407- 464.

[16] N. Kolitz. *Introduction to Elliptic curves and modular forms*. Graduate texts in mathematics-97. Springer-Verlag, 1984.

[17] E. Lindenstrauss. *Some examples how to use measure classification in number theory.* Equidistribution in Number theory, An Introduction, NATO Science Series, Springer, 2007.

[18] Yu. V. Linnik. *Ergodic properties of algebraic fields*, Translated from Russian by M. S. Keane. Springer-Verlag, 1968.

[19] C. T. McMullen. *Minkowski's conjecture, well-rounded lattices and topological dimension.* J. Amer. Math. Soc. Volume 18, no 3, 711-734.

[20] C. T. McMullen. *Hyperbolic manifolds, discrete groups and ergodic theory.* Harvard Course notes, 2007.

[21] T. Miyake. *Modular Forms*, Springer Monographs in Mathematics. Springer-Verlag, 1989.

[22] U. Shapira. *A solution to a problem of Cassels and diophantine properties of cubic numbers*, Ann. of Math.(2) 173 (2011), 543-557.

[23] G. Shimura. *On modular forms of half-integral weight*, Ann. of Math. **97** (1973), 440-481.

[24] P. Sarnak. *Some applications of Modular Forms*, Cambridge Tracts in Mathematics-99, Cambridge Univ. Press, 1990.

[25] G. Watson. *A treatise on Bessel functions*, Cambridge Univ. Press.

[26] H. Weyl. *Über die Gleichverteilung von Zahlen mod. Eins.* Math. Ann. 77 (1916), 313-352.