1

Introduction

The analytic method of Hardy and Littlewood (sometimes called the 'circle method') was developed for the treatment of *additive problems* in the theory of numbers. These are problems which concern the representation of a large number as a sum of numbers of some specified type. The number of summands may be either fixed or unrestricted; in the latter case we speak of *partition problems*. The most famous additive problem is Waring's problem, where the specified numbers are the kth powers, so that the problem is that of representing a large number N as

$$N = x_1^k + x_2^k + \cdots + x_s^k, \tag{1.1}$$

where s and k are given and $x_1, \ldots, x_s$ are positive integers. Almost equally famous is Goldbach's ternary problem, where the specified numbers are the primes, and the problem is that of representing a large number N as

$$N = p_1 + p_2 + p_3.$$

The great achievements of Hardy and Littlewood were followed later by further remarkable progress made by Vinogradov, and it is not without justice that our Russian colleagues now speak of the 'Hardy–Littlewood–Vinogradov method'.

It may be of interest to recall that the genesis of the Hardy–Littlewood method is to be found in a paper of Hardy and Ramanujan [69] in 1917 on the asymptotic behaviour of $p(n)$, the total number of partitions of n. The function $p(n)$ increases like $e^{A\sqrt{n}}$, where A is a certain positive constant; and Hardy and Ramanujan obtained for it an asymptotic series, which, if one stops at the smallest term, gives $p(n)$ with an error $O(n^{-1/4})$. The underlying explanation for this high degree of accuracy, which Hardy describes as 'uncanny', was given by Rademacher [68] in 1937: there is a convergent series which represents $p(n)$ exactly, and this is initially almost the same as the asymptotic series. There is one other group of problems in which the Hardy–Littlewood method leads to exact formulae; these are problems concerning the representation of a number as the sum of a given number of squares. It seems unlikely that there are any such formulae for higher powers.

Waring's problem is concerned with the particular Diophantine equation (1.1). There is no difficulty of principle in extending the Hardy–Littlewood method to deal with more general equations of additive type[1], say

$$N = f(x_1) + f(x_2) + \cdots + f(x_s),$$

where $f(x)$ is a polynomial taking integer values; in particular to the equation

$$N = a_1x_1^k + a_2x_2^k + \cdots + a_sx_s^k. \tag{1.2}$$

It is only in recent years, however, that much progress has been made in adapting the method to Diophantine equations of a general (that is, non-additive) character. An account of these developments will be given later in these lectures, but we shall be concerned at first mainly with Waring's problem and with additive equations of the type (1.2). All work on general Diophantine equations depends heavily on either the methods or the results of the work on additive equations.

Finally, we shall touch on the subject of Diophantine inequalities. Here, too, some results of a general character are now known, but they are less complete and less precise than those for equations.

[1] See the monograph [63].

2

Waring's problem: history

In his *Meditationes algebraicae* (1770), Edward Waring made the statement that every number is expressible as a sum of 4 squares, or 9 cubes, or 19 biquadrates, 'and so on'. By the last phrase, it is presumed that he meant to assert that for every $k \geq 2$ there is some s such that every positive integer N is representable as

$$N = x_1^k + x_2^k + \cdots + x_s^k, \tag{2.1}$$

for $x_i \geq 0$. This assertion was first proved by Hilbert in 1909. Hilbert's proof was a very great achievement, though some of the credit should go also to Hurwitz, whose work provided the starting point. Hurwitz had already proved that if the assertion is true for any exponent k, then it is true for $2k$. I shall not discuss Hilbert's method of proof here; for this one may consult papers by Stridsberg [79], Schmidt [72] or Rieger [71]. It is usual to denote the least value of s, such that every N is representable, by $g(k)$. The exact value of $g(k)$ is now known for all values of k.

The work of Hardy and Littlewood appeared in several papers of the series 'On Partitio Numerorum' (P.N.), the other papers of the series being concerned mainly with Goldbach's ternary problem. In P.N. I [37] they obtained an asymptotic formula for $r(N)$, the number of representations of N in the form (2.1) with $x_i \geq 1$, valid provided $s \geq s_0(k)$, a certain explicit function of k. The asymptotic formula was of the following form:

$$r(N) = C_{k,s} N^{s/k-1} \mathfrak{S}(N) + O(N^{s/k-1-\delta}), \tag{2.2}$$

where $\delta > 0$ and

$$C_{k,s} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)} > 0.$$

In the above formula, $\mathfrak{S}(N)$ is an infinite series of a purely arithmetical nature, which Hardy and Littlewood called the *singular series*. They proved further that

$$\mathfrak{S}(N) \geq \gamma > 0, \tag{2.3}$$

for some γ independent of N, provided that $s \geq s_1(k)$. However they did not at that stage give any explicit value for $s_1(k)$. Thus the formula implies that

$$r(N) \sim C_{k,s} N^{s/k-1} \mathfrak{S}(N) \tag{2.4}$$

as $N \to \infty$, provided $s \geq \max(s_0(k), s_1(k))$, and thereby provided an independent proof of Hilbert's theorem.

