4

Waring's problem: the asymptotic formula

We return to the starting point for our work on Waring's problem, namely (2.7) of Chapter 2:

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha)d\alpha, \tag{4.1}$$

where $T(\alpha)$ is the exponential sum (2.5) from 1 to P and $P \geq [N^{1/k}]$. There is no point in taking P larger than necessary, so we take $P = [N^{1/k}]$. The main term in the asymptotic formula will prove to be of order $N^{s/k-1}$, or P^{s-k}, as indeed it must be if any simple asymptotic formula is valid, for this is the only power of P which is consistent with the fact that there are P^s choices for $x_1, \ldots, x_s$ and the sums $x_1^k + \cdots + x_s^k$ represent numbers of order at most P^k.

Thus we can neglect any set of values of α in the integral (4.1) which can be shown to contribute to the integral an amount which is of strictly lower order than P^{s-k}. We are supposing $s \geq 2^k + 1$, and if we regard the absolute value of the integrand as

$$|T(\alpha)|^{s-2^k} |T(\alpha)|^{2^k},$$

it will follow from Lemma 3.2 that we can neglect any set of α for which $|T(\alpha)| \ll P^{1-\delta}$ for some fixed $\delta > 0$. To obtain such a set of α, we shall use Lemma 3.1.

The general plan in work on Waring's problem and similar problems is to divide the values of α into two sets: the *major arcs*, which contribute to the main term in the asymptotic formula, and the *minor arcs*, the contribution of which is estimated on lines such as those described above, and goes into the error term. The precise line of demarcation between the two sets depends very much on what particular auxiliary results are available, and may to some extent be a matter of personal choice. Generally speaking there are powerful (though somewhat complicated) methods available for the treatment of the major arcs, and the crux of the problem lies with the minor arcs. Having found, in any particular problem, a method which copes successfully with the minor arcs, one usually finds it convenient to enlarge them as far as the method in question will permit, in order to reduce the amount of work needed for the major arcs (even though this work might be relatively straightforward). In the present treatment we can take the major arcs to be few in number and short in length, compared with what is often the case in other work on the subject.

Around every rational number a/q (in its lowest terms) we put an interval

$$\mathfrak{M}_{a,q}: \quad |\alpha - a/q| < P^{-k+\delta}, \tag{4.2}$$

and we do this for

$$1 \le q \le P^{\delta}, \quad 1 \le a \le q, \quad (a,q) = 1. \tag{4.3}$$

These intervals do not overlap, since the distance between their centres is at least $P^{-2\delta}$ and this is much greater than their length. Moreover these intervals are contained in $0 \le \alpha \le 1$ except for the right-hand half of the interval round $1/1$, and for convenience we imagine this interval translated an amount 1 to the left, so that it comes to the right of $\alpha = 0$. The intervals $\mathfrak{M}_{a,q}$ are the major arcs, and their complement relative to $[0,1]$ constitutes the minor arcs, the totality of which we shall denote by $\mathfrak{m}$. In these definitions, δ is some fixed small positive number. It may be remarked that in many applications of the Hardy–Littlewood method, the length of $\mathfrak{M}_{a,q}$ in (4.2) would incorporate a factor q^{-1} as well as a negative power of P, but here this factor is not needed and it is a slight simplification to omit it.

Lemma 4.1. *If $s \ge 2^k + 1$, we have*

$$\int_{\mathfrak{m}} |T(\alpha)|^s d\alpha \ll P^{s-k-\delta'}$$

where δ' is a positive number depending on δ.

