Additive and Combinatorial Number Theory

Ben Green

Mathematical Institute, Oxford
Email address: ben.green@maths.ox.ac.uk

2000 *Mathematics Subject Classification.* Primary

Contents

0.1. Overview

The aim of this course is to present classic results in additive and combinatorial number theory, showing how tools from a variety of mathematical areas may be used to solve number-theoretical problems.

It is divided into two main parts. Part 1 concerns fairly classical additive number theory. Highlights will include the classical theorem of Lagrange that every number is the sum of four squares, results on Waring's problem (every number is the sum of s perfect kth powers, where s is bounded as a function of k).

Part 2 concerns combinatorial number theory, and in particular that part of it which nowadays goes by the name of additive combiantorics. Highlights here are Roth's theorem that sets of integers with positive upper density contain infinitely many 3-term arithmetic progressions, the first interesting case of Szemerédi's theorem. We will also discuss a celebrated theorem of Freiman describing the structure of finite sets of integers A for which the number of distinct sums $\{a + a' : a, a' \in A\}$ is not much larger than the size of A. If time allows, we will hint at the application of this, in the work of Tim Gowers, to Szemerédi's theorem for progressions of length 4.

0.2. Notation

Throughout the course we will be using *asymptotic notation.* This is vital in handling the many inequalities and rough estimates we will encounter. Here is a summary of the notation we will see. We suggest the reader not worry too much about this now; we will gain plenty of practice with this notation. See also the first question on Sheet 0.

- $A \ll B$ means that there is an absolute constant $C > 0$ such that $|A| \leqslant CB$. In this notation, A and B will typically be variable quantities, depending on some other parameter. For example, $x + 1 \ll x$ for $x \geqslant 1$, because $|x + 1| \leqslant 2x$ in this range. It is important to note that the constant C may be different in different instances of the notation.
- $A = O(B)$ means the same thing.
- $A \ll B$ is the same as $B \gg A$.
- $O(A)$ means some quantity bounded in magnitude by CA for some absolute constant $C > 0$. In particular, $O(1)$ simply means a quantity bounded by an absolute positive constant. For example, $\frac{5x}{1+x} = O(1)$ for $x \geqslant 0$.
- $A = o(B)$ means that $|A| \leqslant \varepsilon B$ as some other parameter becomes large enough. The other parameter will usually be clear from context. For example, $\frac{1}{\log x} = o(1)$ (as $x \to \infty$).

- Sometimes we write $o(1)$ by itself to be some quantity tending to zero (as some other parameter, invariably clear from context, tends to infinity). For example, $X^{o(1)}$ means a quantity that is eventually less than X^{ε} for any $\varepsilon > 0$, as $X \to \infty$.
- We will occasionally use $A \asymp B$ to mean that $c_1 A \leqslant B \leqslant c_2 A$ for some $c_1, c_2 > 0$.

We shall adopt the very standard notation

$$e(t) := e^{2\pi i t}.$$

We shall also write $\mathbf{T} = \mathbf{R}/\mathbf{Z}$, and for $\theta \in \mathbf{T}$ we write $\|\theta\|_{\mathbf{T}}$ for the distance of θ from 0. Thus, for example, $\|2/3\|_{\mathbf{T}} = 1/3$.

Finally, we will be using the concept of a *sumset*. If A, B are subsets (usually finite) of some abelian group G then we write

$$A + B := \{a + b : a \in A, b \in B\}$$

and

$$A - B := \{a - b : a \in A, b \in B\}.$$

These definitions extend in an obvious way to more than two summands, for example

$$A_1 + \cdots + A_k := \{a_1 + \cdots + a_k : a_i \in A_i\}.$$

If $A_1 = \cdots = A_k = A$ then we usually write kA for $A_1 + \cdots + A_k$. In particular, $2A = A + A$. We also write, e.g. $2A - 2A$ for $\{a_1 + a_2 - a_3 - a_4 : a_1, \dots, a_4 \in A\}$.

0.3. Quantities

In understanding analytic number theory, it is important to develop a robust intuitive feeling for the rough size of certain quantities. For example, one should be absolutely clear about the fact that, for X large,

$$\log^{10} X \lll e^{\sqrt{\log X}} \lll X^{0.01}.$$

CHAPTER 1

Fourier analysis

Fourier analysis is a pervading theme in this course. The natural habitat for the Fourier transform is a *locally compact abelian group* (LCAG) G. In this course, G will be one of the following examples:

- $G = \mathbf{R}$;
- $G = \mathbf{Z}$;
- $G = \mathbf{Z}/q\mathbf{Z}$ for some $q \in \mathbf{N}$.

We will not be developing a general theory, and in fact we will not even give the definition of an LCAG, though the reader may look it up.

A *character* on G is a continuous homomorphism $\chi : G \to \mathbf{C}^*$. The set of all characters forms an abelian group under pointwise multiplication. This is denoted by $\hat{G}$ and is called the dual group of G. For the groups G listed above, here are the characters on G. Whilst it is true that we are listing *all* characters on each group, we do not need to separately confirm that here (though it is not hard to show, especially, for $G = \mathbf{Z}$ and $G = \mathbf{Z}/q\mathbf{Z}$).

- $\hat{\mathbf{R}} = \mathbf{R}$. An isomorphism $\mathbf{R} \to \hat{\mathbf{R}}$ is given by

$$\xi \mapsto (x \mapsto e(\xi x)). \tag{1.1}$$

- $\hat{\mathbf{Z}} = \mathbf{T}$. An isomorphism $\mathbf{T} \to \hat{\mathbf{Z}}$ is given by

$$\theta \mapsto (n \mapsto e(\theta n)) \tag{1.2}$$

- $\widehat{\mathbf{Z}/q\mathbf{Z}} = \mathbf{Z}/q\mathbf{Z}$. An isomorphism $\mathbf{Z}/q\mathbf{Z} \to \widehat{\mathbf{Z}/q\mathbf{Z}}$ is given by

$$r \mapsto (x \mapsto e(rx/q)). \tag{1.3}$$

In each of the three examples listed above, the group G carries a natural measure μ.

- When $G = \mathbf{R}$, μ is Lebesgue measure;
- When $G = \mathbf{Z}$, μ is counting measure (that is, the measure of each integer is 1);
- When $G = \mathbf{Z}/q\mathbf{Z}$, μ is the *normalised* counting measure (that is, the measure of each element is $\frac{1}{q}$.

Given a "nice" function

$$f : G \to \mathbf{C},$$

its Fourier transform

$$\hat{f} : \hat{G} \to \mathbf{C}$$

is defined by

$$\hat{f}(\chi) := \int f(x)\overline{\chi(x)}d\mu(x).$$

We assume, henceforth, that the duals $\hat{G}$ of $\mathbf{R}, \mathbf{Z}, \mathbf{Z}/q\mathbf{Z}$ have been identified with $\mathbf{R}, \mathbf{T}, \mathbf{Z}/q\mathbf{Z}$ respectively, as in (1.1), (1.2), (1.3) above. Thus, in each case, the Fourier transforms are as follows:

- If $f : \mathbf{R} \to \mathbf{C}$, then
$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e(-\xi x)dx$$
for $\xi \in \mathbf{R}$;
- If $f : \mathbf{Z} \to \mathbf{C}$,
$$\hat{f}(\theta) = \sum_{n=-\infty}^{\infty} f(n)e(-n\theta);$$
- If $f : \mathbf{Z}/q\mathbf{Z} \to \mathbf{C}$,
$$\hat{f}(r) = \frac{1}{q}\sum_{x \in \mathbf{Z}/q\mathbf{Z}} f(x)e(-rx/q).$$

We now turn to some further properties of the Fourier transform in each case.

1.1. Fourier transform on R

The Fourier transform on $\mathbf{R}$ is the most difficult to study from an analytic point of view. We will only need it once in this course (in Section 6.6). There, we will need two facts. First, the the Fourier transform converts convolution to multiplication, that is to say if $f, g : \mathbf{R} \to \mathbf{C}$ are integrable and we define

$$(f * g)(x) := \int f(y)g(x-y)dy$$

then

$$\widehat{f * g} = \hat{f}\hat{g}.$$

It is almost trivial to check this formally:

$$\begin{aligned}\widehat{f * g}(\xi) &= \int\int f(y)g(x-y)e(-\xi x)dxdy \\ &= \int f(y)e(-\xi y)dy \int g(x-y)e(-\xi(x-y))dx \\ &= \hat{f}(\xi)\hat{g}(\xi).\end{aligned}$$

It can be rigorously justified using Fubini's theorem.

Second, the inversion formula, which lies rather deeper: under suitable conditions,

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi) e(\xi x) d\xi.$$

This holds if both f and $\hat{f}$ lie in $L^1(\mathbf{R}) \cap L^2(\mathbf{R})$. This may be proven by first establishing the result for Schwartz functions (see my lecture notes [**2**]) and then using an approximation argument (not given in those lecture notes, but see **??**). We will not go over the details in this course.

1.2. Fourier transform on Z

This plays an absolutely key role throughout the course. All of the functions we will be dealing with are compactly supported (that is, $f(n) = 0$ outside of some finite interval) and so the Fourier transform $\hat{f}(\theta)$ may always be defined, with no issues about convergence.

Here are the basic properties of the Fourier transform.

Proposition 1.2.1. *In the following proposition, $f, g : \mathbf{Z} \to \mathbf{C}$ are two compactly supported functions.*

(i) *We have the inversion formula*

$$f(n) = \int_{\mathbf{T}} \hat{f}(\theta) e(n\theta) d\theta.$$

(ii) *We have the Parseval identity*

$$\sum_n f(n)\overline{g(n)} = \int_{\mathbf{T}} \hat{f}(\theta)\overline{\hat{g}(\theta)} d\theta.$$

(iii) *If the convolution $f * g : \mathbf{Z} \to \mathbf{C}$ is defined by*

$$(f * g)(n) := \sum_m f(m) g(n-m)$$

then $\widehat{f * g}(\theta) = \hat{f}(\theta)\hat{g}(\theta)$.

Proof. All of this is an easy check using the definitions, as well as the fact that

$$\int_{\mathbf{T}} e(m\theta) d\theta = \int_0^1 e(m\theta) d\theta = \begin{cases} 1 & m = 0 \\ 0 & m \in \mathbf{Z} \setminus \{0\}. \end{cases} \tag{1.4}$$

(for (i) and (ii)) and the fact that $e(x+y) = e(x)e(y)$ (for (iii)). □

Remark. Taking $f = g$ in the Parseval identity gives

$$\sum_n |f(n)|^2 = \int_{\mathbf{T}} |\hat{f}(\theta)|^2.$$

1.3. Fourier transform on $\mathbf{Z}/q\mathbf{Z}$

We will use this in a number of places. Here are the basic properties, which parallel those in Proposition 1.2.1 rather closely.

PROPOSITION 1.3.1. *In the following proposition, $f, g : \mathbf{Z}/q\mathbf{Z} \rightarrow \mathbf{C}$ are two compactly supported functions.*

(i) *We have the inversion formula*

$$f(x) = \sum_{r \in \mathbf{Z}/q\mathbf{Z}} \hat{f}(r) e(rx/q).$$

(ii) *We have the Parseval identity*

$$\frac{1}{q} \sum_{x \in \mathbf{Z}/q\mathbf{Z}} f(x)\overline{g(x)} = \sum_{r \in \mathbf{Z}/q\mathbf{Z}} \hat{f}(r)\overline{\hat{g}(r)}.$$

(iii) *If the convolution $f * g : \mathbf{Z}/q\mathbf{Z} \rightarrow \mathbf{C}$ is defined by*

$$(f * g)(x) := \frac{1}{q} \sum_{y \in \mathbf{Z}/q\mathbf{Z}} f(y) g(x - y)$$

then $\widehat{f * g}(r) = \hat{f}(r)\hat{g}(r)$.

Proof. Once again, all of this is an easy check using the definitions, as well as the fact that

$$\sum_r e(rx/q) = \begin{cases} q & x = 0 \\ 0 & x \in (\mathbf{Z}/q\mathbf{Z}) \setminus \{0\}. \end{cases}$$

□

Remark. Taking $f = g$ in the Parseval identity gives

$$\frac{1}{q} \sum_{x \in \mathbf{Z}/q\mathbf{Z}} |f(x)|^2 = \sum_{r \in \mathbf{Z}/q\mathbf{Z}} |\hat{f}(r)|^2.$$

Remark. There is an important observation to be made here, which is that the appropriate measures to take on $\mathbf{Z}/q\mathbf{Z}$ and on the dual $\widehat{\mathbf{Z}/q\mathbf{Z}}$ are different. On the former group (when integrating over the spatial variables x, y) we took the normalised counting measure, whereas on the latter group (integrating over the dual variable r) we took the *un*normalised counting measure. There is a general theory of dual pairs of measures on dual pairs of groups $G, \hat{G}$, but we will not go into it here.

1.4. Convolution and additive number theory

It is worth pausing to comment on the operation of convolution. Suppose, for this discussion, that we are working in $\mathbf{Z}$. Suppose that $f = 1_A$ and that $g = 1_B$, where $A, B \subset \mathbf{Z}$ are finite sets. (That is, $f(n) = 1$ for $n \in A$ and 0 if

$n \notin A$.) Then the convolution $f * g$ is supported (nonzero) precisely on the set $A + B = \{a + b : a \in A, b \in B\}$. For this reason, and as a consequence of the nice behaviour of the Fourier transform with respect to convolution, convolution is a very important operation in additive number theory. We caution that $1_A * 1_B \neq 1_{A+B}$; in fact, $1_A * 1_B(n)$ is the number of representations of n as $a + b$ with $a \in A, b \in B$.

Part 1

Additive Number Theory

CHAPTER 2

Sums of squares

In this chapter, a *square* will mean the square of an integer, which may be zero. Thus the set of squares is $\{0, 1, 4, 9, 16, \dots\}$.

2.1. Sums of two squares

THEOREM 2.1.1. *An odd prime p is the sum of two squares if and only if $p \equiv 1 (\mathrm{mod}\, 4)$.*

Proof. Since all squares are 0 or $1 (\mathrm{mod}\, 4)$, a sum of two squares can only ever be $0, 1$ or 2 modulo 4, and in particular not $3 (\mathrm{mod}\, 4)$.

Conversely, suppose $p \equiv 1 (\mathrm{mod}\, 4)$ is a prime. By basic facts about quadratic residues, -1 is a square modulo p, and so there exist integers x, y (in fact $y = 1$) with $x^2 + y^2 = mp$ for some positive integer m. Suppose that m is the minimal positive integer with this property, and assume as a hypothesis for contradiction that $m > 1$. Replacing x with $-x$ and reducing mod p if necessary, and similarly for y, we may assume that $|x|, |y| < p/2$, and therefore

$$mp < 2(\frac{p}{2})^2,$$

which certainly implies that $m < p$. In particular, at least one of x and y is not divisible by m. Indeed, if not then $m^2 | x^2 + y^2$, implying that $m | p$. Since we are assuming that $m \neq 1$, this would force $m = p$, which we know not to be the case.

Pick a, b with $|a|, |b| \leqslant m/2$ and $x \equiv a (\mathrm{mod}\, m)$, $y \equiv b (\mathrm{mod}\, m)$. Note that $a^2 + b^2 > 0$ since not both of x, y are multiples of m. Note furthermore that

$$a^2 + b^2 \equiv x^2 + y^2 \equiv 0 (\mathrm{mod}\, m);$$

let us write $a^2 + b^2 = rm$, where $r > 0$ is an integer. Note that

$$rm \leqslant 2(\frac{m}{2})^2,$$

and so $r < m$. Observing the identity

$$(x^2 + y^2)(a^2 + b^2) = (xa + yb)^2 + (xb - ya)^2,$$

we have

$$rm^2 p = (xa + yb)^2 + (xb - ya)^2.$$

Now we have

$$xa + yb \equiv x^2 + y^2 \equiv 0(\operatorname{mod} m),$$

and

$$xb - ya \equiv xy - yx \equiv 0(\operatorname{mod} m).$$

Therefore if we define

$$x' := \frac{xa + yb}{m}, y' := \frac{xb - ya}{m},$$

both x' and y' are integers. Furthermore

$$(x')^2 + (y')^2 = rp.$$

Since $0 < r < m$, this is contrary to the supposed minimality of m. Therefore we were wrong to assume that $m > 1$, and the proof is complete. □

**Remarks.* The "descent" argument we gave for Theorem 2.1.1 is one of the most elementary proofs of the theorem. A somewhat different proof goes via algebraic number theory in $\mathbf{Q}(i)$. It is known that the ring of integers $\mathbf{Z}[i]$ is a principal ideal domain (PID), and so if the ideal (p) splits then it must be as a product $(p) = (x + iy)(x - iy)$ of two principal ideals, both of norm p, which then implies that $x^2 + y^2 = p$. But there is a criterion (Dedekind's criterion) asserting that the factorisation of (p) in $\mathbf{Z}[i]$ can be read off from the factorisation of the polynomial $X^2 + 1$ in $\mathbf{F}_p$. In particular, (p) splits if and only if $X^2 + 1$ factors over $\mathbf{F}_p$, that is to say precisely when -1 is a quadratic residue mod p, i.e. $p \equiv 1(\operatorname{mod} 4)$.

Note that, although the above argument is short modulo known results, the usual proof that $\mathbf{Z}[i]$ is a PID proceeds via showing that it is a Euclidean domain, that is to say by a descent procedure quite similar to that used in the proof of Theorem 2.1.1.

2.2. Sums of four squares

THEOREM 2.2.1. *Every natural number is the sum of four squares.*

Proof. We note the identity

$$\begin{aligned}(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) &= (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2 + \\ &+ (x_1y_2 - x_2y_1 + x_3y_4 - x_4y_3)^2 + (x_1y_3 - x_3y_1 + x_4y_2 - x_2y_4)^2 \\ &+ (x_1y_4 - x_4y_1 + x_2y_3 - x_3y_2)^2. \qquad (2.1)\end{aligned}$$

This means that the set of numbers which are the sum of four squares is closed under multiplication. Since $2 = 1^1 + 1^2 + 0^2 + 0^2$, it suffices to show that any odd prime p is in this set.

Now we proceed along very similar lines to the proof of Theorem 2.1.1. First, we claim that there is some $m > 0$ such that

$$mp = x_1^2 + x_2^2 + x_3^2 + x_4^2.$$

To see this, first observe that every element x of $\mathbf{Z}/p\mathbf{Z}$ is a sum of two squares, since the set S of squares $(\mathrm{mod}\, p)$ has size $\frac{1}{2}(p+1)$, and hence S and $x - S$ must intersect. Writing 1 and $-1(\mathrm{mod}\, p)$ as sums of two squares and then adding gives a sum of four squares, not all zero, which is a multiple of p.

Assume that m is minimal with this property, and suppose as a hypothesis for contradiction that $m > 1$.

Replacing x_i with $-x_i$ if necessary, we may assume that $|x_i| < p/2$ (note that p is odd, so the inequality is indeed strict). It follows that

$$mp < 4(\frac{p}{2})^2 = p^2,$$

and so $0 < m < p$.

If m is even, then the x_i may be grouped into two pairs in which the parities are equal, say $x_1 \equiv x_2(\mathrm{mod}\, 2)$, $x_3 \equiv x_4(\mathrm{mod}\, 2)$. But then

$$\frac{1}{2}mp = (\frac{x_1 + x_2}{2})^2 + (\frac{x_1 - x_2}{2})^2 + (\frac{x_3 + x_4}{2})^2 + (\frac{x_3 - x_4}{2})^2,$$

contrary to the minimality of m.

Suppose, then, that m is odd. Not all of the x_i are divisible by m, as this would imply $m^2 | x_1^2 + x_2^2 + x_3^2 + x_4^2 = mp$ and so $m|p$. Since we are assuming $m > 1$, this forces $m = p$, but we have already proved that $m < p$. Pick y_i with $|y_i| < m/2$ and $x_i \equiv y_i(\mathrm{mod}\, m)$, $i = 1, \dots, 4$. This is possible with *strict* inequality, as claimed, since m is odd. Then $y_1^2 + y_2^2 + y_3^2 + y_4^2$ is positive, and also a multiple of m since it is congruent to $x_1^2 + x_2^2 + x_3^2 + x_4^2$. Suppose that $y_1^2 + y_2^2 + y_3^2 + y_4^2 = rm$. Then

$$rm < 4(\frac{m}{2})^2 = m^2$$

and so $r < m$. Now from (2.1), we have

$$rm^2p = (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2 + \dots,$$

where the ... comprises the three other terms in (2.1). One may easily check, using $x_i \equiv y_i(\mathrm{mod}\, m)$, that all four of the bracketed terms are multiples of m. Therefore

$$rp = (\frac{x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4}{m})^2 + \dots,$$

with all the bracketed terms integers. Since $r < m$, this contradicts the supposed minimality of m. □

2.3. Sums of three squares

Theorems about sums of three squares lie a little deeper, at least partly because there is no analogue of the multiplicativity identity (2.1). However, any student of number theory should certainly be aware of the main result on the topic, due to Gauss.

Theorem 2.3.1. *All numbers not of the form* $4^m(8k+7)$ *are the sum of three squares.*

2.4. *Further comments

Sums of squares have a rich theory. Sums of three squares are connected to class numbers. Writing $h(d)$ for the class number of the quadratic field $\mathbf{Q}(\sqrt{d})$, Gauss showed that the number of representations of $d > 3$ as a sum of 3 squares is $ch(d)$, where $c = 12$ if $d \equiv 1, 2 (\bmod 4)$, $c = 24$ if $d \equiv 3 (\bmod 8)$, and $c = 0$ if $d \equiv 7 (\bmod 8)$.

Representations by sums of squares are intimately connected to the theory of modular forms of half integral weight. This leads to beautiful results: for example, the number of ways to write n as a sum of four squares is 8 times the sum of the divisors d of n with $4 \nmid d$.

A good source for more information is [**3**, Section 11.3], where one may find explicit formulae for the number of representations of n as a sum of s squares, $s = 4, 6, 8, 10, 12$. The formula for $s = 8$ is particularly clean, the number of representations in this case being $16 \sum_{d|n} (-1)^{n-d} d^3$.