Hardy and Littlewood introduced the notation $G(k)$ for the least value of s such that every sufficient large N is representable in the form (2.1); this function is really of more significance than $g(k)$, since the latter is affected by the difficulty of representing one or two particular numbers N. In P.N. II [38] and P.N. IV [39], Hardy and Littlewood proved that the asymptotic formula and the lower bound for $\mathfrak{S}(N)$ both hold for $s \geq (k-2)2^{k-1} + 5$, which implies that

$$G(k) \leq (k-2)2^{k-1} + 5.$$

In P.N. VI [40] they found a better upper bound for $G(k)$, though not for the validity of the asymptotic formula, and in particular they proved that $G(4) \leq 19$. The last paper of the series, P.N. VIII [41], was entirely concerned with the singular series and with the congruence problem to which it gives rise.

Hardy and Littlewood took as their starting point the generating function for $r(N)$, that is, the power series

$$\sum_{N=0}^{\infty} r(N) z^N = \left(\sum_{n=0}^{\infty} z^{n^k} \right)^s.$$

They expressed $r(N)$ in terms of this function by means of Cauchy's formula for the coefficients of a power series, using a contour integral taken along the circle $|z| = \rho$, where ρ is slightly less than 1. A helpful technical simplification was introduced by Vinogradov in 1928; this consists of replacing the power series by a finite exponential sum, and the effect is to eliminate a number of unimportant complications that occurred in the original presentation of Hardy and Littlewood.

Write $e(t) = e^{2\pi i t}$. We define $T(\alpha)$, for a real variable α, by

$$T(\alpha) = \sum_{x=1}^{P} e(\alpha x^k), \tag{2.5}$$

where P is a positive integer. Then

$$(T(\alpha))^s = \sum_m r'(m) e(m\alpha), \tag{2.6}$$

where $r'(m)$ denotes the number of representations of m as

$$x_1^k + \cdots + x_s^k, \quad (1 \le x_i \le P).$$

If $P \ge [N^{1/k}]$, where $[\lambda]$ denotes the integer part of any real number λ, then $r'(N)$ is the total number of representations of N in the form (2.1) with $x_i \ge 1$. Consequently $r'(N) = r(N)$. If we multiply both sides of (2.6) by $e(-N\alpha)$ and integrate over the unit interval $[0, 1]$ (or over any interval of length 1), we get

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha. \tag{2.7}$$

This is the starting point of our work on Waring's problem. It corresponds to the contour integral for $r(N)$ used by Hardy and Littlewood, with z replaced by $e^{2\pi i\alpha}$.

Our first aim will be to establish the validity of the asymptotic formula (2.2) for $r(N)$ as $N \to \infty$, subject to the condition $s \ge 2^k + 1$. It is possible to do this in a comparatively simple manner by using an inequality found by Hua in 1938 (Lemma 3.2 below). It may be of interest to observe that no improvement on the condition $s \ge 2^k + 1$ has yet been made for small values of k, as far as the asymptotic formula itself is concerned. For large k it has been shown by Vinogradov that a condition of the type $s > Ck^2 \log k$ is sufficient.

If we prove that the asymptotic formula holds for a particular value of s, say $s = s_1$, it will follow that every large number is representable as a sum of s_1 kth powers, whence $G(k) \le s_1$. But to prove this it is not essential to prove the asymptotic formula for the total number of representations; it would be enough to prove it for some special type of representation as a sum of s_1 kth powers. This makes it possible to get better estimates for $G(k)$ than one can get for the validity of the asymptotic formula. In 1934 Vinogradov proved that $G(k) < Ck \log k$ for large k, and we shall give a proof in Chapter 9. The best known results for small k were found by Davenport in 1939–41 [19].

A new ‘elementary’ proof of Hilbert’s theorem was given by Linnik in 1943 [58], and was selected by Khintchine as one of his ‘three pearls’ [53]. The underlying ideas of this proof were undoubtedly suggested by certain features of the Hardy–Littlewood method, and in particular by Hua’s inequality.