Proof. By a classical result of Dirichlet on Diophantine approximations, every α has a rational approximation a/q satisfying

$$1 \le q \le P^{k-\delta}, \quad |\alpha - a/q| < q^{-1}P^{-k+\delta}. \tag{4.4}$$

Moreover, we have $1 \le a \le q$ whenever $0 < \alpha < 1$. Since the last inequality in (4.4) is stronger than that in (4.2), we should have α in $\mathfrak{M}_{a,q}$ if $q \le P^\delta$. Hence if α is in $\mathfrak{m}$, we must have

$$q > P^\delta.$$

Since $|\alpha - a/q| < q^{-2}$, we can apply Lemma 3.1 to the exponential sum $T(\alpha)$, and since $P^k/q \ge P^\delta$ we get

$$|T(\alpha)| \ll P^{1+\varepsilon-\delta/K},$$

where $K = 2^{k-1}$. Combining this with Lemma 3.2, in the manner indicated earlier, we infer that

$$\begin{aligned}\int_{\mathfrak{m}} |T(\alpha)|^s d\alpha &\ll P^{(s-2^k)(1+\varepsilon-\delta/K)} \int_0^1 |T(\alpha)|^{2^k} d\alpha \\ &\ll P^{s-k-\delta'}\end{aligned}$$

for some $\delta' > 0$ depending on δ. This proves Lemma 4.1. □

It may be noted that instead of appealing to Dirichlet's theorem we could use a simple property of continued fractions: if we take a/q to be the last convergent to α for which $q \le P^{k-\delta}$, we again get (4.4).

We now turn our attention to the major arcs $\mathfrak{M}_{a,q}$. Here α is very near to a/q, with q relatively small. If the exponent of P in (4.2) had been $-k-\delta$ instead of $-k+\delta$, then $T(\alpha)$ would be practically constant on $\mathfrak{M}_{a,q}$, for we should have

$$\left|\alpha x^k - \frac{a}{q}x^k\right| < P^{-k-\delta}P^k = P^{-\delta}.$$

This, of course, is not the case, but nevertheless, the arc $\mathfrak{M}_{a,q}$ is so short that $T(\alpha)$ behaves relatively smoothly in that interval. Just how it varies is seen in the following lemma.

Lemma 4.2. *For α in $\mathfrak{M}_{a,q}$, putting $\alpha = \beta + a/q$, we have*

$$T(\alpha) = q^{-1}S_{a,q}I(\beta) + O(P^{2\delta}), \tag{4.5}$$

where

$$S_{a,q} = \sum_{z=1}^{q} e(az^k/q), \tag{4.6}$$

$$I(\beta) = \int_0^P e(\beta\boldsymbol{\xi}^k)d\boldsymbol{\xi}. \tag{4.7}$$

Proof. We collect together those values of x in the sum defining $T(\alpha)$ which are in the same residue class (mod q), as is natural because αx^k is not far from being periodic in x with period q. This is most conveniently done by putting $x = qy + z$ where $1 \le z \le q$; here y runs through an interval, depending on z, corresponding to the interval $0 < x \le P$. We get

$$T(\alpha) = \sum_{z=1}^{q} e(az^k/q) \sum_{y} e(\beta(qy+z)^k).$$

Now we endeavour to replace the discrete variable y by a continuous variable η, and replace the summation over y by an integration over η. If this can be done, we can then make a change of variable from η to ξ, where $\xi = q\eta + z$; the interval for ξ will be the original interval $0 \le \xi \le P$, and we shall have replaced the summation over y by

$$q^{-1} \int_0^P e(\beta \boldsymbol{\xi}^k) d\boldsymbol{\xi} = q^{-1} I(\beta),$$

the factor q^{-1} coming from $d\eta/d\xi$. Thus we shall get precisely the main term in (4.5).