CHAPTER 3

Waring's problem: an introduction

3.1. Statement of main results

Let $k \geqslant 2$ be an integer. We define $G(k)$ to be the smallest positive integer s such that all except finitely many positive integers may be written as $x_1^k + \cdots + x_s^k$, where $x_1, \dots, x_s$ are non-negative integers. In the preceding chapter, we showed that $G(2) = 4$: indeed *every* positive integer is the sum of four squares, but there are infinitely many n which are not the sum of three squares.

The main result of the first half of the course is the following.

THEOREM 3.1.1. *$G(k)$ is finite, and in fact grows at most exponentially in k.*

We will in fact prove a result which is much more precise than Theorem 3.1.1, obtaining an asymptotic formula or "local–global principle" for the number of representations of N as $x_1^k + \cdots + x_s^k$ (that is, the number of tuples $(x_1, \dots, x_s) \in \mathbf{N}^s$ with $x_1^k + \cdots + x_s^k = N$).

THEOREM 3.1.2. *Let $r_{k,s}(N)$ be the number of representations of N as a sum $x_1^k + \cdots + x_s^k$ with $x_i \in \mathbf{N}$. Suppose that $s > 100^k$. Then*

$$r_{k,s}(N) = \mathfrak{S}_{k,s}(N) N^{s/k-1} + o(N^{s/k-1}). \tag{3.1}$$

Here

$$\mathfrak{S}_{k,s}(N) = \beta_\infty \prod_p \beta_p(N) \tag{3.2}$$

where $\beta_p(N)$ is the p-adic density of solutions *defined by*

$$\beta_p(N) := \lim_{n \to \infty} p^{-n(s-1)} \#\{(x_1, \cdots x_s) \in (\mathbf{Z}/p^n\mathbf{Z})^s : x_1^k + \cdots + x_s^k \equiv N (\operatorname{mod} p^n)\} \tag{3.3}$$

and

$$\beta_\infty := \Gamma(1 + 1/k)^s / \Gamma(s/k).$$

A number of remarks are in order.

1. Included in the statements is the fact that the limit in the definition of the p-adic density (3.3) exists. This is not immediately obvious.

2. The theorem states that $r_{k,s}(N)$ is of order $N^{s/k-1}$, but with a constant $\mathfrak{S}_{k,s}(N)$, usually known as the "singular series", which depends on how easy it is to represent N as a sum of s kth powers "locally", both modulo prime powers (that is, p-adically) and in the reals.

3. One can basically see from the analysis in Section 6.6 that β_∞ is the surface area of the set $\{(x_1, \dots, x_s) : x_1^k + \cdots + x_s^k = 1\}$, and so this term does indeed represent a real-variable or "archimedean" density of solutions.

4. The constant 100 appearing in the statement of Theorem 3.1.2 is certainly not the best one that our method gives. Moreover, with the additional tool of Hua's Lemma (Chapter 8: this may or may not get lectured) one can show that $s \geqslant 2^k + 1$ is enough.

5. Here, and in later chapters, we will not explicitly indicate the fact that error terms such as $o(N^{s/k-1})$ depend on s, k (which we think of as fixed). The same goes for the constants in the $O()$ and $\asymp$ notations.

Theorem 3.1.2 is not of great utility by itself, as we have said nothing about $\mathfrak{S}_{k,s}(N)$. In particular, by itself it does not imply Theorem 3.1.1. In Chapter 7, we will complement it with a proof of the following.

Proposition 3.1.1. *For $s \geqslant k^4$ we have*[1] $\mathfrak{S}_{k,s}(N) \asymp 1$.

On Sheet 2 we will show that the same conclusion holds if $s \geqslant 5$ (when $k = 2$) and if $s \geqslant 9$ (when $k = 3$) and thus when $s \geqslant 2^k + 1$ in all cases.

Proposition 3.1.1, together with Theorem 3.1.2, does immediately show that $G(k)$ is finite, and in fact bounded by 100^k (or, in fact, $2^k + 1$ if one additionally uses Chapter 8 and the calculations on Sheet 2).

3.2. The Hardy–Littlewood method

The first key idea in the proof of Theorem 3.1.2 is to express $r_{k,s}(N)$ using a Fourier transform. This is natural on account of the fact that $r_{k,s}(N)$ is a convolution: in fact, it is the s-fold convolution $1_X * \cdots * 1_X(N)$, where $X = \{n^k : n \leqslant N^{1/k}\}$ is the set of kth powers less than or equal to N. Since the Fourier transform of $1_X * \cdots * 1_X$ is $\hat{1}_X(\theta)^s$, it follows from the inversion formula that

$$r_{k,s}(N) = \int_{\mathbf{T}} \hat{1}_X(\theta)^s e(N\theta) d\theta. \tag{3.4}$$

Since this formula is so fundamental to us, let us write down the proof explicitly, without mentioning convolution and inversion: substituting the definition of $\hat{1}_X(\theta)$, that is to say

$$\hat{1}_X(\theta) = \sum_{n \leqslant N^{1/k}} e(-n^k\theta), \tag{3.5}$$

[1] Recall from the introduction what this means: there are $c_1, c_2 > 0$ such that $c_1 \leqslant \mathfrak{S}_{k,s}(N) \leqslant c_2$; these constants may (and will) depend on k, s.

on the right hand side, we get

$$\int_{\mathbf{T}} \Big(\sum_{n \leqslant N^{1/k}} e(-n^k\theta)\Big)^s e(N\theta)d\theta = \int_{\mathbf{T}} \sum_{n_1,\dots,n_s \leqslant N^{1/k}} e((N - n_1^k - \cdots - n_s^k)\theta)d\theta.$$

Swap the summation over the n_i and the integral over $\mathbf{T}$. Using the orthogonality relation (1.4), we see that the inner integral vanishes unless $N = n_1^k + \cdots + n_s^k$, in which case it equals 1. Thus we do indeed get precisely $r_{k,s}(N)$.

We must now estimate the integral in (3.4), and to this end we must study the Fourier transform (3.5). (This is more usually known in the literature as an *exponential sum*, and we may use that term too, but it *is* a kind of Fourier transform.) The next key observation is that $\hat{1}_X(\theta)$ exhibits two very different types of behaviour, as follows.

- If θ is, or is close to, a rational with small denominator then we can find an asymptotic formula for $\hat{1}_X(\theta)$. For example, suppose that $k = 2$ and consider $\hat{1}_X(\frac{1}{3}) = \sum_{n \leqslant N^{1/2}} e(-n^2/3)$. Noting that $e(-n^2/3) = 1$ if $n \equiv 0 (\operatorname{mod} 3)$ and $e(-1/3)$ if $n \equiv 1, 2 (\operatorname{mod} 3)$, we see that $\hat{1}_X(\frac{1}{3})$ is almost exactly $\frac{N^{1/2}}{3}(1 + 2e(-1/3))$.

  More generally if $\theta = a/q$ the sum $\sum_{n \leqslant N^{1/k}} e(-\theta n^k)$ is periodic with period q, and thus is roughly equal to $\frac{1}{q} N^{1/k} \sum_{n=0}^{q-1} e(-an^k/q)$. Note that this is inclined to be somewhat large: if q is small then we expect it to be comparable to $N^{1/k}$, which is of course the trivial upper bound for $\hat{1}_X(\theta)$.
- If θ is highly irrational then we expect the terms $e(-\theta n^k)$, $n = 1, 2, \dots$ to be somewhat randomly distributed around the unit circle in the complex plane, in which case we expect $\hat{1}_X(\theta)$ to be $o(N^{1/k})$ due to cancellation. Note that we are *not* claiming to have actually proven this.

The two cases are called the *major* and *minor* arcs respectively. Here is a formal definition.

DEFINITION 3.2.1 (Major and minor arcs). Set $\eta := 1/10k$. Define the major arcs $\mathfrak{M}$ to be the union of all the sets $\mathfrak{M}_{a,q}$, over all $q \leqslant N^\eta$ and $a \in (\mathbf{Z}/q\mathbf{Z})^*$, where

$$\mathfrak{M}_{a,q} := \{\theta \in \mathbf{T} : \|\theta - \frac{a}{q}\|_{\mathbf{T}} \leqslant N^{-1+2\eta}\}.$$

Define the minor arcs $\mathfrak{m}$ to be $\mathbf{T} \setminus \mathfrak{M}$.

Remark. There is considerable flexibility in the definition, and we have simply chosen a convenient one. Roughly, the major arc $\mathfrak{M}_{a,q}$ is always something like the set of points where "there is some $q \lessapprox 1$ such that $\|\theta - a/q\| \lessapprox 1/N$". The $\approx$ notation hides the factor N^η (resp. $N^{2\eta}$). Making this smaller simplifies the analysis of the major arcs, but puts more pressure on the analysis of the minor arcs.

We can now divide the task of proving Theorem 3.1.2 into two subtasks, as follows. This division is quite reasonable in view of the informal discussion above.

PROPOSITION 3.2.1 (Major arcs). *Suppose that $s \geqslant 2k+1$. Then*

$$\int_{\mathfrak{M}} \hat{1}_X(\theta)^s e(N\theta)d\theta = \mathfrak{S}_{k,s}(N)N^{s/k-1} + o(N^{s/k-1}). \tag{3.6}$$

PROPOSITION 3.2.2 (Minor arcs). *Suppose that $s > 100^k$. Then*

$$\int_{\mathfrak{m}} \hat{1}_X(\theta)^s e(N\theta)d\theta = o(N^{s/k-1}). \tag{3.7}$$

Remark. Note that one can establish the major arc estimate under much weaker conditions than the minor arc estimate. Improving the minor arc estimate is therefore the main obstacle to obtaining stronger bounds in Waring's problem via the Hardy-Littlewood method.

3.3. *Further comments

In this course, we are giving more-or-less the simplest possible proof of a bound for $G(k)$ using the Hardy-Littlewood method. As previously remarked, the constant 100 can certainly be improved. However, to get a bound better than exponential in k, substantial new ideas are needed.

Asymptotically, the best bound currently known is due to Wooley [**4**], who proved that

$$G(k) \leqslant (1+o(1))k\log k.$$

Improvements to the lower order terms $o(1)$ have been made, as have refinements to the values for small k. However, the only value of $G(k)$ known, other than $G(2)$, is $G(4) = 16$. In particular it is not known whether $G(3)$ is $4, 5, 6$ or 7.

Write[2] $G_{\text{cong}}(k)$ for the least $s \geqslant k+1$ for which there are no congruence obstructions to every large number being the sum of s kth powers. It is natural to conjecture that $G(k) = G_{\text{cong}}(k)$. The quantity $G_{\text{cong}}(k)$ was studied by Hardy and Littlewood. For references, and for a table of values for small k, see [**1**, Chapter 5]. In particular $G_{\text{cong}}(3) = 4$.

The best-known bounds for $G(k)$ do not provide asymptotics for the number of solutions. Asymptotics such as (3.1.2) are known for $s > Ck^2$.

[2] In the literature this is often called $\Gamma(k)$, despite the potential for confusion with the Γ-function.

CHAPTER 4

The minor arcs

Our aim in this chapter is to establish (3.7), the minor arcs estimate. First we note that it is a simple consequence of the following pointwise estimate.

Proposition 4.0.1 (Pointwise estimate). *Set* $\varepsilon := (100)^{-k}$. *We have*

$$\sup_{\theta \in \mathfrak{m}} |\hat{1}_X(\theta)| \ll N^{1/k-\varepsilon}.$$

Indeed, it is obvious that (3.7) then holds for any $s > 100^k$. The main task of this chapter, then, is to establish Proposition 4.0.1.

4.1. Diophantine approximation

It is well-known that if $\alpha \in \mathbf{T}$ is "highly irrational", for example if $\alpha = \sqrt{2}$, then the sequence $(\alpha n)_{n \in \mathbf{N}}$ is very uniformly distributed in $\mathbf{T}$. In this section we prove a result which asserts a kind of converse to this: if the sequence (αn) is *not* close to equidistributed then α is "major arc". Lemmas of this type in this context are normally attributed to Vinogradov.

Here is the statement of the result we shall prove.

Lemma 4.1.1. *The is an absolute constant C with the following property. Suppose that $\alpha \in \mathbf{R}$ and that I is an interval of integers with $\#I = N$. Suppose that δ_1, δ_2 are positive quantities satisfying $\delta_2 > C\delta_1$, and suppose that there are at least $\delta_2 N$ elements $n \in I$ for which $\|\alpha n\|_{\mathbf{T}} \leqslant \delta_1$. Suppose that $N \geqslant C/\delta_2$. Then there is some $q \leqslant C/\delta_2$ such that $\|\alpha q\|_{\mathbf{T}} \leqslant C\delta_1/\delta_2 N$.*

Remark. The proof gives a reasonable value of C such as $C = 128$, as the reader may care to check.

The proof of Lemma 4.1.1 is somewhat fiddly. Let us begin with a very well-known lemma of Dirichlet.

Lemma 4.1.2 (Dirichlet). *Suppose that $Q \in \mathbf{N}$. Let $\alpha \in \mathbf{R}$. Then there is some nonzero $q \leqslant Q$ such that $\|\alpha q\|_{\mathbf{T}} \leqslant 1/Q$.*

Proof. Consider the numbers $\alpha, 2\alpha, \dots, Q\alpha$ as elements of $\mathbf{T}$. By the pigeonhole principle, some two of them, say $i\alpha$ and $j\alpha$ with $i < j$, must satisfy $\|j\alpha - i\alpha\|_{\mathbf{T}} \leqslant \frac{1}{Q}$. Now take $q := j - i$. □

Now we turn to the proof of Lemma 4.1.1.

Proof. [Proof of Lemma 4.1.1]

Step 1. We begin with a simple reduction to the case $I = \{1, \dots, N\}$. Suppose we know that case. In the general case, let n_0 be the smallest element of I with $\|\alpha n_0\|_{\mathbf{T}} \leqslant \delta_1$. There are at least $\delta_2 N - 1 \geqslant \delta_2 N/2$ other (larger) values of $n \in I$ for which $\|\alpha n\|_{\mathbf{T}} \leqslant \delta_1$. For each of them, writing $m := n - n_0$ we have $m \in \{1, \dots, N\}$ and $\|\alpha m\|_{\mathbf{T}} \leqslant \|\alpha n_0\|_{\mathbf{T}} + \|\alpha n\|_{\mathbf{T}} \leqslant 2\delta_1$. Applying the special case $I = \{1, \dots, N\}$ of the lemma (with $\delta' = 2\delta_1$, $\delta_2' = \delta_2/2$) we get the general case, albeit with a worse constant $\tilde{C} = 4C$.

Suppose henceforth that $I = \{1, \dots, N\}$. Write $S \subseteq \{1, \dots, N\}$ for the set of all n such that $\|\alpha n\|_{\mathbf{T}} \leqslant \delta_1 N$; thus $|S| \geqslant \delta_2 N$.

Step 2. We simply apply Dirichlet's lemma with $Q = 4N$. (Taking $Q = 4N$ rather than $Q = N$ is a useful technical convenience later on). We obtain that there is a nonzero $q \leqslant 4N$ such that $\|\alpha q\|_{\mathbf{T}} \leqslant 1/4N$. This conclusion is weaker in both aspects (the bound for q, and the bound for $\|\alpha q\|_{\mathbf{T}}$) than the bound we are aiming for; this is hardly surprising, since we have not used the assumptions of the lemma.

The bound $\|\alpha q\|_{\mathbf{T}} \leqslant 1/4N$ implies that there is some a such that $|\alpha - \frac{a}{q}| \leqslant \frac{1}{4Nq}$. Without loss of generality (decreasing q if necessary) we can assume $(a, q) = 1$. Write $\theta := \alpha - \frac{a}{q}$; thus

$$|\theta| \leqslant \frac{1}{4Nq}. \tag{4.1}$$

The remaining steps consist of "bootstrapping" the rather trivial conclusion of step 2. First, we tighten the bound for q, and then the bound for $|\theta|$.

Step 3. Suppose that $n \in S$. Then, by (4.1), we see that

$$\|\frac{an}{q}\|_{\mathbf{T}} \leqslant \delta_1 + \frac{1}{4q}. \tag{4.2}$$

Now we bound the number of $n \in \{1, \dots, N\}$ satisfying (4.2) in a different way. Divide $\{1, \dots, N\}$ into $\leqslant \frac{N}{q} + 1$ intervals of length q. In each interval, $\frac{an}{q} (\mathrm{mod}\, 1)$ ranges over each rational with denominator q precisely once. At most

$$2q(\delta_1 + \frac{1}{4q}) + 1 < 2(\delta_1 q + 2)$$

of these lie in the interval $\|x\|_{\mathbf{T}} \leqslant \delta_1 + \frac{1}{4q}$. Thus the total number of $n \in \{1, \dots, N\}$ satisfying (4.2) is bounded above by

$$2(\frac{N}{q} + 1)(\delta_1 q + 2) = 2\delta_1 N + 2\delta_1 q + \frac{4N}{q} + 4.$$

It follows that

$$2\delta_1 N + 2\delta_1 q + \frac{4N}{q} + 4 \geqslant \delta_2 N. \tag{4.3}$$

Now

- $2\delta_1 N < \delta_2 N/4$, since we are assuming $\delta_2 > C\delta_1$;
- $2\delta_1 q \leqslant 8\delta_1 N < \delta_2 N/4$, since $q \leqslant 4N$, and we are assuming $\delta_2 > C\delta_1$;
- $4 < \delta_2 N/4$, since we are assuming $N > C/\delta_2$.

Therefore (4.3) forces us to conclude that $4N/q > \delta_2 N/4$, and therefore $q \leqslant 16/\delta_2$. We have succeeded in bootstrapping to a bound on q of the required strength.

Step 4. Recall (4.2), which says that if $n \in S$ then

$$\|\frac{an}{q}\|_{\mathbf{T}} \leqslant \delta_1 + \frac{1}{4q}.$$

However, in the light of Step 3, we have

$$\delta_1 \leqslant \frac{\delta_2}{C} \leqslant \frac{16}{qC} < \frac{1}{2q},$$

and so if $n \in S$ then

$$\|\frac{an}{q}\|_{\mathbf{T}} < \frac{1}{q}.$$

Therefore $\|\frac{an}{q}\|_{\mathbf{T}} = 0$, and every element of S is a multiple of q.

Step 5. It follows from Step 4 that if $n \in S$ then

$$\|\theta n\|_{\mathbf{T}} = \|\alpha n\|_{\mathbf{T}} \leqslant \delta_1.$$

However, since $|\theta| \leqslant 1/4Nq$, for $n \leqslant N$ we have

$$\|\theta n\|_{\mathbf{T}} = |\theta n|.$$

Therefore

$$|\theta n| \leqslant \delta_1 \tag{4.4}$$

for all $n \in S$. Finally, recall that S consists of multiples of q and that $|S| \geqslant \delta_2 N$; therefore there is some $n \in S$ with $|n| \geqslant \delta_2 q N$. Using this n, (4.4) implies that

$$|\theta| \leqslant \frac{\delta_1}{q\delta_2 N},$$

and so finally

$$\|\alpha q\|_{\mathbf{T}} \leqslant |\theta q| \leqslant \frac{\delta_1}{\delta_2 N}.$$

This concludes the proof. □

4.2. Bounds for Weyl sums

In this section we give a bound for exponential sums with polynomial phases, known as *Weyl sums*. The point is that if such a sum is large, then the lead coefficient of the polynomial is “highly rational”. The result is very closely related to Weyl's inequality, the statement of which youwill easily find in the literature. For the purposes of this course I am formulating it slightly differently.

THEOREM 4.2.1. *Set $C_k := 10^k$. Let δ be sufficiently small in terms of k, and suppose that $L > \delta^{-C_k}$. Let $I \subseteq \mathbf{Z}$ be an interval of length at most L. Let $P : \mathbf{Z} \to \mathbf{R}$, $P(x) = \alpha x^k + \cdots$ be a polynomial of degree k. Suppose that $|\sum_{x\in I} e(P(x))| \geqslant \delta L$. Then there is $q \leqslant \delta^{-C_k}$ such that $\|q\alpha\|_{\mathbf{T}} \leqslant \delta^{-C_k} L^{-k}$.*

To make the proof of this more digestible, we will prove the cases $k = 1$ (which is elementary) and $k = 2$ (which has almost all of the ideas of the general case) separately and carefully.

Before commencing either task we isolate a simple general lemma, called an "averaging lemma", from the proof.

LEMMA 4.2.1. *Let X be a finite set and suppose that $b : X \to \mathbf{C}$ is a function such that $|b(x)| \leqslant 1$ for all $x \in X$. Suppose that $|\sum_{x\in X} b(x)| \geqslant \varepsilon|X|$. Then there are at least $\varepsilon|X|/2$ values of $x \in X$ such that $|b(x)| \geqslant \varepsilon|X|/2$.*

Proof. Suppose not. Then

$$|\sum_{x\in X} b(x)| \leqslant \sum_{x\in X} |b(x)| \leqslant \frac{\varepsilon}{2}|X| + (1-\frac{\varepsilon}{2})|X|\frac{\varepsilon}{2} < \varepsilon|X|,$$

contrary to assumption. □

The case $k = 1$. In this case we have the following version of Proposition 4.2.1. Note in particular that there is no q in this case.

PROPOSITION 4.2.1. *Let $\delta > 0$ and $L > 0$. Let $I \subseteq \mathbf{Z}$ be an interval of length at most L. Suppose that $P(x) = \alpha x + \dots$ be a linear polynomial, and suppose that $|\sum_{x\in I} e(P(x))| \geqslant \delta L$. Then $\|\alpha\|_{\mathbf{T}} \leqslant \delta^{-1}L^{-1}$.*

Proof. By summing the geometric series, we have

$$|\sum_{x\in I} e(P(x))| = |\sum_{j=0}^{|I|-1} e(\alpha x)| = |\frac{1-e(|I|\alpha)}{|1-e(\alpha)|} \leqslant \frac{2}{|1-e(\alpha)|}.$$

Now

$$|1-e(\alpha)| = 2|\sin \pi\alpha| \geqslant 4\|\alpha\|_{\mathbf{T}},$$

where in this last step we used the inequality $|\sin t| \geqslant \frac{2}{\pi}|t|$, which is valid for $|t| \leqslant \pi/2$.