3

Weyl's inequality and Hua's inequality

The most important single tool for the investigation of Waring's problem, and indeed many other problems in the analytic theory of numbers, is Weyl's inequality. This was given, in a less explicit form, in Weyl's great memoir [96] of 1916 on the uniform distribution of sequences of numbers to the modulus 1. The explicit form for a polynomial, in terms of a rational approximation to the highest coefficient, was given by Hardy and Littlewood in P.N. I [37].

Lemma 3.1. (Weyl's Inequality) *Let $f(x)$ be a real polynomial of degree k with highest coefficient α:*

$$f(x) = \alpha x^k + \alpha_1 x^{k-1} + \cdots + \alpha_k.$$

Suppose that α has a rational approximation a/q satisfying

$$(a, q) = 1, \quad q > 0, \quad \left|\alpha - \frac{a}{q}\right| \le \frac{1}{q^2}.$$

Then, for any $\varepsilon > 0$,

$$\left|\sum_{x=1}^{P} e(f(x))\right| \ll P^{1+\varepsilon}\left(P^{-\frac{1}{K}} + q^{-\frac{1}{K}} + \left(\frac{P^k}{q}\right)^{-\frac{1}{K}}\right),$$

where $K = 2^{k-1}$ and the implied constant[1] depends only on k and ε.

Note. The inequality gives *some* improvement on the trivial upper bound P provided that $P^\delta \le q \le P^{k-\delta}$ for some fixed $\delta > 0$. If $P \le q \le P^{k-1}$, we get the estimate $P^{1-1/K+\varepsilon}$, and it is under these

[1] We use the Vinogradov symbol $\ll$ to indicate an inequality with an unspecified 'constant' factor. In the present instance, the factor which arises is in reality independent of k, but we do not need to know this.

conditions that Weyl's inequality is most commonly applied. It is obviously impossible to extract any better estimate than this from it. Note that Weyl's inequality fails to give any useful information if q is small, and this is natural because if $f(x) = \alpha x^k$ and α is very near to a rational number with small denominator, the sum is genuinely of a size which approaches P.

Proof. The basic operation in the proof is that of squaring the absolute value of an exponential sum, and thereby relating the sum to an average of similar sums with polynomials of degree one lower. Let

$$S_k(f) = \sum_{x=P_1+1}^{P_2} e(f(x)),$$

where $0 \leq P_2 - P_1 \leq P$, and where the suffix k serves to indicate the degree of $f(x)$. Then

$$\begin{aligned} |S_k(f)|^2 &= \sum_{x_1}\sum_{x_2} e(f(x_2) - f(x_1)) \\ &= P_2 - P_1 + 2\Re \sum_{\substack{x_1, x_2 \\ x_2 > x_1}} e(f(x_2) - f(x_1)). \end{aligned}$$

Put $x_2 = x_1 + y$. Then $1 \leq y < P_2 - P_1$, and

$$f(x_2) - f(x_1) = f(x_1 + y) - f(x_1) = \Delta_y f(x_1),$$

with an obvious notation. Hence

$$|S_k(f)|^2 = P_2 - P_1 + 2\Re \sum_{y=1}^{P} \sum_{x} e\left(\Delta_y f(x)\right),$$

where the summation in x is over an interval depending on y but contained in $P_1 < x \leq P_2$. This interval may, for some values of y, be empty.

In particular,

$$|S_k(f)|^2 \leq P + 2\sum_{y=1}^{P} |S_{k-1}(\Delta_y f)|,$$

where the interval for S_{k-1} is of the nature just described. By repeating the argument we get

$$|S_{k-1}(\Delta_y f)|^2 \leq P + 2\sum_{z=1}^{P} |S_{k-2}\,(\Delta_{y,z} f)\,|,$$

where the interval of summation in S_{k-2} depends on both y and z but is contained in $P_1 < x \leq P_2$. The use of Cauchy's inequality enables us to substitute for S_{k-1} from the second inequality into the first:

$$\begin{aligned} |S_k(f)|^4 &\ll P^2 + P\sum_{y=1}^{P} |S_{k-1}(\Delta_y f)|^2 \\ &\ll P^3 + P\sum_{y=1}^{P}\sum_{z=1}^{P} |S_{k-2}(\Delta_{y,z} f)|. \end{aligned}$$