Putting these facts together gives

$$|\sum_{x\in I} e(P(x))| \leqslant \frac{1}{\|\alpha\|_{\mathbf{T}}}$$

(in fact one could have a 2 in the denominator if one wanted).

If the left-hand side is $\geqslant \delta L$ then we see immediately that $\|\alpha\|_{\mathbf{T}} \leqslant \delta^{-1}L^{-1}$, concluding the proof in this case. □

The case $k = 2$. In this case we have the following result, which implies Theorem 4.2.1 with room to spare.

PROPOSITION 4.2.2. *Let $\delta > 0$, and suppose that $L > 2^{20}\delta^{-4}$. Let $I \subseteq \mathbf{Z}$ be an interval of length at most L. Let $P : \mathbf{Z} \to \mathbf{R}$, $P(x) = \alpha x^2 + \dots$ be a quadratic polynomial. Suppose that $|\sum_{x\in I} e(P(x))| \geqslant \delta L$. Then there is a nonzero $q \ll \delta^{-4}$ such that $\|q\alpha\|_{\mathbf{T}} \ll \delta^{-4}L^{-2}$.*

Proof. The first key idea comes immediately: we square the assumption $|\sum_{x\in I} e(P(x))| \geqslant \delta L$. This gives

$$\sum_{x,y\in I} e(P(y) - P(x)) \geqslant \delta^2 L^2.$$

Now make the change of variables $h := y - x$; this yields

$$\sum_{|h|\leqslant L} \sum_{x\in I_h} e(\partial_h P(x)) \geqslant \delta^2 L^2, \tag{4.5}$$

where $\partial_h P(x) := P(x + h) - P(x)$ and $I_h := I \cap (I - h)$. To this, we apply the averaging principle, Lemma 4.2.1, taking in that lemma $X = \{h : |h| \leqslant L\}$, $\varepsilon = \delta^2/3$, and $b(h) := \frac{1}{L}\sum_{x\in I_h} e(\partial_h P(x))$. The conclusion is that there are $\geqslant \delta^2 L/6$ values of $|h| \leqslant L$ such that $|\sum_{x\in I_h} e(\partial_h P(x))| \geqslant \delta^2 L/6$. Since we are assuming $L > 2^{20}\delta^{-4}$, the contribution from $h = 0$ is negligible and we can assume without loss of generality that there are $\geqslant \delta^2 L/18$ values of $h \in \{1, \dots, L\}$ such that

$$|\sum_{x\in I_h} e(\partial_h P(x))| \geqslant \delta^2 L/6. \tag{4.6}$$

(Alternatively, there are many values of $h \in \{-L, \dots, -1\}$, but the proof is almost identical in this case.)

Now for the second key observation: the derivative $\partial_h P(x)$ is linear, with $\partial_h P(x) = 2\alpha h x + \dots$. Therefore we may apply the case $k = 1$ (that is, Proposition 4.2.1) to (4.6), concluding that if this holds then $\|2\alpha h\|_{\mathbf{T}} \leqslant 6/\delta^2 L$.

Thus, for at least $\delta^2 L/18$ values of $h \in \{1, \dots, L\}$, we have $\|2\alpha h\|_{\mathbf{T}} \leqslant 6/\delta^2 L$.

We are now in the situation covered by the Diophantine Lemma, Lemma **??**, with $\delta_1 := 6/\delta^2 L$ and $\delta^2 := \delta^2/18$. For the lemma to apply, we need $\delta_2 > 64\delta_1$ and $L \geqslant 16/\delta_2$; both of these are consequences of the assumption that $L > 2^{20}\delta^{-4}$. The conclusion of that lemma is then that there is a nonzero

$$q \leqslant \frac{16}{\delta_2} \ll \delta^{-2}$$

such that

$$\|2\alpha q\|_{\mathbf{T}} \leqslant \frac{8\delta_1}{\delta_2 L} \ll \delta^{-4}L^{-2}.$$

Replacing q by $2q$, the proof of Proposition 4.2.2 is complete. □

**The case of general k.* We now give the proof of Theorem 4.2.1. We proceed by induction on k, the result having been proven already for $k = 1, 2$. There are some bookkeeping issues to do with keeping track of exponents, the careful checking of which we leave to the reader: our stated value $C_k = 10^k$ leaves plenty of room. There is one small additional wrinkle beyond the case $k = 2$.

Proof. Exactly as in the case $k = 2$, we square the assumption and change variables, concluding that there are $\geqslant \delta^2 L/18$ values of $h \in \{1, \dots, L\}$ such that

$$|\sum_{x \in I_h} e(\partial_h P(x))| \geqslant \delta^2 L/6. \tag{4.7}$$

Write $H \subseteq \{1, \dots, L\}$ for the set of such h. The derivative $\partial_h P(x)$ is a polynomial of degree $k-1$ with leading coefficient $kh\alpha x^{k-1}$. It follows from the inductive hypothesis that , for each $h \in H$, there is some positive $q_h \ll \delta^{-2C_{k-1}}$ such that

$$\|khq_h\alpha\|_{\mathbf{T}} \ll \delta^{-2C_{k-1}} L^{-(k-1)}.$$

This is the additional wrinkle: q_h may depend on h, whereas in the argument for $k = 2$ it did not.

However, by the pigeonhole principle we may pass to a subset $H' \subset H$,

$$\#H' \gg \delta^{2+2C_{k-1}} L, \tag{4.8}$$

such that $q_{h'}$ does *not* depend on h' for $h' \in H'$. Write q' for this common value. Writing $\alpha' := kq'\alpha$, we have proved the following: there is a set $H' \subseteq \{1, \dots, L\}$,

$$|H'| \gg \delta^{2+2C_{k-1}} L, \tag{4.9}$$

such that if $h \in H'$ then

$$\|h\alpha\|_{\mathbf{T}} \ll \delta^{-2C_{k-1}} L^{-(k-1)}. \tag{4.10}$$

We are once again in the situation described by the Diophantine lemma, Lemma 4.1.1. In that lemma, take

$$\delta_1 = C\delta^{-2C_{k-1}} L^{-(k-1)}$$

and

$$\delta_2 = c\delta^{2+2C_{k-1}},$$

where c, C are the implied constants in (4.9), (4.10).

We leave it to the reader to check that (if δ is sufficiently small in terms of k) the assumptions of that lemma, namely that $L \geqslant 16/\delta_2$ and that $\delta_2 > 64\delta_1$, are satisfied.

The lemma states that there is a nonzero q'',

$$q'' \ll \delta_2^{-1} \ll \delta^{-2-2C_{k-1}}$$

such that

$$\|\alpha' q''\|_{\mathbf{T}} \ll \delta_1 \delta_2^{-1} L^{-1} \ll \delta^{-2-4C_{k-1}} L^{-k}.$$

Take $q := kq'q''$. Then

$$\|\alpha q\|_{\mathbf{T}} = \|\alpha' q''\|_{\mathbf{T}} \ll \delta^{-2-4C_{k-1}} L^{-k} \tag{4.11}$$

and

$$q \ll \delta^{-2-4C_{k-1}}. \tag{4.12}$$

The k case of Theorem 4.2.1 follows immediately from (4.11) and (4.12), noting that $C_k > 2 + 4C_{k-1}$ by a substantial margin. □

4.3. The pointwise estimate

We may now establish Proposition 4.0.1, the pointwise estimate for $\hat{1}_X(\theta)$ on the minor arcs.

Proof. [Proof of Proposition 4.0.1] As in Proposition 4.0.1, let $\varepsilon := (100)^{-k}$. Suppose that

$$|\hat{1}_X(\theta)| \geqslant N^{1/k-\varepsilon}. \tag{4.13}$$

We will show that $\theta \in \mathfrak{M}$ (the major arcs), which of course implies Proposition 4.0.1.

Set $\delta := N^{-\varepsilon}$. Then (4.13) is equivalent to

$$|\sum_{n \leqslant N^{1/k}} e(-\theta n^k)| \geqslant \delta N^{1/k}.$$

We now apply Proposition 4.2.1 with $P(x) = \theta x^k$ and $L = N^{1/k}$. The conclusion is that there is some $q \ll N^{\varepsilon 10^k}$ such that $\|q\theta\|_{\mathbf{T}} \ll N^{-1+\varepsilon 10^k}$. If N is large enough then, due to the choice of ε, this (by a very large margin) implies that θ does indeed lie in the major arcs $\mathfrak{M}$.

□

CHAPTER 5

Gauss sums and integrals

We now being our study of $\hat{1}_X(\theta)$ at the major arcs $\mathfrak{M}$. The most basic question is what happens when θ is actually equal to a rational $\frac{a}{q}$, where $(a,q) = 1$. Then, we have

$$\hat{1}_X(\theta) = \sum_{n \leqslant N^{1/k}} e(-\frac{a}{q}n^k) \approx \frac{N^{1/k}}{q} \sum_{b \in \mathbf{Z}/q\mathbf{Z}} e(-\frac{ab^k}{q}).$$

This last expression comes from splitting $n \leqslant N^{1/k}$ according to the residue class b of $n (\mathrm{mod}\, q)$, and in fact the $\approx$ is an $=$ if N is a multiple of q.

In the light of this, it makes sense to first study sums like the one over $b \in \mathbf{Z}/q\mathbf{Z}$. These are called Gauss sums.

Definition 5.0.1. Suppose that $a \in \mathbf{Z}/q\mathbf{Z}$. Then the Gauss sum $G_{a,q}$ is defined by

$$G_{a,q} := \frac{1}{q} \sum_{b \in \mathbf{Z}/q\mathbf{Z}} e(-\frac{ab^k}{q}).$$

We remark that (clearly) the Gauss sum also depends on k, but we will suppress explicit mention of this dependence; k will be clear from context.

Gauss sums are basically discrete Fourier transforms of the kth powers modulo q, and so are very natural mathematical objects in their own right in the context of this course.

The *trivial bound* for Gauss sums is $|G_{a,q}| \leqslant 1$, and this is sharp when $a = 0$. Much of the effort in this chapter will be directed towards improving this when a is coprime to q, and in particular we will establish the following result.

Proposition 5.0.1. *Suppose that* $a \in (\mathbf{Z}/q\mathbf{Z})^*$ *.Then we have* $|G_{a,q}| \ll q^{-1/k+o(1)}$.

We remark that with a little more care, the $o(1)$ in the exponent of Proposition 5.0.1 may be removed, but otherwise this bound is sharp in general.

5.1. Multiplicativity

The first key property of Gauss sums is quite elementary.

Lemma 5.1.1. *Suppose that* q_1, q_2 *are coprime and that* $a_i \in (\mathbf{Z}/q_i\mathbf{Z})^*$, $i = 1, 2$. *Then* $G_{a_1,q_1} G_{a_2,q_2} = G_{a_1q_2+a_2q_1, q_1q_2}$.

Proof. By a simple change of variables we may write

$$G_{a_1,q_1}G_{a_2,q_2} = \frac{1}{q_1q_2}\sum_{x_1\in\mathbf{Z}/q_1\mathbf{Z}}\sum_{x_2\in\mathbf{Z}/q_2\mathbf{Z}} e\big(-\frac{a_1}{q_1}(q_2x_1)^k - \frac{a_2}{q_2}(q_1x_2)^k\big).$$

Now from the binomial theorem it follows immediately that

$$(a_1q_2 + a_2q_1)(q_2x_1 + q_1x_2)^k \equiv a_1q_2(q_2x_1)^k + a_2q_1(q_1x_2)^k (\operatorname{mod} q_1q_2),$$

and hence that

$$e\big(\frac{a_1q_2 + a_2q_1}{q_1q_2}(q_2x_1 + q_1x_2)^k\big) = e\big(\frac{a_1}{q_1}(q_2x_1)^k + \frac{a_2}{q_2}(q_1x_2)^k\big).$$

Thus

$$G_{a_1,q_1}G_{a_2,q_2} = \sum_{x_1,x_2} e\big(-\frac{a_1q_2 + a_2q_1}{q_1q_2}(q_2x_1 + q_1x_2)^k\big) = G_{a_1q_2+a_2q_1,q_1q_2},$$

as stated. □

As a consequence of this lemma is that we may reduce the proof of Proposition 5.0.1 to the case in which q is a prime power. In that case, we will in fact prove a stronger result, without the $o(1)$ in the exponent.

LEMMA 5.1.2. *Suppose that q is a prime power and that $a \in (\mathbf{Z}/q\mathbf{Z})^*$. Then $|G_{a,q}| \leqslant 6kq^{-1/k}$.*

Whilst this is essentially optimal for general prime powers, when $q = p$ is prime one can prove a stronger bound. We will do this in Section 5.2 below.

To conclude this section, let us show how Lemma 5.1.2 implies Proposition 5.0.1.

Proof. [Proof of Proposition 5.0.1] By Lemmas 5.1.1 and 5.1.2, we immediately obtain

$$|G_{a,q}| \leqslant (6k)^{\omega(q)}q^{-1/k},$$

where $\omega(q)$ is the number of distinct prime factors of q. However, $\omega(q) = o(\log q)$ (the smallest number with m distinct prime factors is $p_1 \cdots p_m$, the product of the m smallest primes, which has size $e^{m\log m}$ and so $(4k)^{\omega(q)} = q^{o(1)}$). □

5.2. Prime moduli and Waring's problem (mod p)

As a warmup to the proof of Lemma 5.1.2, we look at the case in which q is a prime. The proof is easier in this case and the bound is sharper, but it is based on similar ideas. Moreover, we can apply this result to get "Waring's problem (mod p)", which we will need later on when analysing the singular series.

LEMMA 5.2.1. *Suppose that p is a prime and that $a \in (\mathbf{Z}/p\mathbf{Z})^*$. Then $|G_{a,p}| \leqslant kp^{-1/2}$.*

Proof. Note the invariance property $G_{a,p} = G_{ab^k,p}$ whenever $(b,p) = 1$. The equation $b^k \equiv 1(\operatorname{mod} p)$ has at most k solutions (since $\mathbf{Z}/p\mathbf{Z}$ is a field) so, for fixed a, no element of $(\mathbf{Z}/p\mathbf{Z})^*$ can be written as ab^k in more than k ways. Therefore

$$(p-1)|G_{a,p}|^2 = \sum_{b \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{ab^k,p}|^2 \leqslant k \sum_{r \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{r,p}|^2. \tag{5.1}$$

To estimate the sum on the right, extend it to all $r \in \mathbf{Z}/p\mathbf{Z}$, expand, and use the orthogonality relations. This gives

$$\begin{aligned}\sum_r |G_{r,p}|^2 &= \frac{1}{p^2}\sum_{x,y}\sum_r e(-\frac{r}{p}(x^k - y^k)) \\ &= \frac{1}{p}\#\{x,y \in \mathbf{Z}/p\mathbf{Z} : x^k \equiv y^k(\operatorname{mod} p)\}. \qquad (5.2)\end{aligned}$$

The number of pairs x, y with $x^k = y^k(\operatorname{mod} p)$ is at most $1 + k(p-1)$; once $x \neq 0$ is fixed, there are at most k choices for y, and moreover if $x = 0$ then $y = 0$. Deploying this in (5.2) and subtracting off the contribution from $r = 0$, it follows that

$$\sum_{r \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{r,p}|^2 \leqslant \frac{1}{p}(1 + k(p-1)) - 1 = \frac{(k-1)(p-1)}{p} < \frac{k(p-1)}{p}. \tag{5.3}$$

Finally, comparing with (5.1) gives the claimed bound. □

The next lemma solves Waring's problem modulo p, at least if p is sufficiently large. We show that just three kth powers are required.

Lemma 5.2.2. *Suppose that $p \geqslant k^4$. Then there are $x_1, x_2, x_3 \in \mathbf{Z}/p\mathbf{Z}$, not all zero, such that $x_1^k + x_2^k + x_3^k \equiv N(\operatorname{mod} p)$.*

Proof. By the orthogonality relations, the number T of triples satisfying

$$x_1^k + x_2^k + x_3^k \equiv N(\operatorname{mod} p) \tag{5.4}$$

is

$$\frac{1}{p}\sum_{x_1,x_2,x_3}\sum_a e\Big(\frac{a(x_1^k + x_2^k + x_3^k - N)}{p}\Big) = p^2 \sum_{a \in \mathbf{Z}/p\mathbf{Z}} G_{a,p}^3 e(\frac{aN}{p}).$$

The contribution from $a = 0$ is p^2, and therefore

$$T = p^2 + p^2 \sum_{a \in (\mathbf{Z}/p\mathbf{Z})^*} G_{a,p}^3 e(\frac{aN}{p}) \geqslant p^2 - p^2 \sum_{a \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{a,p}|^3. \tag{5.5}$$

Now we use the bound of Lemma 5.2.1, but also the bound for the second moment obtained in the proof, namely (5.3). It follows that

$$\sum_{a \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{a,p}|^3 \leqslant \sup_{a \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{a,p}| \sum_{a \in (\mathbf{Z}/p\mathbf{Z})^*} |G_{a,p}|^2$$
$$\leqslant \frac{k}{\sqrt{p}} \cdot (k-1).$$

Hence, from (5.5),

$$T \geqslant p^2\big(1 - \frac{k^2}{\sqrt{p}}\big) + kp^{3/2}.$$

If $p \geqslant k^4$, the first term is non-negative. The second term is certainly $\geqslant 2$. Thus $T \geqslant 2$, and so there are at least two solutions to (5.4), at least one of which does not have $x_1 = x_2 = x_3 = 0$. □

Remark. If you complete Example Sheet 2, you will see that a similar result can be established with just two summands, instead of three.

5.3. Prime power moduli

We turn now to the estimation of Gauss sums $G_{a,q}$ in the case q a prime power, the aim being to prove Lemma 5.1.2. The arguments are similar to those in the prime case, but more involved. The chief difficulty is in estimating the analogue of (5.2).

We begin with two simple lemmas about finite abelian groups.

Lemma 5.3.1. *In any cyclic group $\mathbf{Z}/m\mathbf{Z}$ (written additively) there are at most k solutions x to $kx = 0$.*

Proof. The $\times k$ map which sends x to kx is a homomorphism. Its image has size at least m/k, since the images of the elements $0, 1, \dots \lceil m/k \rceil - 1$ are all distinct. The kernel therefore has size at most k. □

Lemma 5.3.2. *Let q be an odd prime power. Then there are at most k kth roots of unity in $(\mathbf{Z}/q\mathbf{Z})^*$. If q is a power of two then there are at most $2k$ kth roots of unity in $(\mathbf{Z}/q\mathbf{Z})^*$.*

Proof. Suppose that $q = p^\nu$. We recall some facts about the group $(\mathbf{Z}/q\mathbf{Z})^*$, the proofs of which may be found in standard elementary number theory texts. This group has order $p^{\nu-1}(p-1)$. When p is odd, it is cyclic. When $p = 2$, it is isomorphic to the product of $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2^{\nu-2}\mathbf{Z}$. The result therefore follows from Lemma 5.3.1. □

Remark. When $q = 8$, $k = 2$ there *are* $2k$ kth roots of unity, $1, 3, 5$ and 7.

The next lemma, which generalises (5.2) to prime powers, is not quite so straightforward.

LEMMA 5.3.3. *Let $k \geqslant 3$, and let q be a prime power. Then the number of pairs $x, y \in \mathbf{Z}/q\mathbf{Z}$ with $x^k = y^k$ (in $\mathbf{Z}/q\mathbf{Z}$) is at most $8kq^{2(1-1/k)}$.*

Proof. Suppose that $x = p^\lambda t$ with $0 < t < p^{\nu-\lambda}$ and that $y = p^\mu u$ with $0 < u < p^{\nu-\mu}$, and with both t and u coprime to p. Then either (1) $\lambda, \mu \geqslant \nu/k$ (so both x^k and y^k are divisible by p^ν), or else (2) $\lambda = \mu$ and $t^k \equiv u^k (\operatorname{mod} p^{\nu-k\lambda})$.

The number of pairs satisfying (1) is at most $p^{2\nu(1-1/k)}$.

To count the number of pairs satisfying (2), choose t arbitrarily (at most $p^{\nu-\lambda}$ choices), then note that u lies in one of at most $2k$ residue classes modulo $p^{\nu-k\lambda}$. This gives at most $2kp^{(k-1)\lambda}$ choices for u, given t, so the number of pairs satisfying (2) for a given λ is at most $2kp^{\nu+(k-2)\lambda}$. We must now sum this over $\lambda < \nu/k$. Since $k \geqslant 3$, we have a geometric series dominated by the larger values of λ, so the sum is certainly at most $4kp^{\nu+(k-2)\nu/k} = 4kp^{2\nu(1-1/k)}$. □

We now turn to the proof of Lemma 5.1.2.

Proof. [Proof of Lemma 5.1.2] We will handle the cases $k = 2$ and $k \geqslant 3$ separately.

Case $k = 2$. Then

$$|G_{a,q}|^2 = \frac{1}{q^2} \sum_{x,y \in \mathbf{Z}/q\mathbf{Z}} e(\frac{a(y^2 - x^2)}{q}) = \frac{1}{q^2} \sum_{h \in \mathbf{Z}/q\mathbf{Z}} e(\frac{ah^2}{q}) \sum_{x \in \mathbf{Z}/q\mathbf{Z}} e(\frac{2ahx}{q}),$$

where here we made the substitution $y = x + h$. Using the orthogonality relations for the inner sum over x and the triangle inequality, we obtain

$$|G_{a,q}|^2 \leqslant \frac{1}{q} \#\{h \in \mathbf{Z}/q\mathbf{Z} : 2ah \equiv 0 (\operatorname{mod} q)\}.$$

Since a is coprime to q, this is just the number of $h \in \mathbf{Z}/q\mathbf{Z}$ with $2h \equiv 0$, which is either 1 (if q is odd) or 2 (if q is even). This completes the proof in the case $k = 2$.

Case $k \geqslant 3$. Write $q = p^\nu$.

Note the invariance property $G_{a,p^\nu} = G_{ab^k,p^\nu}$ whenever $(b,p) = 1$. By Lemma 5.3.2, no element of $(\mathbf{Z}/q\mathbf{Z})^*$ is ab^k in more than $2k$ ways. Therefore

$$p^{\nu-1}(p-1)|G_{a,p^\nu}|^2 = \sum_{b \in (\mathbf{Z}/p^\nu\mathbf{Z})^*} |G_{ab^k,p^\nu}|^2 \leqslant 2k \sum_{r \in \mathbf{Z}/p^\nu\mathbf{Z}} |G_{r,p^\nu}|^2. \tag{5.6}$$

Note carefully that the sum over r on the right really is over all of $\mathbf{Z}/p^\nu\mathbf{Z}$, not just $(\mathbf{Z}/p^\nu\mathbf{Z})^*$; we may do this since all the terms are positive. Expanding out we obtain

$$\sum_r |G_{r,p^\nu}|^2 = \frac{1}{p^{2\nu}} \sum_{x,y} \sum_r e(-\frac{r}{p^\nu}(x^k - y^k))$$

$$= \frac{1}{p^\nu} \#\{x, y \in \mathbf{Z}/p^\nu\mathbf{Z} : x^k = y^k\}. \tag{5.7}$$

By Lemma 5.3.3, this is bounded by $8kp^{\nu-2\nu/k}$. Comparing (5.6) and (5.7), and using the bound $p - 1 \geqslant \frac{1}{2}p$, we therefore have

$$\frac{1}{2}p^\nu |G_{a,p^\nu}|^2 \leqslant 2k \cdot 8k \cdot p^{\nu-2\nu/k},$$

which quickly implies Lemma 5.1.2. □

5.4. Integrals

In addition to the discrete Gauss sums mentioned above, we will also need the following integrals.

Definition 5.4.1. For a real parameter t, define

$$I(t) := \int_0^{N^{1/k}} e(-tx^k)dx.$$

Evidently $I(t)$ depends on N and on k, as well as on t, but we will suppress this. It is a kind of Fourier transform of the kth power function on $\mathbf{R}$.

Obviously we have the trivial bound $|I(t)| \leqslant N^{1/k}$. We also have the following less trivial bound, somewhat analogous to the Gauss sum bound (but not containing any arithmetic information).

Lemma 5.4.1. *We have $|I(t)| \ll |t|^{-1/k}$.*

Proof. Suppose that $t > 0$ (the proof in the case $t < 0$ is almost identical). Making the substitution $w = tx^k$ in the definition of $I(t)$, we get

$$I(t) = \frac{1}{k}t^{-1/k} \int_0^{Nt} e(-w)w^{-1+1/k}dw.$$

Therefore it suffices to prove that

$$|\int_0^Z e(-w)w^{-1+1/k}dw| = O(1) \tag{5.8}$$

uniformly in Z. A bound for the integral from 1 to Z follows quickly by integration by parts:

$$\int_1^Z e(-w)w^{-1+1/k} = O(1) - (2\pi i)^{-1}(-1 + \frac{1}{k}) \int_1^Z e(-w)w^{-2+1/k}dw,$$

and the integral on the right is bounded by

$$\int_1^Z w^{-2+1/k} dw \ll 1.$$

We also have

$$|\int_0^1 e(-w) w^{-1+1/k} dw| \leqslant \int_0^1 w^{-1+1/k} dw \ll_k 1,$$

and the claim (5.8), and hence the lemma, follows. □

CHAPTER 6

The major arcs

Notation. Recall that $\eta = 1/10k$ is the exponent appearing in Definition 3.2.1, the definition of major and minor arcs.

Our aim in this chapter is to establish Proposition 3.6, that is to say the estimate

$$\int_{\mathfrak{M}} \hat{1}_X(\theta)^s e(N\theta) d\theta = \mathfrak{S}_{k,s}(N) N^{s/k-1} + o(N^{s/k-1})$$

under the assumption that $s \geqslant 2k+1$. This is a somewhat lengthy task, but it does split into manageable parts. The first thing to note is that the integral over $\mathfrak{M}$ splits into a sum over $q \leqslant N^\eta$ and $a \in (\mathbf{Z}/q\mathbf{Z})^*$ of the separate integrals

$$\int_{\mathfrak{M}_{a,q}} \hat{1}_X(\theta)^s e(N\theta) d\theta.$$

This is *almost* completely obvious from the definition of $\mathfrak{M}$ as $\bigcup_{a,q} \mathfrak{M}_{a,q}$, but one does need to check that the $\mathfrak{M}_{a,q}$ are disjoint. This is easy: if $\theta \in \mathfrak{M}_{a,q} \cap \mathfrak{M}_{a',q'}$ then

$$\|\frac{a}{q} - \frac{a'}{q'}\|_{\mathbf{T}} \leqslant \|\theta - \frac{a}{q}\|_{\mathbf{T}} + \|\theta - \frac{a'}{q'}\|_{\mathbf{T}} \leqslant 2N^{-1+2\eta}.$$

On the other hand,

$$\|\frac{a}{q} - \frac{a'}{q'}\|_{\mathbf{T}} \geqslant \frac{1}{qq'} \geqslant N^{-2\eta}.$$

Since $\eta \leqslant \frac{1}{10}$, this certainly leads to a contradiction for N large.

6.1. A single point of a major arc

At the beginning of Chapter 5, we related the Fourier transform $\hat{1}_X(\theta)$ when $\theta = \frac{a}{q}$ to a Gauss sum.

In this section we spread the net a little wider, looking at the case $\theta \approx \frac{a}{q}$, or in other words at $\hat{1}_X(\theta)$ for $\theta \in \mathfrak{M}_{a,q}$. Here is the main result. Here, recall the definition of Gauss sums $G_{a,q}$ (Definition 5.0.1) and the integrals $I(t)$ (Definition 5.4.1).

Proposition 6.1.1. *For $\theta \in \mathfrak{M}_{a,q}$ we have*

$$\hat{1}_X(-\theta) = G_{a,q} I(\theta - \frac{a}{q}) + O(N^{4\eta}).$$

Before beginning the proof, let us appraise the nature of this task in the simplest case, where $q = 1$ and $a = 0$. Then the Gauss sum is simply 1, and the statement is that

$$\sum_{n \leqslant N^{1/k}} e(-\theta n^k) \approx \int_0^{N^{1/k}} e(-\theta x^k) dx,$$

provided that θ is suitably small. This is a very familiar kind of statement, in which an integral is approximated by a sum. There are many ways, some quite sophisticated[1], to prove statements of this type. In our case we can get away with a very simple argument based on the fact that the integrand is slowly-varying. We turn now to the detailed argument, which we recommend the reader follow through in the case $q = 1$, $a = 0$.

Proof. Let $\theta = \frac{a}{q} + t$. Then

$$\hat{1}_X(\theta) = \sum_{n \leqslant N^{1/k}} e(-(\frac{a}{q} + t)n^k).$$

Splitting the sum over n into residue classes b mod q, we get

$$\hat{1}_X(\theta) = \sum_{b \in \mathbf{Z}/q\mathbf{Z}} e(-\frac{ab^k}{q}) \sum_{\substack{n \leqslant N^{1/k} \\ n \equiv b (\operatorname{mod} q)}} e(-tn^k). \tag{6.1}$$

We claim that

$$q \sum_{\substack{n \leqslant N^{1/k} \\ n \equiv b (\operatorname{mod} q)}} e(-tn^k) = I(t) + O(N^{4\eta}). \tag{6.2}$$

Once this is shown, it then follows from (6.1) that

$$\begin{aligned} \hat{1}_X(\theta) &= \sum_{b \in \mathbf{Z}/q\mathbf{Z}} e(-\frac{ab^k}{q})\big(\frac{1}{q} I(t) + O(\frac{N^{4\eta}}{q})\big) \\ &= G_{a,q} I(t) + O(N^{4\eta}), \end{aligned}$$

which is the claimed result.

It remains to establish (6.2). Splitting up into intervals of length q, we have

$$I(t) = \sum_{\substack{n \leqslant N^{1/k} \\ n \equiv b (\operatorname{mod} q)}} \int_n^{n+q} e(-tx^k) dx + O(q),$$

[1] The Euler–Maclaurin summation formula is most relevant here; the Poisson summation formula is important in other contexts.

where the $O(q)$ term comes from the endpoints. It follows that

$$\begin{aligned}\big|I(t) - q \sum_{\substack{n \leqslant N^{1/k} \\ n \equiv b (\operatorname{mod} q)}} e(-tn^k)\big| \\ &\leqslant N^{1/k} \sup_{n \leqslant N^{1/k}} \big| \int_n^{n+q} e(-tx^k) dx - q e(-tn^k) \big| + O(q) \\ &\leqslant q N^{1/k} \sup_{n \leqslant N^{1/k}} \sup_{0 \leqslant x - n \leqslant q} |e(-tx^k) - e(-tn^k)| + O(q).\end{aligned}$$

But if n, x satisfy the stated inequalities then by the binomial theorem and the fact that $|t| \leqslant N^{-1+2\eta}$, $q \leqslant N^{\eta}$ we have

$$tx^k = tn^k + O(tqn^{k-1}) = tn^k + O(N^{3\eta - 1/k}).$$

Therefore

$$|e(-tx^k) - e(-tn^k)| = O(N^{3\eta - 1/k}).$$

The estimate (6.2) then follows, using the fact that $q \leqslant N^{\eta}$ again.

□

6.2. Integrating over a major arc

Proposition 6.1.1 gives an expansion for $\hat{1}_X(\theta)$ at a single point $\theta \in \mathfrak{M}_{a,q}$. The next step is to work out the contribution this gives when (after raising to the power s and multiplying by $e(N\theta)$) we integrate over the whole major arc $\mathfrak{M}_{a,q}$.

Here is the answer.

Proposition 6.2.1. *We have*

$$\int_{\mathfrak{M}_{a,q}} \hat{1}_X(\theta)^s e(N\theta) d\theta = G_{a,q}^s e(Na/q) \int_{-\infty}^{\infty} I(t)^s e(Nt) dt + o(N^{s/k - 1 - 2\eta}).$$

Remark. The error term is not best possible; it is designed simply to be $o(N^{s/k-1})$ when summed over a, q, which we shall do in the next section.

Proof. We must first raise the conclusion of Proposition 6.1.1 to the power s. Using the trivial bound $|G_{a,q} I(\theta - \frac{a}{q})| \leqslant N^{1/k}$, together with the binomial theorem, we have

$$\hat{1}_X(\theta)^s = G_{a,q}^s I(\theta - \frac{a}{q})^s + O(N^{(s-1)/k + 4\eta}).$$

Integrating over $\mathfrak{M}_{a,q}$, which has length $N^{-1+2\eta}$, it follows that

$$\begin{aligned}\int_{\mathfrak{M}_{a,q}} & \hat{1}_X(\theta)^s e(N\theta)d\theta \\ &= G_{a,q}^s e(Na/q)\int_{\theta\in\mathfrak{M}_{a,q}} I(\theta-\frac{a}{q})^s e(N(\theta-a/q)) + O(N^{-1+(s-1)/k+6\eta}) \\ &= G_{a,q}^s e(Na/q)\int_{|t|\leqslant N^{-1+2\eta}} I(t)^s e(Nt)dt + O(N^{-1+(s-1)/k+6\eta}).\end{aligned}$$

The error term is bounded as required in the proposition, by the choice of η ($=1/10k$). This is almost what we want, except that the integral over t must be extended to $\pm\infty$. Using the trivial bound $|G_{a,q}|\leqslant 1$, it is enough to show that

$$\int_{N^{-1+2\eta}}^{\infty} |I(t)|^s dt = o(N^{s/k-1-2\eta}). \tag{6.3}$$

(as well as a corresponding bound down to $-\infty$, proved the same way). This follows from Lemma 5.4.1, that is to say the bound $|I(t)|\ll|t|^{-1/k}$, since then

$$\int_{N^{-1+2\eta}}^{\infty}|I(t)|^s dt \ll \int_{N^{-1+2\eta}}^{\infty} t^{-s/k}dt = \frac{(N^{1-2\eta})^{s/k-1}}{s/k-1} = o(N^{s/k-1-2\eta}).$$

In the last step we used the fact that $s\geqslant 2k+1$, but this is certainly not the critical use of this assumption, which will come later. □

6.3. The sums $A(q)$

The next main task in our development is to take the result of the last section and sum it over all major arcs. When we do this (in Section 6.4 below) a certain sum $A(q)$ will appear. This sum turns out to be natural in the theory and will reappear several times later. For these reasons, we derive basic bounds and properties for this quantity now.

Definition 6.3.1. Let $q\geqslant 1$ be an integer. We define

$$A(q) := \sum_{a\in(\mathbf{Z}/q\mathbf{Z})^*} G_{a,q}^s e(Na/q),$$

where $G_{a,q}$ is the Gauss sum over kth powers.

Remark. $A(q)$ depends on k, s and N, as well as on q, but we are thinking of these quantities as fixed for the duration of the argument, so we suppress this dependence to ease the notation.

Lemma 6.3.1 (Bounds for $A(q)$). *We have the following bounds, uniformly in N.*

(i) $|A(q)|\ll q^{1-s/k+o(1)}$ *(uniformly for all integers q. If q is a prime power, we can omit the $o(1)$ term;*

(ii) *If* $s \geqslant k+1$, $\sum_j |A(p^j)| \ll 1$, *uniformly for all primes* p;

(iii) *If* $s \geqslant 2k+1$, $\sum_q |A(q)| \ll 1$.

Proof. (i) The first statement follows immediately from Proposition 5.0.1, that is to say the bound $|G_{a,q}| \ll q^{-1/k+o(1)}$, and the triangle inequality. The fact that we can omit the $o(1)$ in the prime power case reflects the fact that we can omit the $o(1)$ in the bound for Gauss sums, in the prime power case, that is to say Lemma 5.1.2.

(ii) By (i) (the prime power case) $|A(p^j)| \ll p^{j(1-s/k)}$, and this is $\ll p^{-j/k}$ if $s \geqslant k+1$. When summed over j, this is a geometric series with ratio $p^{-1/k}$. This converges rapidly, with $p = 2$ being the worst case.

(iii) By (i) (the general q case) $|A(q)| \ll q^{-1-1/k}$ when $s \geqslant 2k+1$. This converges when summed over q. □

Additionally, $A(q)$ is multiplicative.

Lemma 6.3.2. *Suppose that* $(q_1, q_2) = 1$. *Then* $A(q_1q_2) = A(q_1)A(q_2)$.

Proof. We have

$$\begin{aligned} A(q_1)A(q_2) &= \sum_{a_1 \in (\mathbf{Z}/q_1\mathbf{Z})^*} \sum_{a_2 \in (\mathbf{Z}/q_2\mathbf{Z})^*} G^s_{a_1,q_1} G^s_{a_2,q_2} e(a_1N/q_1)e(a_2N/q_2) \\ &= \sum_{a_1 \in (\mathbf{Z}/q_1\mathbf{Z})^*} \sum_{a_2 \in (\mathbf{Z}/q_2\mathbf{Z})^*} G^s_{a_1q_2+a_2q_1,q} e(\frac{a_1q_2+a_2q_1}{q_1q_2}N) \\ &= \sum_{a \in (\mathbf{Z}/q\mathbf{Z})^*} G^s_{a,q} e(aN/q) \\ &= A(q_1q_2). \end{aligned}$$

□

Corollary 6.3.1. *Suppose that* $s \geqslant 2k+1$. *Then*

$$\sum_q A(q) = \prod_p (\sum_{j=0}^{\infty} A(p^j)). \tag{6.4}$$

Proof. (Sketch) Formally, this is obvious from Lemma 6.3.2 and unique factorisation into primes. The bounds of Lemma 6.3.1 guarantee that everything converges absolutely and that the rearrangement is permissible. We omit a detailed justification. (Hint for the interested reader: start on the right hand side and truncate the product to $p \leqslant P$ and the sums to $j \leqslant J$. Then let $J \to \infty$, then $P \to \infty$.) □

6.4. Integrating over all major arcs

In this section we return to our main line of argument. Recall that we have an estimate for the integral of $\hat{1}_X(\theta)^s e(N\theta)$ over a single major arc $\mathfrak{M}_{a,q}$ (Proposition 6.2.1). We wish to sum this over all a, q. Here is the result of doing this. (It may be helpful to recall the definitions of $I(t)$ (Definition 5.4.1) and of $A(q)$ (Definition 6.3.1).)

Proposition 6.4.1. *We have*

$$\int_{\mathfrak{M}} \hat{1}_X(\theta)^s e(N\theta)d\theta = \int_{-\infty}^{\infty} I(t)^s e(Nt)dt \sum_q A(q) + o(N^{s/k-1}).$$

Proof. Sum the result of Proposition 6.2.1 over all the major arcs $\mathfrak{M}_{a,q}$, that is to say over all $q \leqslant N^\eta$ and all $a \in (\mathbf{Z}/q\mathbf{Z})^*$. This being at most $N^{2\eta}$ values of a, q, the error term remains $o(N^{s/k-1})$ after performing this sum.

Hence

$$\int_{\mathfrak{M}} \hat{1}_X(\theta)^s e(N\theta)d\theta = \int_{-\infty}^{\infty} I(t)^s e(Nt)dt \sum_{q \leqslant N^\eta} \sum_{a \in (\mathbf{Z}/q\mathbf{Z})^*} G^s_{a,q} e(Na/q) + o(N^{s/k-1})$$

$$= \int_{-\infty}^{\infty} I(t)^s e(Nt)dt \sum_{q \leqslant N^\eta} A(q) + o(N^{s/k-1}). \tag{6.5}$$

To finish the proof of Proposition 6.4.1, all we need do is show that the sum over q may be extended all the way to ∞ without enlarging the error term; that is, it is enough to show that

$$\int_{-\infty}^{\infty} I(t)^s e(Nt)dt \sum_{N^\eta < q < \infty} A(q) = o(N^{s/k-1}). \tag{6.6}$$

Using Lemma 5.4.1 and the trivial bound $|I(t)| \leqslant N^{1/k}$, we have

$$|\int_{-\infty}^{\infty} I(t)^s e(Nt)dt| \ll \int_{-\infty}^{\infty} \min(N^{s/k}, |t|^{-s/k})dt \ll N^{s/k-1}$$

(consider the integrals over $|t| \leqslant \frac{1}{N}$ and $|t| > \frac{1}{N}$ separately). Since $\sum_q |A(q)| \ll 1$ (Lemma 6.3.1 (iii)), (6.6) follows. □

6.5. The remaining task

Let us compare the result of Proposition 6.4.1 with our goal, the major arcs estimate Proposition 3.2.1.

To complete the proof of Proposition 3.2.1, it is enough to show that

$$\int_{-\infty}^{\infty} I(t)^s e(Nt)dt \sum_q A(q) = \mathfrak{S}_{k,s}(N) N^{s/k-1},$$

which follows if we can show that

$$\int_{-\infty}^{\infty} I(t)^s e(Nt)dt = \beta_\infty N^{s/k-1} = \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)} N^{s/k-1}, \tag{6.7}$$

and that

$$\sum_q A(q) = \prod_p \beta_p(N). \tag{6.8}$$

(Implicit in (6.8) is the assertion that $\beta_p(N)$ exists.)

Note that these are *formulae*; there are no error terms. This suggests the proofs should be more or less formal calculations, and that suggestion is correct. We give the details in the next two sections.

6.6. *The archimedean prime

In this section we prove (6.7).

Making two obvious substitutions ($t = u/N$ in (6.7) and $x = N^{1/k}y^{1/k}$ in the definition of $I(t)$) we may immediately reduce (6.7) to the task of proving that

$$\int_{-\infty}^{\infty} \big(\frac{1}{k}\int_0^1 y^{-1+1/k} e(-uy)dy\big)^s e(u)du = \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)}. \tag{6.9}$$

This can be proven using the basic facts about the Fourier transform on $\mathbf{R}$, as described in Section 1.1. We will proceed formally, leaving the verification that the analytic conditions of the results of Section 1.1 (which we did not, in any case, carefully state or prove) are valid.

We have

$$\frac{1}{k}\int_0^1 y^{-1+1/k} e(-uy)dy = \hat{f}(u),$$

where $f : \mathbf{R} \to \mathbf{R}$ is the function

$$f(y) := \frac{1}{k} y^{-1+1/k} 1_{[0,1]}(y).$$

Therefore the left-hand side is

$$\int_{-\infty}^{\infty} \hat{f}(u)^s e(u)du.$$

Noting that $\hat{f}^s$ is the Fourier transform of the s-fold autoconvolution $f * f * \cdots * f$, and assuming the Fourier inversion formula holds, this equals

$$(f * \cdots * f)(1) = k^{-s} \int_{y_i \geqslant 0} (y_1 \cdots y_{s-1}(1 - y_1 - \cdots - y_{s-1}))^{-1+1/k} dy_1 \ldots dy_{s-1}.$$

On the other hand,

$$\begin{aligned}\Gamma(1/k)^s &= \big(\int_0^\infty e^{-v}v^{1/k-1}dv\big)^s\\ &= \int_0^\infty \cdots \int_0^\infty e^{-v_1-\cdots-v_s}(v_1\cdots v_s)^{1/k-1}dv_1\cdots dv_s.\end{aligned}$$

Make the substitution $z = v_1 + \cdots + v_s$, $y_i = v_i/z$, $i = 1, \dots, s-1$. We have $\partial v_i/\partial z = y_i$, $i = 1, \dots, s-1$, and $\partial v_i/\partial y_j = 1$ when $i = j$ and -1 when $i = s$. Therefore the Jacobean is

$$\begin{vmatrix} y_1 & \dots & y_{s-1} & 1-y_1-\cdots-y_{s-1}\\ z & 0 & 0 & -z\\ \vdots & z & 0 & -z\\ 0 & 0 & z & -z\end{vmatrix}$$

Adding the first $(s-1)$ columns to the last shows that this is z^{s-1}, and so we have

$$\begin{aligned}\Gamma(1/k)^s &= \int_{y_i\geqslant 0}\big(\int_0^\infty e^{-z}z^{s/k-1}dz\big)(y_1\cdots y_{s-1}(1-y_1-\cdots-y_{s-1}))^{1/k-1}\\ &= \Gamma(s/k)\int_{y_i\geqslant 0}(y_1\cdots y_{s-1}(1-y_1-\cdots-y_{s-1}))^{1/k-1}.\end{aligned}$$

Putting all this together gives

$$(f*\cdots*f)(1) = k^{-s}\frac{\Gamma(1/k)^s}{\Gamma(s/k)} = \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)},$$

which concludes the proof of (6.9). (This is basically a well-known evaluation of what is called the *Beta integral* in terms of Γ-functions.)