The process can be continued, and the general inequality established in this way is

$$|S_k(f)|^{2^\nu} \ll P^{2^\nu-1} + P^{2^\nu-\nu-1}\sum_{y_1=1}^{P}\cdots\sum_{y_\nu=1}^{P} |S_{k-\nu}(\Delta_{y_1,\ldots,y_\nu} f)|\,. \quad (3.1)$$

This is readily proved by induction on ν, using again Cauchy's inequality together with the basic operation described above which expresses $|S_{k-\nu}|^2$ in terms of $S_{k-\nu-1}$. It is to be understood that the range of summation for x in $S_{k-\nu}$ in (3.1) is an interval depending on $y_1, \ldots, y_\nu$, but contained in $P_1 < x \leq P_2$.

At this point we interpolate a remark which will be useful in the proof of Lemma 3.2. This is that if, at the last stage of the proof of (3.1), we apply the basic operation in its original form, we get

$$|S_k(f)|^{2^\nu} \ll P^{2^\nu-1} + P^{2^\nu-\nu-1}\sum_{y_1=1}^{P}\cdots\sum_{y_\nu=1}^{P} \Re S_{k-\nu}(\Delta_{y_1,\ldots,y_\nu} f). \quad (3.2)$$

Here again, the range for x in $S_{k-\nu}$ depends on $y_1, \ldots, y_\nu$ and may sometimes be empty.

Returning to (3.1), we take $\nu = k-1$ and in the original S_k we take $P_1 = 0$, $P_2 = P$. We observe that

$$\Delta_{y_1,\ldots,y_{k-1}} f(x) = k!\alpha y_1 \cdots y_{k-1} x + \beta,$$

say, where β is a collection of terms independent of x. Hence

$$\left|S_1(\Delta_{y_1,\ldots,y_{k-1}} f)\right| = \left|\sum_x e(k!\alpha y_1 \cdots y_{k-1} x)\right|.$$

The sum on the right, taken over any interval of x of length at most P, is of the form

$$\left|\sum_{x=x_1}^{x_2-1} e(\lambda x)\right| \leq \frac{2}{|1-e(\lambda)|} = \frac{1}{|\sin \pi\lambda|} \ll \frac{1}{\|\lambda\|},$$

where $\|\lambda\|$ denotes the distance of λ from the nearest integer. This fails if λ is an integer, and indeed gives a poor result if λ is very near to an integer, but we can supplement it by the obvious upper bound P. Hence (3.1) gives

$$|S_k(f)|^K \ll P^{K-1} + P^{K-k} \sum_{y_1=1}^{P} \cdots \sum_{y_{k-1}=1}^{P} \min(P, \|k!\alpha y_1 \cdots y_{k-1}\|^{-1}).$$

We now appeal to a result in elementary number theory, which enables us to collect together all the terms in the sum for which $k!y_1 \cdots y_{k-1}$ has a given value, say m. The number of such terms is $\ll m^{\varepsilon}$. To prove this, it suffices to show that

$$d(m) \ll m^{\varepsilon}, \tag{3.3}$$

for any integer m, where $d(m) = \sum_{d|m} 1$ is the usual divisor function. Indeed there are at most $d(m)$ possibilities for each of $y_1, \ldots, y_{k-1}$. To establish (3.3) we suppose that $m = p_1^{\lambda_1} p_2^{\lambda_2} \cdots$, and note that

$$\frac{d(m)}{m^{\varepsilon}} = \prod_i \frac{\lambda_i + 1}{p_i^{\varepsilon\lambda_i}} \leq \prod_{p_i \leq 2^{1/\varepsilon}} \frac{\lambda_i + 1}{2^{\varepsilon\lambda_i}} \leq C(\varepsilon),$$

since $2^{-\varepsilon\lambda}(\lambda + 1)$ is bounded above for $\lambda > 0$.

Collecting terms as mentioned above, we get

$$|S_k(f)|^K \ll P^{K-1} + P^{K-k+\varepsilon} \sum_{m=1}^{k!P^{k-1}} \min(P, \|\alpha m\|^{-1}).$$

It remains to estimate the last sum in terms of the rational approximation a/q to α which was mentioned in the enunciation. We divide the sum over m into blocks of q consecutive terms (with perhaps one incomplete block), the number of such blocks being

$$\ll \frac{P^{k-1}}{q} + 1.$$

Consider the sum over any one block, which will be of the form

$$\sum_{m=0}^{q-1} \min(P, \|\alpha(m_1 + m)\|^{-1}),$$