6.7. The non-archimedean primes

In this section we establish (6.8), that is to say that

$$\sum_q A(q) = \prod_p \beta_p(N).$$

Recall (Corollary 6.3.1) that

$$\sum_q A(q) = \prod_p(\sum_j A(p^j)).$$

Therefore it is sufficient, and extremely natural, to try and prove that

$$\beta_p(N) = \sum_j A(p^j) \tag{6.10}$$

(with the existence of $\beta_p(N)$ being part of this statement).

Let us recall the definition of the p-adic density $\beta_p(N)$, namely

$$\beta_p(N) = \lim_{n\to\infty} \beta_{p,n}(N),$$

where $\beta_{p,n}(N)$ is the $(\mathbf{Z}/p^n\mathbf{Z})$-density

$$\beta_{p,n}(N) := p^{-(s-1)n}\#\{(x_1,\dots,x_s) \in (\mathbf{Z}/p^n\mathbf{Z})^s : x_1^k + \dots + x_s^k \equiv N (\operatorname{mod} p^n)\}.$$

We will show that

$$\beta_{p,n}(N) = \sum_{j \leqslant n} A(p^j). \tag{6.11}$$

Since $\sum_j |A(p^j)|$ converges (Lemma 6.3.1 (ii)), letting $n \to \infty$ establishes (6.10) and, at the same time, the existence of $\beta_p(N)$.

The remaining task, then, is to prove (6.11). We do this now.

First, observe that

$$\beta_{p,n}(N) = \sum_{a \in \mathbf{Z}/p^n\mathbf{Z}} G^s_{a,p^n} e(aN/p^n). \tag{6.12}$$

This follows immediately by substituting in the definition of the Gauss sum and using the orthogonality relations on $\mathbf{Z}/p^n\mathbf{Z}$. Split the sum over a according to the highest power p^{n-j} of p dividing a, thus a ranges over $p^{n-j}a'$ with $a' \in (\mathbf{Z}/p^j\mathbf{Z})^*$. It is easy to check that $G_{a,p^n} = G_{a',p^j}$, and of course $e(aN/p^n) = e(a'N/p^j)$, and so the contribution from a particular j is precisely $A(p^j)$. Summing over j establishes (6.11).

This concludes the proof of (6.8), and hence the proof of Proposition 3.2.1.

CHAPTER 7

The singular series

The analysis of the last three chapters has provided us with a formula for $r_{k,s}(N)$, the number of ways of writing N as a sum of s kth powers. The formula is given in Theorem 3.1.2. As it stands, this formula is not very useful, since we have yet to say anything substantive about the singular series $\mathfrak{S}_{k,s}(N)$.

In this chapter we make good this omission by proving Proposition 3.1.1, namely the statement that $\mathfrak{S}_{k,s}(N) \asymp 1$ for $s \geqslant k^4$.

Our analysis is rather crude, and with more refined arguments one may obtain a similar result for a larger range of s quite easily.

7.1. Bounding the p-adic densities

Recall from the last chapter the $(\mathbf{Z}/p^n\mathbf{Z})$-densities

$$\beta_{p,n}(N) := p^{-(s-1)n}\{(x_1, \dots, x_s) \in (\mathbf{Z}/p^n\mathbf{Z})^s : x_1^k + \cdots + x_s^k \equiv N (\operatorname{mod} p^n)\},$$

from which we define the p-adic density

$$\beta_p(N) = \lim_{n\rightarrow\infty} \beta_{p,n}(N).$$

In the last chapter we showed that this exists, and we also derived the formula (6.10), that is to say

$$\beta_p(N) = \sum_j A(p^j). \tag{7.1}$$

Recall that the quantities $A(q)$ are defined, and their basic properties developed, in Section 6.3. (Recall also the $\beta_p(N)$ depend on k and s, but we suppress explicit mention of this dependence.)

In this section, we show that the p-adic densities $\beta_p(N)$ are bounded above and below uniformly in N, at least when $s \geqslant k^4$. (This condition can certainly be weakened with more effort, especially for specific values of k: see Example Sheet 2.) The lower bound is the crux of the matter. The idea is to "lift" solutions to $x_1^k + \cdots + x_k^s \equiv N (\operatorname{mod} p)$ (which exist under suitable conditions by Lemma 5.2.2) to solutions modulo larger powers of p.

The following lemma (which is closely related to a special case of *Hensel's lemma*) drives this lifting procedure.

Lemma 7.1.1. *As usual, let $k \geqslant 2$ be an integer. Suppose that p is a prime and that $(a,p) = 1$. Let γ be the maximal exponent of p which divides k (thus $k = p^{\gamma}k_0$, with k_0 coprime to p). Then if a is a kth power modulo $p^{2\gamma+1}$, it is a kth power modulo all higher powers of p.*

Proof. Let $n \geqslant 2\gamma+1$, and suppose it is known that a is a kth power $(\operatorname{mod} p^n)$. We will show that a is a kth power $(\operatorname{mod} p^{n+1})$. Suppose that $x^k \equiv a (\operatorname{mod} p^n)$. For any integer t, the binomial theorem tells us that

$$(x+tp^{n-\gamma})^k \equiv x^k + k_0 t x^{k-1} p^n + \binom{k}{2} t^2 x^{k-2} p^{2(n-\gamma)} + \binom{k}{3} t^3 x^{k-3} p^{3(n-\gamma)} + \dots \tag{7.2}$$

Since $n \geqslant 2\gamma+1$, we have $2(n-\gamma) \geqslant n+1$, and so all except the first two terms are $0 (\operatorname{mod} p^{n+1})$. That is,

$$(x+tp^{n-\gamma})^k \equiv x^k + k_0 t x^{k-1} p^n (\operatorname{mod} p^{n+1}).$$

As t cycles through $0,1,\dots,p-1$, the right hand side assumes all of the p elements of $\mathbf{Z}/p^{n+1}\mathbf{Z}$ congruent to $a (\operatorname{mod} p^n)$. In particular, for some value of t, it is equal to $a (\operatorname{mod} p^{n+1})$. □

Proposition 7.1.1. *We have the following bounds.*

(i) *Suppose that $s \geqslant 2k+1$. Then $\beta_p(N) = 1 + O(p^{-1-1/k})$, where the $O()$ is uniform in p and N.*

(ii) *Suppose that $s \geqslant k^4$. Then $\beta_p(N) \gg 1$, uniformly in p and N.*

Remark. It is worth remarking that (ii) does not follow immediately from (i), since the implied constant in the $O()$ notation may be large, in which case (i) cannot tell us that $\beta_p(N) \neq 0$, at least for small p.

Proof. For (i), we use (7.1) to obtain

$$\beta_p(N) = 1 + \sum_{j \geqslant 1} A(p^j).$$

By Lemma 6.3.1 (i) (the prime power case), $|A(p^j)| \ll p^{-(1+1/k)j}$, so the bound follows immediately by summing the geometric series.

(ii) For p sufficiently large (as a function of k,s) this follows from (i). Therefore it suffices to prove (ii) for each of the remaining (small) primes p separately, with a bound which *can* depend on p.

As in Lemma 7.1.1, let γ be the maximal exponent of p which divides k. We claim that if $s \geqslant k^4$ then there is a solution to

$$y_1^k + \dots + y_s^k \equiv N (\operatorname{mod} p^{2\gamma+1}) \tag{7.3}$$

with $y_1 \neq 0$. If $\gamma = 0$, this follows immediately from Lemma 5.2.2 (in fact with $s = 3$) when $p \geqslant k^4$. Suppose that $p < k^4$. We may clearly assume that $1 \leqslant N \leqslant p < k^4$, and so in this case we have the trivial solution $y_1 = \cdots = y_N = 1$, $y_{N+1} = \cdots = y_s = 0$. This proves the claim when $\gamma = 0$.

Suppose that $\gamma \geqslant 1$. Then, since $p^\gamma | k$, we have $p^{2\gamma+1} \leqslant p^{3\gamma} \leqslant k^3 < s$. We may assume that $1 \leqslant N \leqslant p^{2\gamma+1} < s$, and so we may again take the trivial solution $y_1 = \cdots = y_N = 1$, $y_{N+1} = \cdots = y_s = 0$.

The claim is proven in all cases.

Now suppose that $n \geqslant 2\gamma + 1$. We are going to show that there are many solutions to

$$x_1^k + \cdots + x_s^k \equiv N (\operatorname{mod} p^n) \tag{7.4}$$

which "lift" the solution (7.3). To create these solutions, pick arbitrary $x_2, \dots, x_s \in \mathbf{Z}/p^n\mathbf{Z}$ with $x_i \equiv y_i (\operatorname{mod} p^{2\gamma+1})$. There are $(p^{n-2\gamma-1})^{s-1}$ choices for these x_i. For any such choice we have

$$N - x_2^k - \cdots - x_s^k \equiv y_1^k (\operatorname{mod} p^{2\gamma+1}).$$

Thus, by Lemma **??**, $N - x_2^k - \cdots - x_s^k$ is a kth power modulo p^n, and so there is a choice of $x_1 \in \mathbf{Z}/p^n\mathbf{Z}$ such that (7.4) holds. It follows that there are at least $(p^{n-2\gamma-1})^{s-1}$ solutions to (7.4), and so

$$\beta_{p,n}(N) \geqslant p^{-(s-1)n}(p^{n-2\gamma-1})^{s-1} = p^{-(s-1)(2\gamma+1)}$$

for all $n \geqslant 2\gamma + 1$. Taking limits as $n \to \infty$, $\beta_p(N) \geqslant p^{-(s-1)(2\gamma+1)}$, and so indeed $\beta_p(N) \gg_{p,k,s} 1$, uniformly in N.

As previous remarked, this is enough to conclude the proof of (ii). □

7.2. Bounding the singular series

Finally, we are ready to complete the last outstanding task from Chapter 3, the proof of Proposition 3.1.1. We recall the statement below.

The proof is just an application of what we have already shown and the following fact about infinite products, which is of a fairly standard type.

Lemma 7.2.1. *Let C be a constant, and suppose that $x_1, x_2, \dots$ is a sequence of real numbers (not necessarily positive) such that*

(i) $\frac{1}{C} \leqslant 1 + x_i \leqslant C$ *for all i;*
(ii) $|x_i| \leqslant \frac{1}{10}$ *for all $i \geqslant C$;*
(iii) $\sum |x_i| \leqslant C$.

Then $\prod_i (1 + x_i) \asymp_C 1$.

Proof. The product over $i < C$ is acceptable by (i), so we can restrict attention to the product over $i \geqslant C$. If $|t| \leqslant \frac{1}{10}$, $1 + t \leqslant e^{2|t|}$ (an easy calculus check) and so by (ii), (iii)

$$\prod_{i \geqslant C} (1 + x_i) \leqslant e^{2\sum_{i \geqslant C} |x_i|} \leqslant e^{2C}.$$

Similarly, using instead the lower bound $1 + t \geqslant e^{-2|t|}$,

$$\prod_{i \geqslant C} (1 + x_i) \geqslant e^{-2C}.$$

This concludes the proof. □

Proposition 3.1.1. For $s \geqslant k^4$ we have $\mathfrak{S}_{k,s}(N) \asymp 1$.

Proof. Recall that, by definition,

$$\mathfrak{S}_{k,s}(N) = \beta_\infty \prod_p \beta_p(N). \tag{7.5}$$

The archimedean factor β_∞ is clearly $\asymp 1$, so we need only prove that

$$\prod_p \beta_p(N) \asymp 1. \tag{7.6}$$

This follows from Lemma 7.2.1, taking the x_i to be the $\beta_p(N) - 1$, and C to be some suitably large constant (large enough in terms of k, s). Of the hypotheses in Lemma 7.2.1, (i) follows from Proposition 7.1.1 (ii), part (ii) follows from Proposition 7.1.1 (i), and finally part (iii) is a consequence of Proposition 7.1.1 (i) and the fact that $\sum_p p^{-1-1/k} < \infty$. □

CHAPTER 8

Hua's Lemma

This chapter may or may not be lectured, depending on time. Last year it was lectured right at the end of the course (after Part B: Additive Combinatorics). It is listed in the schedules as examinable, but if I do not lecture it, it will not be.

The main aim is to prove *Hua's Lemma*, which is the following statement.

We will be dealing with 2^k-tuples of integers. For various reasons it is convenient to index them with the cube $\{0,1\}^k$. If $\omega = (\omega_1, \dots, \omega_k) \in \{0,1\}^k$ then we write $|\omega| := \omega_1 + \cdots + \omega_k$.

Theorem 8.0.1 (Hua's Lemma). *Let $k \geqslant 2$ be an integer, and let D be a (typically much larger) integer. Then the number of 2^k-tuples $(x_\omega)_{\omega \in \{0,1\}^k}$ with $x_\omega \in \{1, \dots, D\}$ for all ω and*

$$\sum_\omega (-1)^\omega x_\omega^k = 0$$

is $\ll D^{2^k - k + o(1)}$.

The main reason for proving this is that it allows us, relatively easily, to show that the asymptotic formula Theorem 3.1.2 holds with the rather better bound $s \geqslant 2^k + 1$. We will comment on this in Section 8.4.

8.1. The divisor bound

A key ingredient in the proof of Hua's lemma is the *divisor bound*, which find widespread use throughout analytic number theory and related areas. Recall that $\tau(n)$ denotes the number of divisors of the positive integer n.

Lemma 8.1.1 (Divisor bound). *We have $\tau(n) \ll n^{o(1)}$.*

Proof. Let $\varepsilon > 0$. Let the prime factorisation of n be $n = p_1^{a_1} \cdots p_k^{a_k}$, where $p_1 < p_2 < \cdots < p_k$. Then

$$\tau(n) = (a_1 + 1) \cdots (a_k + 1).$$

Let $p_{k'}$ be the largest prime factor of n less than $2^{1/\varepsilon}$. If $i > k'$ then

$$a_i + 1 \leqslant 2^{a_i} \leqslant (p_i^{a_i})^\varepsilon. \qquad (8.1)$$

Now since $\lim_{a\to\infty}\frac{a+1}{2^{a\varepsilon}}=0$, there is some constant $C(\varepsilon)$ so that

$$a_i+1\leqslant C(\varepsilon)(2^{a_i})^{\varepsilon}\leqslant C(\varepsilon)(p_i^{a_i})^{\varepsilon} \tag{8.2}$$

for *all* i. Applying (8.2) for $i\leqslant k'$, and (8.1) for $i>k'$, we have

$$\tau(n)\leqslant C(\varepsilon)^{2^{1/\varepsilon}}n^{\varepsilon}.$$

Since ε was arbitrary, the result follows. □

Remark. Being a little more careful, one can obtain more explicit bounds, the best possible (up to the $o(1)$ term) being

$$\tau(n)\ll\exp((\log 2+o(1))\frac{\log n}{\log\log n}).$$

We observe that this is not *that* small. It is certainly not bounded by a power of $\log n$. However the moments of $\tau(n)$ *are* logarithmically bounded, in fact

$$\frac{1}{N}\sum_{n\leqslant N}\tau(n)^k\ll(\log N)^{2^k-1},$$

and estimates of this type are often used in practice.

8.2. Hua's lemma: combinatorial bounds

A key ingredient in the proof of Hua's lemma is the following purely combinatorial fact and its corollary. They apply to any function $f:\{1,\dots,D\}\to\mathbf{C}$. In the next section, we will specialise to the case $f(x)=x^k$.

Lemma 8.2.1. *Let $f:\{1,\dots,D\}\to\mathbf{C}$ be a function. Let j,d be non-negative integers with $j<d$. Suppose that $t\in\mathbf{Z}$. Write $S_{j,d}(t)$ for the number of 2^d-tuples $(x_\omega)_{\omega\in\{0,1\}^d}$ such that*

- *We have*

$$\sum_{\omega}(-1)^{|\omega|}f(x_\omega)=t$$

  and
- *The "first" 2^j coordinates x_ω lie in a parallelepiped, that is to say there are $h_1,\dots,h_j$ such that*

$$x_\omega=x_0+\omega_1h_1+\cdots+\omega_jh_j$$

  for all ω such that $\omega_{j+1}=\cdots=\omega_d=0$.

Then

$$S_{j,d}(0)^2\leqslant(2D)^jS_{0,d}(0)S_{j+1,d}(0).$$

Remark. It may be observed that only the case $t = 0$ features in the conclusion of this lemma. However, it is convenient to introduce the slightly more general notation, both for the proof of Lemma 8.2.1, and for use later on.

Proof. Write $M_j(h_1, \dots, h_j; t)$ for the number of 2^{d-1}-tuples $(x_\omega)_{\omega \in \{0,1\}^{d-1}}$ such that

$$\sum_{\omega \in \{0,1\}^{d-1}} (-1)^{|\omega|}(-1)^{|\omega|} f(x_\omega) = t,$$

and for which the first 2^j coordinates lie in a paralleleipiped with sidelengths $h_1, \dots, h_j$. Then we have the following three identities (8.3), (8.4), (8.5). We give proofs in words; very carefully notated justifications are a bit tedious and left to the reader.

$$S_{j,d}(0) = \sum_{t,h_1,\dots,h_j} M_j(h_1, \dots, h_j; t) M_0(t) \tag{8.3}$$

(A 2^d-tuple with a j-parallelepiped whose ±1 sum over f is zero can be decomposed into two 2^{d-1} tuples, one containing a j-paralleleipiped, the other not. They have the same ±1-sum, t.)

$$S_{j+1,d}(0) = \sum_{t,h_1,\dots,h_j} M_j(h_1, \dots, h_j, t)^2 \tag{8.4}$$

(A $(j+1)$-parallelepiped is the same thing as a union of two j-parallelepipeds with the same sidelengths. Decompose a 2^d-tuple containing a $(j+1)$-parallelpiped into two 2^{d-1}-tuples, each containing j-parallelepipeds with the same sidelengths.)

$$S_{0,d}(0) = \sum_t M_0(t)^2. \tag{8.5}$$

(A 2^d-tuple whose ±1-sum over f is zero is the union of two 2^{d-1}-tuples, with the ±1-sum of f over each of them having the same value t. Alternatively, this follows immediately from either (8.3) or (8.4) with $j = 0$, noting in the latter case that $S_{1,d}(0) = S_{0,d}(0)$.)

By the Cauchy-Schwarz inequality,

$$\sum_{t,h_1,\dots,h_j} M_j(h_1, \dots, h_j; t) M_0(t) \leqslant (2D)^j \sum_{h_1,\dots,h_j} \big(\sum_t M_j(h_1, \dots, h_j; t) M_0(t)\big)^2.$$

By the Cauchy-Schwarz inequality again,

$$\big(\sum_t M_j(h_1, \dots, h_j; t) M_0(t)\big)^2 \leqslant \big(\sum_t M_j(h_1, \dots, h_j; t)^2\big)\big(\sum_t M_0(t)^2\big).$$

Combining these inequalities with (8.3), (8.4), (8.5) gives the result. □

Corollary 8.2.1. *With notation as in Lemma 8.2.1,*

$$S_{0,d}(0) \ll D^{2^j-j-1} S_{j,d}(0). \tag{8.6}$$

In particular,

$$S_{0,m+1}(0) \ll D^{2^m-m-1} S_{m,m+1}(0). \tag{8.7}$$

Proof. By induction on j, the result being clear when $j = 0$ or 1. For the general case, apply the induction hypothesis and Lemma 8.2.1 and we obtain

$$\begin{aligned} S_{0,d}(0)^2 &\ll D^{2(2^j-j-1)} S_{j,d}(0)^2 \\ &\ll D^j D^{2(2^j-j-1)} S_{0,d}(0) S_{j+1}(0). \end{aligned}$$

Dividing through by $S_{0,d}(0)$ gives the case $j + 1$ of (8.6), so the inductive step is complete.

The second inequality, (8.7), is simply the special case $d = m + 1$, $j = m$ of the first. □

8.3. Hua's lemma: arithmetic input

We have gone as far as we can for general functions f, and we now specialise to the case $f(x) = x^k$. In this case, we can supplement Lemma 8.2.1 with the following bounds.

Lemma 8.3.1. *Let $f(x) = x^k$, and let all other notation be as in Lemma 8.2.1. Let $m \leqslant k$. Then*

$$S_{m,m}(t) \ll D^{o(1)}, \tag{8.8}$$

uniformly for $|t| \leqslant 2^k D$, $t \neq 0$. Additionally,

$$S_{m,m}(0) \ll D^{m+o(1)}. \tag{8.9}$$

Proof. Both parts rely on the fact that

$$\sum_{\omega \in \{0,1\}^m} (-1)^{|\omega|} f(x + \omega_1 h_1 + \cdots + \omega_m h_m) = h_1 \cdots h_m p(x; h_1, \dots, h_m), \tag{8.10}$$

where p is a polynomial of degree $k - m$ for each fixed choice of $h_1, \dots, h_m$. This follows from m applications of the fact that if P is a polynomial of degree d then its derivative $\partial_h P(x) := P(x) - P(x + h)$ is a polynomial of degree $d - 1$ (which may depend on h). The reader will note that we used exactly the same fact in the proof of Weyl's inequality.

For (8.8), we use the divisor bound. If the RHS of (8.10) is equal to t then $h_1, \dots, h_m$ are all divisors of t, which means there are $D^{o(1)}$ choices for each of them. The value of $p(x; h_1, \dots, h_m)$ is then fixed, which means that there are at most $\deg p = k - m$ choices for x.

For (8.9), if the RHS of (8.10) is zero then either one of the h_is is zero (giving $\ll D^m$ choices for x and the other h_js) or $p(x; h_1, \dots, h_m) = 0$, giving $\deg p = k-m$ choices for x, for each of the $\ll D^m$ choices of $h_1, \dots, h_m$. □

We are now ready to prove Hua's lemma itself.

Proof. [Proof of Theorem 8.0.1] In the notation of Lemma 8.2.1, what we need to show is that, with $f(x) = x^k$,

$$S_{0,k}(0) \ll D^{2^k - k + o(1)}. \tag{8.11}$$

To do this, we prove that in fact

$$S_{0,m}(0) \ll D^{2^m - m + o(1)} \tag{8.12}$$

for $1 \leqslant m \leqslant k$, the case $m = k$ being the one we are interested in.

We do this by induction on m, the case $m = 1$ being obvious. For the inductive step, first observe that

$$\begin{aligned} S_{m,m+1}(0) &= \sum_t S_{m,m}(t) S_{0,m}(t) \quad \text{(follows from the definition)} \\ &= S_{m,m}(0) S_{0,m}(0) + \sum_{t \neq 0} S_{m,m}(t) S_{0,m}(t) \\ &\ll D^{m+o(1)} \cdot D^{2^m - m + o(1)} + D^{o(1)} \cdot D^{2^m} \\ &\ll D^{2^m + o(1)}. \end{aligned}$$

In the penultimate step, we used four bounds: (8.9), the inductive hypothesis (8.11), (8.8), and finally

$$\sum_t S_{0,m}(t) = D^{2^m},$$

which is clear from the definitions.

To complete the proof, we apply (8.7), obtaining

$$S_{0,m+1}(0) \ll D^{2^m - m - 1} S_{m,m+1}(0) \ll D^{2^m - m - 1} \cdot D^{2^m + o(1)} = D^{2^{m+1} - m - 1 + o(1)}.$$

This completes the inductive step in the proof of (8.12). Therefore (8.12) holds for all $m \leqslant k$, and in particular for $m = k$ which, as remarked in (8.11), is equivalent to Hua's Lemma. □

8.4. Consequences for Waring's problem

We use the notation of Section 3. Hua's lemma is equivalent to the bound

$$\int_{\mathbf{T}} |\hat{1}_X(\theta)|^{2^k} d\theta \ll N^{\frac{2^k}{k}-1+o(1)}. \tag{8.13}$$

Indeed, the left-hand side may be expanded out using Fourier analysis (Parseval and the formula for convolution) and it equals exactly the number of tuples counted in Hua's Lemma, with $D = N^{1/k}$.

Using this, one may quickly obtain the minor arcs bound Proposition 3.2.2 under the much tighter condition $s \geqslant 2^k + 1$.

PROPOSITION 8.4.1. *Let notation by as in Section 3. Suppose that $s \geqslant 2^k + 1$. Then*

$$\int_{\mathfrak{m}} \hat{1}_X(\theta)^s e(N\theta) d\theta = o(N^{s/k-1}).$$

Proof. Recall that in Proposition 4.0.1 we obtained the pointwise estimate

$$\sup_{\theta \in \mathfrak{m}} |\hat{1}_X(\theta)| \ll N^{1/k-\varepsilon}.$$

We showed that $\varepsilon = (100)^{-k}$ was acceptable, but the precise value is no longer relevant.

It follows from this and Hua's lemma that

$$\begin{aligned}\int_{\mathfrak{m}} \hat{1}_X(\theta)^s e(N\theta) d\theta &\leqslant \int_{\mathfrak{m}} |\hat{1}_X(\theta)|^s d\theta \\ &\leqslant \sup_{\theta \in \mathfrak{m}} |\hat{1}_X(\theta)|^{s-2^k} \int_{\mathbf{T}} |\hat{1}_X(\theta)|^{2^k} d\theta \\ &\ll N^{(s-2^k)(\frac{1}{k}-\varepsilon)} \cdot N^{\frac{2^k}{k}-1+o(1)} \ll N^{\frac{s}{k}-1-\frac{\varepsilon}{2}}.\end{aligned}$$

This concludes the proof.

□

The whole of the analysis of the major arcs only requires the much weaker condition $s \geqslant 2k+1$. Therefore the asymptotic formula (3.1.2) is valid for $s \geqslant 2^k+1$.

We have shown that the singular series $\mathfrak{S}_{k,s}(N)$ is $\asymp 1$ for $s \geqslant k^4$, and hence for $s \geqslant 2^k + 1$ when $k \geqslant 4$. It turns out that the same conclusion is true for $k = 2$ and $k = 3$. This requires a little calculation and a couple of further lemmas, and may be found on Sheet 2.

In particular, with all of these results in place one has $G(k) \leqslant 2^k + 1$.

Part 2

Additive Combinatorics

CHAPTER 9

Roth's theorem on progressions of length 3

In this chapter our aim is to prove the following theorem of Roth from 1953.

Theorem 9.0.1 (Roth's theorem). *There is an absolute constant C such that any subset $A \subset \{1, \dots, N\}$ with cardinality at least $CN/\log\log N$ contains a nontrivial three-term arithmetic progression (that is to say, a triple $x, x+d, x+2d$ with $d \neq 0$).*

Note, in particular, that $1/\log\log N$ is eventually smaller than any fixed positive constant.

Throughout this chapter we will assume that N is sufficiently large (meaning bigger than some absolute constant which we shall not specify precisely).

9.1. The density increment strategy

Roth's theorem proceeds via the so-called *density increment strategy,* and the key proposition which drives this is the following.

Proposition 9.1.1. *Suppose that $0 < \alpha < 1$ and that $N \geqslant (8/\alpha)^{10}$. Suppose that $P \subset \mathbf{Z}$ is an arithmetic progression of length N and that $A \subset P$ is a set with cardinality at least αN. Then one of the following two alternatives holds:*

(i) *A contains a nontrivial 3-term progression;*
(ii) *There is an arithmetic progression P' of length $N' \geqslant N^{1/5}$ such that, writing $A' := A \cap P'$ and $\alpha' := |A'|/|P'|$, we have $\alpha' \geqslant \alpha + \frac{\alpha^2}{112}$.*

Theorem 9.0.1 follows by iterating this proposition. Set $P_0 := \{1, \dots, N\}$ and let us suppose that we have a set $A \subset P_0$ with $|A| = \alpha N$ and containing no nontrivial 3-term progression. Then we attempt to use Proposition 9.1.1 repeatedly to obtain a sequence $P_0, P_1, P_2, \dots$ of progressions together with sets $A_i := A \cap P_i$. The length of P_i will be $N_i \geqslant N^{(1/5)^i}$ and the densities $\alpha_i := |A_i|/|P_i|$ will satisfy $\alpha_{i+1} > \alpha_i + c\alpha_i^2$.

Now this iteration cannot last too long: after C/α steps the density has already doubled, after a further $C/2\alpha$ steps it has doubled again, and so on. Since no set can have density greater than one, there can be no more than $2C/\alpha$ steps in total. We conclude that our applications of Proposition 9.1.1 must have been invalid, which can ony mean that the condition $N_i > C\alpha_i^{-C}$ was violated. Since

$$N_i > N^{(1/5)^i} \geqslant N^{(1/5)^{2C/\alpha}}$$

and (very crudely)

$$\alpha_i \geqslant \alpha,$$

we infer the bound

$$N^{(1/5)^{2C/\alpha}} \leqslant C\alpha^{-C}.$$

Rearranging gives

$$\log\log N \leqslant \log\log(C\alpha^{-C}) + \frac{2C}{\alpha} \leqslant \frac{C'}{\alpha},$$

which immediately gives the claimed bound.

Remark. The most important parameter by far is the number of times we performed the iteration, which was roughly $O(1/\alpha)$.

9.2. A large Fourier coefficient

We turn now to the details of the density increment strategy. We begin with a very simple observation, which is that we may assume without loss of generality that $P = [N] = \{1, \dots, N\}$. We may always reduce to this case by an affine rescaling.

We will first establish the following alternative version of Proposition 9.1.1, in which the conclusion of part (ii) is different, asserting the existence of a large Fourier coefficient of the function

$$f_A := 1_A - \alpha 1_{[N]},$$

the so-called *balanced function* of A. In the next section, we will show that a large Fourier coefficient implies a density increment as in the original formulation of Proposition 9.1.1.

PROPOSITION 9.2.1. *Suppose that $0 < \alpha < 1$ and that $N \geqslant 4/\alpha^2$. Suppose that $A \subset [N]$ is a set with cardinality at least αN. Then one of the following two alternatives holds:*

(i) *A contains a nontrivial 3-term progression;*
(ii) *The balanced function f_A has a large Fourier coefficient: specifically, there is some $\theta \in \mathbf{T}$ such that $|\hat{f}_A(\theta)| \geqslant \alpha^2 N/28$.*

Proof. If $f_1, f_2, f_3 : \mathbf{Z} \to \mathbf{R}$ are three finitely-supported functions then we introduce the operator

$$T(f_1, f_2, f_3) := \sum_{x,d} f_1(x) f_2(x+d) f_3(x+2d).$$

This counts the number of 3-term progressions weighted by the functions f_i. In particular,

$$T(1_A, 1_A, 1_A) = \#\{\text{number of 3-term progressions in } A\}. \tag{9.1}$$

Note carefully that this count includes "trivial" progressions with $d = 0$. However, A has precisely αN trivial progressions, so if option (i) does not hold then

$$T(1_A, 1_A, 1_A) = \alpha N \leqslant \alpha^3 N^2/4. \tag{9.2}$$

For the inequality on the right we used the assumption that $N \geqslant 4/\alpha^2$.

Note that T is a trilinear operator. Thus we may write $1_A = f_A + \alpha 1_{[N]}$ and expand $T(1_A, 1_A, 1_A)$ as a sum of eight terms,

$$T(1_A, 1_A, 1_A) = \alpha^3 T(1_{[N]}, 1_{[N]}, 1_{[N]}) + \cdots + T(f_A, f_A, f_A). \tag{9.3}$$

Each of the seven "error terms" denoted by the ellipsis $\cdots$ contains at least one copy of f_A. Let us look at the first term $\alpha^3 T(1_{[N]}, 1_{[N]}, 1_{[N]})$. It is quite simple to evaluate this exactly: the number of (x, d) with $x, x + d, x + 2d \in [N]$ is precisely the number of pairs $(n_1, n_2) \in [N] \times [N]$ with n_1, n_2 having the same parity, since we then have, uniquely, $x = n_1$ and $d = \frac{1}{2}(n_2 - n_1)$, and $x + d$ automatically lies in $[N]$. This is $N^2/2$ if N is even, and $(N^2 + 1)/2$ if N is odd, thus at least $N^2/2$ in all cases. Thus

$$\alpha^3 T(1_{[N]}, 1_{[N]}, 1_{[N]}) \geqslant \alpha^3 N^2/2.$$

It follows that if option (i) does not hold (and hence we have (9.2)) then the sum of the seven error terms in (9.3) is at least $\alpha^3 N^2/4$. Thus one of these terms is at least $\alpha^3 N^2/28$, that is to say

$$|T(f_1, f_2, f_3)| \geqslant \alpha^3 N^2/28, \tag{9.4}$$

where each f_i is either $\alpha 1_{[N]}$ or f_A, and at least one of them is f_A.

Now we come to the key idea: there is a formula for $T(f_1, f_2, f_3)$ in terms of the Fourier transform:

$$T(f_1, f_2, f_3) = \int_{\mathbf{T}} \hat{f}_1(\theta)\hat{f}_2(-2\theta)\hat{f}_3(\theta)d\theta. \tag{9.5}$$

Once written down, it is very easy to check this by substituting the definition of the Fourier transforms on the right-hand side.

Thus if (9.4) holds then

$$\Big|\int_{\mathbf{T}} \hat{f}_1(\theta)\hat{f}_2(-2\theta)\hat{f}_3(\theta)d\theta\Big| \geqslant \alpha^3 N^2/28. \tag{9.6}$$

Suppose that $f_3 = f_A$; the analysis of other possibilities is very similar. Then

$$\sup_{\theta \in \mathbf{T}} |\hat{f_A}(\theta)| \int_{\mathbf{T}} |\hat{f}_1(\theta)||\hat{f}_2(-2\theta)|d\theta \geqslant \alpha^3 N^2/28.$$

By the Cauchy-Schwarz inequality,

$$\sup_{\theta \in \mathbf{T}} |\hat{f_A}(\theta)| \Big(\int_{\mathbf{T}} |\hat{f}_1(\theta)|^2 d\theta \Big)^{1/2} \Big(\int_{\mathbf{T}} |\hat{f}_2(\theta)| d\theta \Big)^{1/2} \geqslant \alpha^3 N^2/28. \tag{9.7}$$

However, by Parseval's identity we have

$$\int_{\mathbf{T}} |f_i(\theta)|^2 = \sum_n |f_i(n)|^2.$$

One may easily check that the RHS is $\alpha^2 N$ if $f_i = \alpha 1_{[N]}$ and $\alpha(1-\alpha)N$ if $f_i = f_A$, and so certainly at most αN in either case. Thus from (9.7) we obtain

$$\sup_{\theta \in \mathbf{T}} |\hat{f}_A(\theta)| \geqslant \alpha^2 N/28,$$

which is precisely option (ii) in the proposition. □

Remarks. The above proof depended crucially on "observing" the Fourier identity (9.5). One could easily create this identity starting from the definition of $T(f_1, f_2, f_3)$ by writing each f_i using the Fourier inversion formula. That Fourier analysis should be useful in this problem could perhaps be suggested by the observation that $T(f_1, f_2, f_3) = f_1 * g * f_3(0)$, where $g(-2x) = f_2(x)$, and g vanishes on odd numbers.

9.3. From a large Fourier coefficient to a density increment

In this section, we show how option (ii) in Proposition 9.2.1 (the balanced function f_A has a large Fourier coefficient) may be replaced by option (ii) in Proposition 9.1.1 (a density increment on a progression). The crucial technical ingredient is the following.

Here, if $F : \mathbf{Z} \to \mathbf{C}$ is a function and $S \subset \mathbf{Z}$ a finite set, we write $\operatorname{diam}_S(F) := \sup_{x,x' \in S} |F(x) - F(x')|$.

Lemma 9.3.1. *Suppose that $\theta \in \mathbf{T}$. Then we may partition $[N]$ into progressions P_i, each of length at least $N^{1/5}$, such that* $\operatorname{diam}_{P_i}(e(\theta x)) \leqslant N^{-1/5}$ *for all i.*

Proof. Throughout this argument we will assume that N is sufficiently large. Let $Q := \lfloor N^{1/2} \rfloor$. By a well-known application of the pigeonhole principle due to Dirichlet, there is some positive $d \leqslant Q$ such that $\|d\theta\| \leqslant 1/Q$. (Consider $\theta, 2\theta, \cdots, Q\theta$ as elements of $\mathbf{T}$; some two of these, say $j_1\theta$ and $j_2\theta$, lie within $1/Q$ of one another. Take $d := |j_1 - j_2|$.)

If P is any progression with common difference d and length $\leqslant 3N^{1/5}$ then, by the triangle inequality,

$$\operatorname{diam}_P(e(\theta x)) \leqslant 3N^{1/5}|e(\theta d) - 1| \leqslant 20N^{1/5}/Q < N^{-1/5},$$

where here we used the inequality

$$|e(t)-1| = 2|\sin \pi t| \leqslant 2\pi \|t\|_{\mathbf{R}/\mathbf{Z}}.$$

Now observe that $[N]$ can be partitioned into progressions P_i with common difference d and lengths in the range $[N^{1/5}, 3N^{1/5}]$. To do this, first partition $[N]$ into progressions of common difference d, each of length $\sim N/d \gg N^{1/2}$. Then proceed along each such progression from left to right, partitioning into progressions of length $\lceil N^{1/5} \rceil$ until we have a leftover progression of length $\leqslant N^{1/5}$. Amalgamate this with the preceding one. □

The following result, together with Proposition 9.2.1, immediately implies Proposition 9.1.1, and hence completes the proof of Roth's theorem.

PROPOSITION 9.3.1. *Suppose that* $|\hat{f}_A(\theta)| \geqslant \alpha^2 N/28$*, that* $N \geqslant (8/\alpha)^{10}$*, and let* $[N] = \bigcup_i P_i$ *be a partition as above. Then there is some* i *such that* $|A \cap P_i| \geqslant (\alpha + \frac{\alpha^2}{112})|P_i|$.

Proof. Since the P_i partition $[N]$, we obviously have

$$\sum_i |\sum_{x\in P_i} f_A(x) e(-\theta x)| \geqslant \frac{\alpha^2}{28} N.$$

By the triangle inequality and the bound $|f_A(x)| \leqslant 1$, the left-hand side is at most

$$\begin{aligned}\sum_i |\sum_{x\in P_i} f_A(x)| + \sum_i |P_i| \operatorname{diam}_{P_i}(e(\theta x)) &\leqslant \sum_i |\sum_{x\in P_i} f_A(x)| + N^{4/5} \\ &\leqslant \sum_i |\sum_{x\in P_i} f_A(x)| + \frac{\alpha^2}{56} N,\end{aligned}$$

the last step following from our assumption on N. It follows that

$$\sum_i |\sum_{x\in P_i} f_A(x)| \geqslant \frac{\alpha^2}{56} N.$$

Since $\sum_{x\in[N]} f_A(x) = 0$, we have

$$\sum_i \big(|\sum_{x\in P_i} f_A(x)| + \sum_{x\in P_i} f_A(x)\big) \geqslant \frac{\alpha^2}{56} N = \frac{\alpha^2}{56} \sum_i |P_i|,$$

so there must be some i such that

$$|\sum_{x\in P_i} f_A(x)| + \sum_{x\in P_i} f_A(x) \geqslant \frac{\alpha^2}{56} |P_i|,$$

which implies that

$$\sum_{x\in P_i} f_A(x) \geqslant \frac{\alpha^2}{112} |P_i|,$$

or in other words that

$$|A \cap P_i| \geqslant (\alpha + \frac{\alpha^2}{112})|P_i|.$$

This concludes the proof.

□

CHAPTER 10

Sumsets

Recall that if A is a set of integers then

$$A + A := \{a_1 + a_2 : a_1, a_2 \in A\}.$$

10.1. The cardinality of sumsets and Freiman's theorem

Suppose A has size n. How big is $A+A$? Trivially, it has size at most $\frac{1}{2}n(n+1)$, that being the number of pairs (a_1, a_2), with (a_1, a_2) and (a_2, a_1) counted the same.

On the other hand, it has size at least $2n - 1$. Writing $a_1 < \dots < a_n$ for the elements of A, we have

$$a_1 + a_1 < a_1 + a_2 < \dots < a_1 + a_n < a_2 + a_n < \dots < a_n + a_n,$$

a listing of $2n - 1$ distinct elements of A.

Equality can occur in both bounds. For example if $A = \{1, 2, \dots, 2^{n-1}\}$ then all the sums $a_1 + a_2$ are distinct (except for the trivial relations $a_1 + a_2 = a_2 + a_1$). If $A = \{1, \dots, n\}$ then $A + A = \{2, \dots, 2n\}$, a set of size $2n - 1$.

A highlight of the rest of the course is a theorem of Freiman, which gives an answer to the following question.

Question 10.1.1. What is the structure of A if $|A + A| \leqslant K|A|$?

We say that A has *doubling constant at most* K. Typically, we will have in mind that K is fixed (say $K = 10$) and $n = |A|$ is very large.

Before stating the theorem, let us give some progressively more complicated motivating examples.

Example 10.1.1 (Progression). Let A be any arithmetic progression of length n. Then $|A + A| = 2n - 1$.

Example 10.1.2 (Subsets of progressions). Let P be a progression of length Cn, and let $A \subset P$ be an arbitrary set of size n. Then $|A + A| \leqslant 2Cn$.

Example 10.1.3 (2-dimensional progression). Suppose that $L_1 L_2 = n$, and consider a set A of the form

$$A := \{x_0 + \ell_1 x_1 + \ell_2 x_2 : 0 \leqslant \ell_1 < L_1, 0 \leqslant \ell_2 < L_2\}.$$

If the x_i are suitably widely spaced, the elements described here are all distinct and $|A| = n$. In this case we say that A is *proper*. We have

$$A + A = \{2x_0 + \ell_1' x_1 + \ell_2' x_2 : 0 \leqslant \ell_1' < 2L_1 - 1, 0 \leqslant \ell_2' < 2L_2 - 1\},$$

and so certainly

$$|A + A| \leqslant 4|A|.$$

Example 10.1.4 (d-dimensional progression). The same as above, but with d parameters $L_1, \dots, L_d$: thus

$$A = \{x_0 + l_1 x_1 + \cdots + l_d x_d : 0 \leqslant l_i < L_i\}. \tag{10.1}$$

Now we have $|A + A| \leqslant 2^d |A|$.

Example 10.1.5 (Subsets of multidimensional progressions). Let P be a proper d-dimensional progression of size Cn. Let $A \subset P$ be an arbitrary set of size n. Then

$$|A + A| \leqslant |P + P| \leqslant 2^d |P| = 2^d Cn.$$

The final example gives a somewhat large class of sets with doubling constant at most K (pick any parameters d, C with $2^d C \leqslant K$).

Freiman's theorem is the result that the above examples are the only ones.

Theorem 10.1.1 (Freiman). *Suppose that $A \subset \mathbf{Z}$ is a finite set with $|A + A| \leqslant K|A|$. Then A is contained in a generalised progression P of dimension $\ll_K 1$ and size $\ll_K |A|$.*

The *size* of a generalised progression as in (10.1) is defined to be $L_1 \cdots L_d$. This is at least the cardinality of the progression, but is strictly bigger than it if the progression fails to be proper.

Freiman's theorem states that A is contained in a proper progression of dimension at most $d(K)$ and size at most $C(K)|A|$, where $d(), C()$ are functions of K only. In this course we will not be concerned with bounds, but the argument we give leads to a bound for $d(K)$ that is exponential in K, and a bound for $C(K)$ that is doubly exponential in K. This is quite far from the truth; in fact, it does not require a vast amount of further effort to remove an exponential from both of these bounds, but we will not do so here.

Many other refinements are possible, but again we will not cover them here. For example, one can insist that P be proper if desired.

10.2. Ruzsa's triangle inequality and covering lemma

In the proof of Freiman's theorem, we will need some estimates for the size of sumsets. There is a huge literature on this topic, from which we isolate a few key results. All of the results we shall state are valid for finite subsets of arbitrary abelian groups, and for brevity it is usual to call these "additive sets". In fact, many of the results (but not all) remain true without the assumption of commutativity, but we shall not cover that topic in this course.

In this section we prove two elegant results of Ruzsa, which are surprisingly useful despite their apparent simplicity.

Lemma 10.2.1 (Ruzsa triangle inequality). *Suppose that U, V, W are finite additive sets. Then*

$$|V - W||U| \leqslant |V - U||U - W|.$$

Proof. We will define a map $\phi : (V - W) \times U \to (V - U) \times (U - W)$, and prove that it is an injection, which implies the result. Given $d \in V - W$ select a pair $v_d \in V, w_d \in W$ for which $d = v_d - w_d$ (there may be more than one such pair, but for each d we make a definite choice). Then define

$$\phi(d, u) = (v_d - u, u - w_d)$$

for each $d \in V - W$ and $u \in U$. To prove that ϕ is an injection, suppose that $(x, y) \in \mathrm{im}(\phi) \subset (V - U) \times (U - W)$. If $\phi(d, u) = (x, y)$ then $x + y = (v_d - u) + (u - w_d) = v_d - w_d = d$, and therefore we can determine d and hence v_d and w_d from (x, y). And we also determine u as $u = -x + v_d$ $(= y - w_d)$. □

Remark. If we define

$$d(U, V) := \log \frac{|U - V|}{|U|^{1/2}|V|^{1/2}}$$

then the Ruzsa triangle inequality may be written

$$d(V, W) \leqslant d(U, V) + d(U, W).$$

This explains the term "triangle inequality". Note that, although the triangle inequality is satisfied, d is not a true distance. This is because $d(U, V) = 0$ neither implies, nor is implied by, $U = V$.

Lemma 10.2.2 (Ruzsa's covering lemma). *Suppose that A and B are finite additive sets and that $|A + B| \leqslant K|A|$. Then B may be covered by k translates of $A - A$, for some $k \leqslant K$. That is, there is a set X, $|X| \leqslant K$, such that*

$$B \subset (A - A) + X.$$

Proof. Choose $X \subset B$ maximal so that $\{A + x : x \in X\}$ are disjoint. The union of these sets contains exactly $|A||X|$ elements, and all of these elements lie in $A + B$. Therefore $|X| \leqslant K$. Now, if $b \in B$ then $A + b$ intersects $A + x$ for some $x \in X$, because of the maximality of X, and so $b \in A - A + x$. Hence, $B \subset (A - A) + X$. □

10.3. Petridis's inequality

In this section and the next we develop inequalities controlling the size of sums of three or more sets. A beautiful way to do this was discovered surprisingly recently by Petridis. His result is stated as Corollary 10.3.1 below. We give an elegant rephrasing of his proof which was given by Tao on the blog of Tim Gowers.

Let B be a set in some abelian group G. Let K be a real number, and consider the function ϕ on subsets of G defined by

$$\phi(A) := |A+B| - K|A|. \tag{10.2}$$

LEMMA 10.3.1. *ϕ is* submodular*, that is to say it satisfies*

$$\phi(A \cup A') + \phi(A \cap A') \leqslant \phi(A) + \phi(A').$$

Proof. Write $\sigma(A) := A + B$. Observe that

$$\sigma(A \cap A') = \sigma(A) \cap \sigma(A'),$$

and that

$$\sigma(A \cap A') \subseteq \sigma(A) \cap \sigma(A').$$

Therefore

$$\begin{aligned} |\sigma(A) \cup \sigma(A')| &= |\sigma(A)| + |\sigma(A')| - |\sigma(A) \cap \sigma(A')| \\ &\leqslant |\sigma(A)| + |\sigma(A')| - |\sigma(A \cap A')|, \end{aligned}$$

that is to say $|\sigma|$ satisfies the submodularity property

$$|\sigma(A) \cap \sigma(A')| + |\sigma(A) \cap \sigma(A')| \leqslant |\sigma(A)| + |\sigma(A')|.$$

Since the function $|A|$ satisfies

$$|A \cup A'| + |A \cap A'| = |A| + |A'|,$$

the result follows immediately. □

LEMMA 10.3.2. *Let ϕ be any submodular function. Suppose that $A_1, \dots, A_n$ are sets with the following property: $\phi(A_i) = 0$, and $\phi(Z_i) \geqslant 0$ for every subset $Z_i \subseteq A_i$. Then $\phi\big(\bigcup_{i=1}^n A_i\big) \leqslant 0$.*

Proof. By the assumptions and submodularity, for any i and for any set S, we have

$$\phi(A_i \cup S) \leqslant \phi(A_i \cup S) + \phi(A_i \cap S) \leqslant \phi(A_i) + \phi(S) = \phi(S).$$

The result then follows immediately by induction on n. □

Proposition 10.3.1 (Petridis). *Let A, B be sets in some abelian group. Suppose that $|A+B| = K|A|$ and that $|Z+B| \geqslant K|Z|$ for all $Z \subseteq A$. Then, for any further set S in the group, $|A+B+S| \leqslant K|A+S|$.*

Proof. Apply Lemma 10.3.2 with the particular function ϕ defined in (10.2) above. Take the A_i to be the translates $A+s$ of A by elements of s. It is easy to check that the hypotheses of Lemma 10.3.2 hold. Observe that $\bigcup_{i=1}^n A_i = A+S$, and so the Lemma implies that $\phi(A+S) \leqslant 0$, or in other words $|A+B+S| \leqslant K|A+S|$. □

It is convenient to apply Petridis' inequality in the following form.

Corollary 10.3.1. *Let A, B be sets in some abelian group. Suppose that $|A+B| \leqslant K|A|$. Let $X \subseteq A$ be a non-empty set for which the ratio $|X+B|/|X|$ is minimal. Then for any further set S we have*

$$|S+X+B| \leqslant K|S+X|.$$

Proof. Apply Proposition 10.3.1 with A replaced by X. □

10.4. The Plünnecke–Ruzsa inequality

The most widely applicable result about higher-order sumsets is the Plünnecke–Ruzsa inequality.

Theorem 10.4.1 (Plünnecke–Ruzsa). *Suppose that A and B are additive sets with $|A+B| \leqslant K|A|$. Let $k, \ell \geqslant 0$ be integers. Then $|kB - \ell B| \leqslant K^{k+\ell}|A|$.*

The original proof was quite long and involved a fair amount of machinery from graph theory. Nowadays, it can be deduced quickly from Petridis's inequality.

Lemma 10.4.1. *Suppose that A and B are finite additive sets for which $|A+B| \leqslant K|A|$. Then there exists $X \subset A$ for which $|X+kB| \leqslant K^k|X|$.*

Proof. Let X be the subset of A for which the ratio $|X+B|/|X|$ is minimal. By Petridis's inequality (Corollary 10.3.1) with $S = (k-1)B$, we have

$$|X+kB| = |X+(k-1)B+B| \leqslant K|X+(k-1)B|.$$

The result then follows by induction on k. □

Proof. [Proof of Theorem 10.4.1]. Suppose that A and B are finite additive sets for which $|A+B| \leqslant K|A|$. By Ruzsa's Triangle Inequality with U, V, W replaced by $X, -kB, -\ell B$, respectively, and then Lemma 10.4.1, we have

$$|kB - \ell B|\ |X| \leqslant |X+kB| \cdot |X+\ell B| \leqslant K^{k+\ell}|X|^2.$$

Thus, since $X \subset A$, $|kB - \ell B| \leqslant K^{k+\ell}|X| \leqslant K^{k+\ell}|A|$. □

CHAPTER 11

Freiman homomorphisms and Ruzsa's model lemma

11.1. Freiman homomorphisms

In his remarkably insightful 1966 book [**5**], Freiman made an attempt to treat additive number theory by analogy with the way Klein treated geometry: as well as sets $A, B, \cdots$ of integers, one should study maps between them and, most particularly, properties invariant under natural types of map. This was doubtless regarded as somewhat eccentric at the time, but the notion of Freiman homomorphism is now quite important in additive combinatorics.

Definition 11.1.1. Suppose that $s \geqslant 2$ is an integer. Suppose that A, B are additive sets. Then we say that a map $\phi : A \to B$ is a Freiman s-homomorphism if we have

$$\phi(a_1) + \cdots + \phi(a_s) = \phi(a_1') + \cdots + \phi(a_s')$$

whenever

$$a_1 + \cdots + a_s = a_1' + \cdots + a_s'.$$

It is obvious that any group homomorphism restricts to a Freiman homomorphism (of arbitrary order) on any subset. However, the notion is much more general. For example, any map whatsoever from $A = \{1, 10, 100, 1000\}$ to another additive set is a Freiman 2-homomorphism, simply because A has no nontrivial relations of the form $a_1 + a_2 = a_1' + a_2'$.

The map ϕ is said to be a Freiman s-isomorphism if it has an inverse ϕ^{-1} which is also a Freiman s-homomorphism. We caution that, contrary to what is often expected in more algebraic situations, a one-to-one Freiman homomorphism need not be a Freiman isomorphism. For example, the obvious map

$$\phi : \{0, 1\}^n \to (\mathbf{Z}/2\mathbf{Z})^n$$

is a Freiman homomorphism of all orders (it is induced from the natural group homomorphism $\mathbf{Z}^n \to (\mathbf{Z}/2\mathbf{Z})^n$). However, it is not a Freiman 2-isomorphism as $(\mathbf{Z}/2\mathbf{Z})^n$ contains a great many more additive relations than $\{0, 1\}^n$.

The following lemma records some basic facts about Freiman isomorphisms.

Lemma 11.1.1. *Suppose that A, B, C are additive sets. Let $s \geqslant 2$ be an integer. Then we have the following.*

(i) *Suppose that $\phi : A \to B$ and $\psi : B \to C$ are Freiman s-homomorphisms. Then so is the composition $\psi \circ \phi$.*

(ii) *Suppose that $\phi : A \to B$ is a Freiman s-homomorphism. Then it is also a Freiman s'-homomorphism for every s' satisfying $2 \leqslant s' \leqslant s$.*

(iii) *Suppose that $\phi : A \to B$ is a Freiman s-homomorphism and let $k, l \geqslant 0$ be integers. Then ϕ induces a Freiman s'-homomorphism $\tilde{}:kA - lA \to kB - lB$, for any integer $s' \leqslant s/(k+l)$.*

(iv) *The above three statements also hold with "homo" replaced by "iso" throughout.*

(v) *Suppose that P is a generalised progression and that $\phi : P \to B$ is a Freiman 2-homomorphism. Then $\phi(P)$ is a generalised progression of the same dimension. If ϕ is a Freiman 2-isomorphism, and if P is proper, then so is $\phi(P)$.*

(vi) *Let $\pi_m : \mathbf{Z} \to \mathbf{Z}/m\mathbf{Z}$ be the natural map. Then π_m is a Freiman s-isomorphism when restricted to $(t, t + \frac{m}{s}] \cap \mathbf{Z}$, for any $t \in \mathbf{R}$.*

Proof. The first four parts of this are very straightforward once one has understood the definitions, and we will not go over them carefully in lectures. Perhaps (iii) requires some further comment: one should define $\tilde{\phi} : kA - lA \to kB - lB$ by

$$\tilde{\phi}(a_1 + \cdots + a_k - a'_1 - \cdots - a'_l) = \phi(a_1) + \cdots + \phi(a_k) - \phi(a'_1) - \cdots - \phi(a'_l).$$

One must then check that this is well-defined and is a Freiman homomorphism of the order claimed.

To prove (v), let $\phi : P \to \phi(P)$ be a Freiman 2-homomorphism. Suppose that $P = \{x_0 + l_1x_1 + \cdots + l_dx_d : 0 \leqslant l_i < L_i\}$. Set $y_0 = \phi(x_0)$, and define $y_1, \dots, y_d$ by $y_0 + y_i = \phi(x_0 + x_i)$ for $i = 0, 1, \dots, d$; we claim that $\phi(x_0 + l_1x_1 + \cdots + l_dx_d) = y_0 + l_1y_1 + \cdots + l_dy_d$ for all $l_1, \dots, l_d$ satisfying $0 \leqslant l_i < L_i$. This may be established by induction on $l_1 + \cdots + l_d$, noting that we have defined the y_i in such a way that it holds whenever $l_1 + \cdots + l_d = 0$ or 1. To obtain the statement for $(l_1, \dots, l_d) = (1, 1, 0, \dots, 0)$, for example, one may use the relation

$$x_0 + (x_0 + x_1 + x_2) = (x_0 + x_1) + (x_0 + x_2)$$

to conclude that

$$\phi(x_0) + \phi(x_0 + x_1 + x_2) = \phi(x_0 + x_1) + \phi(x_0 + x_2)$$

and hence that $\phi(x_0 + x_1 + x_2) = y_0 + y_1 + y_2$, as required.

Finally, we comment on (vi). Since π_m is a group homomorphism, it is also a Freiman homomorphism. Its restriction to any interval of length at most m is a bijection. Suppose that $x_1, \ldots, x_s, x'_1, \ldots, x'_s$ satisfy $t < x_i, x'_i \leqslant t + \frac{m}{s}$ and that $\pi_m(x_1) + \cdots + \pi_m(x_s) = \pi_m(x'_1) + \cdots + \pi_m(x'_s)$, that is to say $x_1 + \cdots + x_s = x'_1 + \cdots + x'_s (\operatorname{mod} m)$. Then, since $|x_1 + \cdots + x_s - x'_1 - \cdots - x'_s| < m$, we must have $x_1 + \cdots + x_s = x'_1 + \cdots + x'_s$. □

11.2. Ruzsa's model lemma

In this section we prove a remarkable lemma of Imre Ruzsa. It asserts that a subset of $\mathbf{Z}$ with small doubling has a large piece which is Freiman isomorphic to a dense subset of a cyclic group $\mathbf{Z}/m\mathbf{Z}$. In that setting the tools of harmonic analysis become much more powerful, unlike for arbitrary subsets of $\mathbf{Z}$ (even those of small doubling) which could well be highly "spread out". Here is Ruzsa's lemma.

PROPOSITION 11.2.1. *Suppose that $A \subset \mathbf{Z}$ is a finite set and that $s \geqslant 2$ is an integer. Let $m \geqslant |sA - sA|$ be an integer. Then there is a set $A' \subset A$ with $|A'| \geqslant |A|/s$ which is Freiman s-isomorphic to a subset of $\mathbf{Z}/m\mathbf{Z}$.*

Proof. By translating A is necessary, we may assume that A consists of positive integers. Let q be a prime number greater than all elements of A, and consider the composition $\phi_\lambda := \pi_m \circ \pi_q^{-1} \circ D_\lambda \circ \pi_q$ of maps

$$\mathbf{Z} \xrightarrow{\pi_q} \mathbf{Z}/q\mathbf{Z} \xrightarrow{D_\lambda} \mathbf{Z}/q\mathbf{Z} \xrightarrow{\pi_q^{-1}} \{1, \ldots, q\} \xrightarrow{\pi_m} \mathbf{Z}/m\mathbf{Z}$$

where π_q, π_m are reduction mod q and m respectively and D_λ is multiplication (dilation) by $\lambda \in (\mathbf{Z}/q\mathbf{Z})^\times$.

Now π_q, D_λ and π_m are Freiman homomorphisms of any order. By Proposition (11.1.1) (vi), π_q^{-1} is a Freiman homomorphism on any $\pi_q(I_j)$, $j = 0, 1, \ldots, s-1$, where $I_j := \{n \in \mathbf{Z} : \frac{jq}{s} < n \leqslant \frac{(j+1)q}{s}\}$. Since the $\pi_q(I_j)$ partition $\mathbf{Z}/q\mathbf{Z}$, it follows from the pigeonhole principle that for each λ there is some j such that, if we define

$$A'_\lambda := \{a \in A : D_\lambda(\pi_q(a)) \in \pi_q(I_j)\},$$

then $|A'_\lambda| \geqslant |A|/s$. By the preceding discussion, ϕ_λ is a Freiman s-homomorphism when restricted to A'_λ.

Everything we have said so far holds for an arbitrary λ. To conclude the proof we show that there is a choice of λ for which ϕ_λ is invertible when restricted to A'_λ, and for which its inverse is also a Freiman s-homomorphism. For this choice of λ, ϕ_λ will then be a Freiman s-isomorphism when restricted to A'_λ. Suppose, by contrast, that for every $\lambda \in (\mathbf{Z}/q\mathbf{Z})^*$ there are $a_i, a'_i \in A'_\lambda$ with

$$d_\lambda := a_1 + \cdots + a_s - a'_1 - \cdots - a'_s \neq 0 \qquad (11.1)$$

but

$$\phi_\lambda(a_1) + \cdots + \phi_\lambda(a_s) = \phi_\lambda(a_1') + \cdots + \phi_\lambda(a_s'). \tag{11.2}$$

Write

$$x_\lambda := \sum_{i=1}^{s} \pi_q^{-1}(D_\lambda(\pi_q(a_i)) - \sum_{i=1}^{s} \pi_q^{-1}(D_\lambda(\pi_q(a_i')).$$

Then (11.2) implies that $x_\lambda \equiv 0 (\operatorname{mod} m)$. Without loss of generality (switching the a_i, a_i' if necessary), $x_\lambda \geqslant 0$. Since all the a_i, a_i' lie in A_λ', it follows that

$$x_\lambda \in s(I_j - I_j) \subset (-q, q),$$

and so in fact $0 \leqslant x_\lambda < q$. However,

$$\pi_q(x_\lambda) = D_\lambda(\pi_q(d_\lambda)) = \pi_q(\lambda d_\lambda),$$

which is not zero since $d_\lambda \neq 0$ and we are assuming q is very large. It follows that

$$x_\lambda = \pi_q^{-1}(\pi_q(\lambda d_\lambda)).$$

Thus we conclude the following: for every $\lambda \in (\mathbf{Z}/q\mathbf{Z})^*$, there is some $d = d_\lambda \in (sA - sA) \setminus \{0\}$ such that $\pi_q^{-1}(\pi_q(\lambda d)) \equiv 0 (\operatorname{mod} m)$.

To get a contradiction, Let us fix d and ask about values of λ for which $d = d_\lambda$: lacking imagination, we call them "bad for d". As λ ranges over $(\mathbf{Z}/q\mathbf{Z})^\times$, $\pi_q(\lambda d)$ of course covers $(\mathbf{Z}/q\mathbf{Z})^\times$ uniformly, and hence the "unwrapped" set $\pi_q^{-1} \circ \pi_q(\lambda d)$ covers each point of $\{1, \dots, q-1\}$ precisely once. The number of elements y in this interval for which $\pi_m(y) = 0$ (that is to say y is divisible by m) is at most $(q-1)/m$. Since each d lies in the set $(sA - sA) \setminus \{0\}$, it follows that the number of λ which are bad for *some* d is at most

$$\frac{q-1}{m}\big(|sA - sA| - 1\big) < q - 1,$$

the inequality being a consequence of the assumption that $m \geqslant |sA - sA|$. This is contrary to what we proved before, namely that *every* λ is bad for some d. □

In our proof of Freiman's theorem, we will use the following corollary.

Corollary 11.2.1. *Suppose that $A \subset \mathbf{Z}$ is a finite set with doubling constant K. Then there is a prime $q \leqslant 2K^{16}|A|$ and a subset $A' \subset A$ with $|A'| \geqslant |A|/8$ such that A' is Freiman 8-isomorphic to a subset of $\mathbf{Z}/q\mathbf{Z}$.*

Proof. By the Plünnecke–Ruzsa inequality, Theorem 10.4.1, we have $|8A - 8A| \leqslant K^{16}|A|$. Now by Bertrand's postulate there is a prime p satisfying $|8A - 8A| \leqslant q \leqslant 2|8A - 8A|$. This prime of course satisfies the bound $q \leqslant 2K^{16}|A|$, and by the preceding proposition there is a subset $A' \subset A$ with $|A'| \geqslant |A|/8$ which is Freiman 8-isomorphic to a subset of $\mathbf{Z}/q\mathbf{Z}$. □

CHAPTER 12

Freiman's theorem

In this chapter we prove Freiman's theorem. We begin by proving some results about dense subsets of cyclic groups, since that is the situation that Corollary 11.2.1 puts us in.

12.1. Bogolyubov's lemma

DEFINITION 12.1.1. Suppose that $R = \{r_1, \dots, r_k\}$ is a set of nonzero elements of $\mathbf{Z}/q\mathbf{Z}$ and that $\varepsilon > 0$ is a parameter. Then we define the *Bohr set* $B(R, \varepsilon)$ with frequency set R and width ε by

$$B(R, \varepsilon) := \{x \in \mathbf{Z}/q\mathbf{Z} : \|\frac{r_i x}{q}\|_{\mathbf{T}} \leqslant \varepsilon \text{ for } i = 1, 2, \dots, k\}.$$

The parameter k is said to be the *dimension* of the Bohr set.

PROPOSITION 12.1.1 (Bogolyubov's lemma). *Let $S \subset \mathbf{Z}/q\mathbf{Z}$ be a set of size σq. Then $2S - 2S$ contains a Bohr set of dimension at most $4/\sigma^2$ and width at least $\frac{1}{10}$.*

Proof. We use harmonic analysis on $\mathbf{Z}/q\mathbf{Z}$. Consider the function $f := 1_S * 1_S * 1_{-S} * 1_{-S}$. This is supported on $2S - 2S$, that is to say if $f(x) > 0$ then $x \in 2S - 2S$. Note also that $\hat{1}_{-S}(r) = \overline{\hat{1}_S(r)}$, and so $\hat{f}(r) = |\hat{1}_S(r)|^4$. By the Fourier inversion formula and the fact that f is real, we have

$$f(x) = \sum_r |\hat{1}_S(r)|^4 e(rx/q) = \sum_r |\hat{1}_S(r)|^4 \cos(2\pi rx/q). \tag{12.1}$$

Let R be the set of all $r \neq 0$ for which $|\hat{1}_S(r)| \geqslant \sigma^{3/2}/2$. By Parseval's identity we have

$$|R|\frac{\sigma^3}{4} \leqslant \sum_{r \in R} |\hat{1}_S(r)|^2 \leqslant \sum_r |\hat{1}_S(r)|^2 = \frac{1}{q} \sum_{x \in \mathbf{Z}/q\mathbf{Z}} 1_S(x)^2 = \sigma,$$

and so

$$|R| \leqslant 4/\sigma^2. \tag{12.2}$$

We claim that $B(R, \frac{1}{10}) \subset 2S - 2S$, to which end it suffices to show that $f(x) > 0$ for $x \in B(R, \frac{1}{10})$. To do this, we will use the formula (12.1). We split the sum over r into three pieces: the term $r = 0$, the terms with $r \in R$, and all other terms.

Clearly

$$|\hat{1}_S(0)|^4 = \sigma^4.$$

If $r \in R$ then $\cos(2\pi rx/q) \geqslant 0$, so the sum of these terms is nonnegative. Finally,

$$\sum_{r \notin R \cup \{0\}} |\hat{1}_S(r)|^4 \cos(2\pi rx/q) \geqslant -\sum_{r \notin R \cup \{0\}} |\hat{1}_S(r)|^4 \geqslant -\frac{\sigma^3}{4} \sum_r |\hat{1}_S(r)|^2 = -\frac{\sigma^4}{4},$$

the last step being a further applucation of Parseval's identity. Combining all of this we obtain

$$f(x) \geqslant \sigma^4 + 0 - \frac{\sigma^4}{4} > 0,$$

as required. □

12.2. Generalised progressions in Bohr sets

It is by no means obvious what has been gained in proving Proposition 12.1.1. The answer is that a Bohr set $B(R, \varepsilon)$ has a great deal of structure, in particular containing a large generalised progression. The key proposition is as follows.

Proposition 12.2.1. *Let $R \subset \mathbf{Z}/q\mathbf{Z}$ be a set of size k, not containing zero. Let $0 < \varepsilon < \frac{1}{2}$. Then the Bohr set $B(R, \varepsilon)$ contains a proper generalised progression of dimension d and cardinality at least $(\varepsilon/k)^k q$.*

Proof. In the proof of this we will rely on a result from the geometry of numbers, Minkowski's second theorem. This is stated as Proposition 12.2.2 below. The proof will not be lectured and is not examinable, but it is given in Appendix **??**. To state the theorem, we need some terminology. We will have a centrally symmetric (that is, $x \in K$ implies $-x \in K$) convex body $K \subset \mathbf{R}^d$, and a lattice[1] $\Lambda \subset \mathbf{R}^d$. The determinant $\det(\Lambda)$ is the volume of a fundamental region of Λ. We define the *successive minima* $\lambda_1, \dots, \lambda_d$ of K with respect to Λ as follows: λ_j is the infimum of those λ for which the dilate λK contains j linearly independent elements of Λ.

Proposition 12.2.2 (Minkowski's second theorem). *We have $\lambda_1 \cdots \lambda_d \operatorname{vol}(K) \leqslant 2^d \det(\Lambda)$.*

Returning to the proof of Proposition 12.2.1, let $R = \{r_1, \dots, r_k\}$ and consider the lattice

$$\Lambda = q\mathbf{Z}^k + (r_1, \dots, r_k)\mathbf{Z}.$$

[1] A lattice is a discrete and cocompact subgroup of $\mathbf{R}^d$. It is a theorem that every lattice is of the form $\mathbf{Z}v_1 \oplus \mathbf{Z}v_2 \oplus \cdots \oplus \mathbf{Z}v_d$ for linearly independent $v_1, \dots, v_d$, which are then called an *integral basis* for Λ. The set $\mathcal{F} := \{x_1 v_1 + \cdots + x_d v_d : 0 \leqslant x_i < 1\}$ is then called a fundamental region for Λ; note that translates of it by Λ precisely cover $\mathbf{R}^d$. Note that the v_i (and hence $\mathcal{F}$) are not uniquely determined by Λ, but it turns out that the volume of $\mathcal{F}$ is.

Since q is prime, this may be written as a direct sum $q\mathbf{Z}^k \oplus \{0, 1, \dots, q-1\} \cdot (r_1, \dots, r_k)$. Thus Λ has index q as a subgroup of $q\mathbf{Z}^k$, and from this and the fact that $\det(q\mathbf{Z}^k) = q^k$ it follows that $\det(\Lambda) = q^{k-1}$ (see Lemma A.0.1).

Take $K \subset \mathbf{R}^k$ to be the box $\{\mathbf{x} : \|\mathbf{x}\|_\infty \leqslant \varepsilon q\}$. Let $\lambda_1, \dots, \lambda_k$ be the successive minima of K with respect to Λ. Since K is closed, $\lambda_j K$ contains j linearly independent elements of Λ. We may, by choosing each element in turn, select a basis $\mathbf{b}_1, \dots, \mathbf{b}_k$ for $\mathbf{R}^k$ with $\mathbf{b}_j \in \Lambda \cap \lambda_j K$ for all j. (Such a basis is called a *directional basis*; we should caution that, whilst the $\mathbf{b}_j$ are linearly independent elements of Λ, they need not form an integral basis for Λ. An example is presented on Sheet 3.) Thus $\mathbf{b}_j \in \Lambda$ and $\|\mathbf{b}_j\|_\infty \leqslant \lambda_j \varepsilon q$. Set $L_j := \lceil 1/\lambda_j k \rceil$ for $j = 1, \dots, k$. Then if $0 \leqslant l_j < L_j$ we have $\|l_j \mathbf{b}_j\|_\infty \leqslant \varepsilon q/k$ and therefore

$$\|l_1\mathbf{b}_1 + \dots + l_k\mathbf{b}_k\|_\infty \leqslant \varepsilon q.$$

Now each $\mathbf{b}_i$ lies in Λ and hence is congruent to $x_i(r_1, \dots, r_k)(\operatorname{mod} q)$ for some x_i, $0 \leqslant x_i < q$. Abusing notation slightly, we think of these x_i as lying in $\mathbf{Z}/q\mathbf{Z}$. The preceding observation implies that

$$\|\frac{(l_1x_1 + \dots + l_kx_k)r_i}{q}\|_{\mathbf{T}} \leqslant \varepsilon$$

for each i, or in other words the GAP $\{l_1x_1 + \dots + l_kx_k : 0 \leqslant l_i < L_i\}$ is contained in the Bohr set $B(R, \varepsilon)$.

It remains to prove a lower bound on the size of this progression and also to establish its properness. The lower bound on the size is easy: it is at least $k^{-k}(\lambda_1 \cdots \lambda_k)^{-1}$ which, by Minkowski's Second Theorem and the fact that $\det(\Lambda) = q^{k-1}$ and $\operatorname{vol}(K) = (2\varepsilon q)^k$, is at least $(\varepsilon/k)^k q$.

To establish the properness, suppose that

$$l_1x_1 + \dots + l_kx_k = l_1'x_1 + \dots + l_k'x_k(\operatorname{mod} q),$$

where $|l_i|, |l_i'| < \lceil 1/k\lambda_i \rceil$. Then the vector

$$\mathbf{b} = (l_1 - l_1')\mathbf{b}_1 + \dots + (l_k - l_k')\mathbf{b}_k$$

lies in $q\mathbf{Z}^k$ and furthermore

$$\|\mathbf{b}\|_\infty \leqslant \sum_{i=1}^k 2\lfloor \frac{1}{\lambda_i k} \rfloor \|\mathbf{b}_i\|_\infty \leqslant 2\varepsilon q.$$

Since we are assuming that $\varepsilon < 1/2$ it follows that $\mathbf{b} = 0$ and hence, due to the linear independence of the $\mathbf{b}_i$, that $l_i = l_i'$ for all i. Therefore the progression is indeed proper. □

12.3. Freiman's theorem: conclusion of the proof

In this section, we conclude the proof of Freiman's theorem. Let us begin by stating it again.

Theorem 12.3.1 (Freiman). *For any K, there are constants $d(K), C(K)$ with such that the following is true. Suppose that $A \subset \mathbf{Z}$ is a finite set with $|A + A| \leqslant K|A|$. Then A is contained in a proper d-dimensional progression P of dimension at most $d(K)$ and size at most $C(K)|A|$.*

Proof. By Corollary 11.2.1, the corollary of Ruzsa's model lemma, there is a prime $q \leqslant 2K^{16}|A|$ and a subset $A' \subset A$ with $|A'| \geqslant |A|/8$ such that A' is Freiman 8-isomorphic to a subset $S \subset \mathbf{Z}/q\mathbf{Z}$. If $\sigma := |S|/q$ then we have $\sigma \geqslant \frac{1}{16}K^{-16}$.

By Bogolyubov's lemma, Proposition 12.1.1, $2S - 2S$ contains a Bohr set of dimension at most $2^{10}K^{32}$ and width at least $\frac{1}{10}$.

By Proposition 12.2.1, that Bohr set (and hence $2S - 2S$) contains a proper generalised progression P of dimension at most $K^{O(1)}$ and cardinality at least $\exp(-K^{O(1)})q$. (We could keep track of exact constants, but this becomes a little tedious).

Now A' is Freiman 8-isomorphic to S, and so by Lemma 11.1.1 (iii), $2A' - 2A'$ is Freiman 2-isomorphic to $2S - 2S$. The inverse of this Freiman isomorphism restricts to a Freiman isomorphism $\phi : P \to \phi(P) \subset 2A' - 2A'$. By Lemma 11.1.1 (v), $Q = \phi(P)$ is also a proper generalised progression, of the same dimension and size as P. Therefore we have shown that $2A - 2A$ contains a proper generalised progression Q of dimension $K^{O(1)}$ and

$$|Q| \geqslant \exp(-K^{O(1)})|A|. \tag{12.3}$$

To finish the argument, we apply the covering lemma, Lemma 10.2.2, to the sets Q and A. Since

$$Q + A \subset (2A - 2A) + A = 3A - 2A,$$

the Plünnecke–Ruzsa inequality and (12.3) imply that

$$|Q + A| \leqslant K^5|A| \leqslant \exp(K^{O(1)})|Q|.$$

By Lemma 10.2.2, there is some set $Y = \{y_1, \dots, y_m\}$,

$$m \leqslant \exp(K^{O(1)}), \tag{12.4}$$

such that

$$A \subset (Q - Q) + Y.$$

Suppose that

$$Q = \{x_0 + l_1x_1 + \cdots + l_dx_d : 0 \leqslant l_i < L_i\}$$

and that

$$Y = \{y_1, \ldots, y_m\}.$$

Then

$$\begin{aligned}(Q - Q) + Y \subset \{&\tilde{x}_0 + l_1 x_1 + \cdots + l_d x_d + l'_1 y_1 + \cdots + l'_m y_m, 0 \leqslant l_i < 2L_i, 0 \leqslant l'_j < 2\} \\ &= \tilde{Q}\end{aligned}$$

where

$$\tilde{x}_0 = -(L_1 x_1 + \cdots + L_d x_d).$$

Note that $\tilde{Q}$ is a generalised progression of dimension $d + m$ and that

$$\operatorname{size}(\tilde{Q}) = 2^{d+m} L_1 \cdots L_d = 2^{d+m}|Q| \leqslant 2^{d+m}|2A - 2A| \ll_K |A|,$$

the penultimate step following since $Q \subset 2A - 2A$.

The dominant term in the bound is 2^m, which is double exponential in K. □

APPENDIX A

Geometry of numbers

The main goal of this section is to prove Minkowski's second theorem. First we briefly go over some standard properties of the determinant of a lattice.

Lemma A.0.1. *If $q \in \mathbf{N}$ then $\det(q\mathbf{Z}^d) = q^d$. If Λ, Λ' are two lattices with $\Lambda' \subset \Lambda$, then $\det(\Lambda')/\det(\Lambda) = [\Lambda : \Lambda']$, where the latter quantity is the index of Λ' as a subgroup of Λ, that is to say the number of cosets of Λ' needed to cover Λ.*

Now let us recall the statement of Minkowski's Second theorem, and let us also state Minkowski's *first* theorem. In both of these results, $K \subset \mathbf{R}^d$ is a centrally symmetric convex body, and $\Lambda \subset \mathbf{R}^d$ a lattice. The successive minima of K with respect to Λ are $\lambda_1, \dots, \lambda_d$.

Theorem A.0.1 (Minkowski I). *Suppose that $\mathrm{vol}(K) > 2^d \det(\Lambda)$. Then K contains a nonzero point of Λ.*

Theorem A.0.2 (Minkowski II). *We have $\lambda_1 \cdots \lambda_d \,\mathrm{vol}(K) \leqslant 2^d \det(\Lambda)$.*

Let us remark that Minkwoski I is a consequence of Minkowski II. To see this, note that if $\mathrm{vol}(K) > 2^d \det(\Lambda)$ then Minkowski II implies that $\lambda_1 \cdots \lambda_d < 1$. Since $\lambda_1 \leqslant \cdots \lambda_d$, this implies that $\lambda_1 < 1$. By the definition of λ_1, it follows that K contains at least one nonzero point of Λ.

Minkowski I is a very straightforward consequence of the following result, *Blichfeldt's lemma*, which is also an ingredient in the proof of Minkowski II.

Lemma A.0.2 (Blichfeldt's lemma). *Suppose that $K \subset \mathbf{R}^d$, and suppose that $\mathrm{vol}(K) > \det(\Lambda)$. Then there are two distinct points $\mathbf{x}, \mathbf{y} \in K$ with $\mathbf{x} - \mathbf{y} \in \Lambda$.*

Remark. Note that here K is not required to be either centrally symmetric or convex.

Proof. By considering the sets $K \cap B(0, R)$, as $R \to \infty$, whose volumes tend to that of K, we may assume that K lies inside some ball $B(0, R)$. Now let us suppose that the conclusion is false: then no translate of K contains two points of Λ, or in other words

$$\sum_{\mathbf{x}} 1_K(\mathbf{x} - \mathbf{t}) 1_\Lambda(\mathbf{x}) \leqslant 1$$

for all $\mathbf{t} \in \mathbf{R}^d$. Let R' be much bigger than R, and average this last inequality over $\mathbf{t}$ lying in the ball $B(0, R')$ to obtain

$$\sum_x 1_\Lambda(\mathbf{x})\Big(\frac{1}{\mathrm{vol}(B(0,R'))}\int_{B(0,R')} 1_K(\mathbf{x}-\mathbf{t})d\mathbf{t}\Big) \leqslant 1.$$

Since $K \subset B(0, R)$, the inner integral equals $\mathrm{vol}(K)$ if $\|x\| \leqslant R' - R$, and therefore

$$\sum_{\mathbf{x}} 1_\Lambda(\mathbf{x})1_{B(0,R'-R)}(\mathbf{x})d\mathbf{x} \leqslant \frac{\mathrm{vol}(B(0,R')}{\mathrm{vol}(K)},$$

and hence

$$\text{(A.1)} \quad \frac{1}{\mathrm{vol}(B(0,R'-R))}\sum_{\mathbf{x}} 1_\Lambda(\mathbf{x})1_{B(0,R'-R)}(\mathbf{x})d\mathbf{x} \leqslant \frac{\mathrm{vol}(B(0,R')}{\mathrm{vol}(B(0,R'-R))}\cdot\frac{1}{\mathrm{vol}(K)}.$$

However it is “clear” by tiling with fundamental parallelepipeds that

$$\lim_{r\to\infty}\frac{1}{\mathrm{vol}(B(0,r))}\sum_{\mathbf{x}} 1_\Lambda(\mathbf{x})1_{B(0,r)}(\mathbf{x}) = \frac{1}{\det(\Lambda)},$$

and moreover

$$\lim_{R'\to\infty}\frac{\mathrm{vol}(B(0,R')}{\mathrm{vol}(B(0,R'-R))} = 1.$$

Comparing with (A.1) immediately leads to

$$\frac{1}{\det(\Lambda)} \leqslant \frac{1}{\mathrm{vol}(K)},$$

contrary to assumption. □

Although we will not formally need it in what follows, let us pause to give the simple deduction of Minkowski I.

Proof. [Proof of Minkowski I] By Blichfeldt's lemma, the set $\frac{1}{2}K = \{\frac{1}{2}\mathbf{x} : \mathbf{x} \in \mathbf{R}^d\}$ contains two distinct points of Λ; thus there are $\mathbf{x}, \mathbf{y} \in K$ with $\frac{1}{2}(\mathbf{x}-\mathbf{y}) \in \Lambda$. However, since K is convex and centrally symmetric we have $\frac{1}{2}(\mathbf{x}-\mathbf{y}) \in K$. □

Now we turn to the proof of Minkowski II.

Proof. [Proof of Minkowski II] It is technically convenient to assume that K is *open*; this we may do by passing from K to the interior K°. Take a directional basis $\mathbf{b}_1, \dots, \mathbf{b}_d$ for Λ with respect to K. Since K is open, $\lambda_k K \cap \Lambda$ is spanned (over $\mathbf{R}$) by the vectors $\mathbf{b}_1, \dots, \mathbf{b}_{k-1}$. Indeed if it were not then we could choose some further linearly independent vector $\mathbf{b} \in \lambda_k K \cap \Lambda$, and by the openness of K this would in fact lie in $(\lambda_k - \varepsilon)K \cap \Lambda$ for some $\varepsilon > 0$, contrary to the definition of λ_k.

Write each given $\mathbf{x}$ in coordinates relative to the basis vectors $\mathbf{b}_i$ as $x_1\mathbf{b}_1 + \cdots + x_d\mathbf{b}_d$. We now define some rather unusual maps $\phi_j : K \to K$, by mapping $\mathbf{x} \in K$ to the centre of gravity of the slice of K which contains $\mathbf{x}$ and is parallel to the subspace spanned by $\mathbf{b}_1, \cdots, \mathbf{b}_{j-1}$ (for $j = 1$, $\phi_1(\mathbf{x}) = \mathbf{x}$). Next, we define a map $\phi : K \to \mathbf{R}^d$ by

$$\phi(\mathbf{x}) := \sum_{j=1}^{d} (\lambda_j - \lambda_{j-1})\phi_j(\mathbf{x}),$$

where we are operating with the convention that $\lambda_0 = 0$. Let us make a few further observations concerning the ϕ_j and ϕ. In coordinates we have $\phi_j(\mathbf{x}) = \sum_i c_{ij}(\mathbf{x})\mathbf{b}_i$, where $c_{ij}(\mathbf{x}) = x_i$ for $i \geqslant j$, and $c_{ij}(\mathbf{x})$ depends only on $x_j, \cdots, x_d$ for $i < j$. It follows that

$$\phi(\mathbf{x}) = \sum_{i=1}^{d} \mathbf{b}_i(\lambda_i x_i + \psi_j(x_{i+1}, \cdots, x_d))$$

for certain continuous functions ψ_j. It follows easily that

$$\text{vol}(\phi(K)) = \lambda_1 \cdots \lambda_d \,\text{vol}(K), \tag{A.2}$$

the Jacobian of the transformation $x_i' = \lambda_i x_i + \psi_i(x_{i+1}, \ldots, x_d)$ being $\lambda_1 \cdots \lambda_d$.

Suppose, as a hypothesis for contradiction, that $\lambda_1 \cdots \lambda_d \,\text{vol}(K) > 2^d \det(\Lambda)$. By Blichfeldt's lemma and (A.2), this means that $\phi(K)$ contains two elements $\phi(\mathbf{x})$ and $\phi(\mathbf{y})$ which differ by an element of $2 \cdot \Lambda = \{2\lambda : \lambda \in \Lambda\}$, and this means that $\frac{1}{2}(\phi(\mathbf{x}) - \phi(\mathbf{y})) \in \Lambda$. Write $\mathbf{x} = \sum_i x_i \mathbf{b}_i$ and $\mathbf{y} = \sum_i y_i \mathbf{b}_i$, and suppose that k is the largest index such that $x_k \neq y_k$. Then we have $\phi_i(\mathbf{x}) = \phi_i(\mathbf{y})$ for $i > k$, so that

$$\begin{aligned}\frac{\phi(\mathbf{x}) - \phi(\mathbf{y})}{2} &= \sum_{j=1}^{d} (\lambda_j - \lambda_{j-1})\big(\frac{\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})}{2}\big) \\ &= \sum_{j=1}^{k} (\lambda_j - \lambda_{j-1})\big(\frac{\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})}{2}\big).\end{aligned}$$

This has two consequences. First of all the convexity of K implies that $\frac{1}{2}(\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})) \in K$ for all j, and hence (again by convexity) $\frac{1}{2}(\phi(\mathbf{x}) - \phi(\mathbf{y})) \in \lambda_k K$. Secondly we may easily evaluate the coefficient of $\mathbf{b}_k$ when $\frac{1}{2}(\phi(\mathbf{x}) - \phi(\mathbf{y}))$ is written in terms of our directional basis: it is exactly $\lambda_k(x_k - y_k)/2$. In particular this is nonzero, which means that $\frac{1}{2}(\phi(\mathbf{x}) - \phi(\mathbf{y}))$ lies in Λ and $\lambda_k K$, but not in the span of $\mathbf{b}_1, \cdots, \mathbf{b}_{k-1}$. This is contrary to the observation made at the start of the proof. □

Bibliography

[1] H. Davenport, *Analytic methods for diophantine equations and diophantine inequalities* Second edition. With a foreword by R. C. Vaughan, D. R. Heath-Brown and D. E. Freeman. Edited and prepared for publication by T. D. Browning. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2005. xx+140 pp

[2] B. J. Green, *Course notes for C3.8*, available at
`http://people.maths.ox.ac.uk/greenbj/papersprimenumbers.pdf`

[3] H. Iwaniec, *Topics in classical automorphic forms*, Graduate Studies in Math. **17**, Springer.

[4] T. D. Wooley, *Large improvements in Waring's problem,* Ann. of Math. (2) **135** (1992), no. 1, 131–164.

[5] G. A. Freiman, *Foundations of a structural theory of set addition*. Translated from the Russian. Translations of Mathematical Monographs, Vol 37. American Mathematical Society, Providence, R. I., 1973. vii+108 pp.