We have to estimate the difference between the sum over y and the corresponding integral over η. For the present purpose a very crude argument is good enough. If $f(y)$ is any differentiable function, we have

$$|f(\eta) - f(y)| \le \frac{1}{2} \max |f'(\eta)| \quad \text{for} \quad |\eta - y| \le \frac{1}{2}.$$

Hence, on dividing any interval $A < \eta < B$ into intervals of length 1 together with two possible broken intervals, we obtain

$$\left| \int_A^B f(\eta) d\eta - \sum_{A<y<B} f(y) \right| \ll (B-A) \max |f'(\eta)| + \max |f(\eta)|.$$

In our case, $f(\eta) = e(\beta(q\eta + z)^k)$, whence

$$\max |f'(\eta)| \ll q|\beta| P^{k-1}, \quad \max |f(\eta)| = 1.$$

Also $B - A \ll P/q$. Hence the error in replacing the sum over y by the integral over η is

$$\ll Pq^{-1} q|\beta| P^{k-1} + 1 \ll P^{\delta},$$

since $|\beta| < P^{-k+\delta}$ by (4.2). Multiplying by q, which is $\le P^{\delta}$, to allow for the outside summation over z, we obtain the error term in (4.5). □

Later, in Lemma 9.1, we shall meet a more effective method for replacing a sum by the corresponding integral.

Lemma 4.3. *If* $\mathfrak{M}$ *denotes the totality of the major arcs* $\mathfrak{M}_{a,q}$*, then*

$$\int_{\mathfrak{M}} (T(\alpha))^s e(-N\alpha)d\alpha = P^{s-k}\mathfrak{S}(P^\delta, N)J(P^\delta) + O(P^{s-k-\delta'}) \tag{4.8}$$

for some $\delta' > 0$*, where*

$$\mathfrak{S}(P^\delta, N) = \sum_{q \le P^\delta} \sum_{\substack{a=1 \\ (a,q)=1}}^{q} (q^{-1}S_{a,q})^s e(-Na/q), \tag{4.9}$$

$$J(P^\delta) = \int_{|\gamma|<P^\delta} \left(\int_0^1 e(\gamma \boldsymbol{\xi}^k) d\boldsymbol{\xi} \right)^s e(-\gamma)d\gamma. \tag{4.10}$$

Proof. We first raise to the power s the expression (4.5) for $T(\alpha)$, valid on an individual major arc $\mathfrak{M}_{a,q}$. Since

$$|q^{-1}S_{a,q}I(\beta)| \le P$$

trivially, we get

$$(T(\alpha))^s = (q^{-1}S_{a,q})^s(I(\beta))^s + O(P^{s-1+2\delta}). \tag{4.11}$$

Multiplying by $e(-N\alpha)$ and integrating over $\mathfrak{M}_{a,q}$, that is, over $|\beta| < P^{-k+\delta}$, the main term in the last expression gives

$$(q^{-1}S_{a,q})^s e(-Na/q) \int_{|\beta|<P^{-k+\delta}} (I(\beta))^s e(-N\beta)d\beta$$

The integral here is independent of q and a, and therefore summation over q and a satisfying (4.3) gives

$$\mathfrak{S}(P^\delta, N) \int_{|\beta|<P^{-k+\delta}} (I(\beta))^s e(-N\beta)d\beta.$$

In the integrand we can replace N by P^k with a negligible error. Indeed we have $N - P^k \ll P^{k-1}$, so that

$$|e(-\beta N) - e(-\beta P^k)| \ll |\beta|P^{k-1} \ll P^{-1+\delta},$$

and the error in the integral is $\ll P^{-k+\delta}P^sP^{-1+\delta}$. Since a crude estimate for $|\mathfrak{S}(P^\delta, N)|$ is $P^{2\delta}$, this leads to a final error $P^{s-k-1+4\delta}$, which is negligible. The integral is now

$$\int_{|\beta|<P^{-k+\delta}} \left(\int_0^P e(\beta \boldsymbol{\xi}^k) d\boldsymbol{\xi} \right)^s e(-P^k \beta) d\beta,$$

and on putting $\xi = P\xi'$ and $\beta = P^{-k}\gamma$, this becomes

$$P^{s-k} J(P^\delta).$$

Thus we have obtained the main term in the result (4.8).

It remains to estimate the effect of the error term in (4.11). Integrated over $|\beta| < P^{-k+\delta}$, it becomes $\ll P^{s-k-1+3\delta}$. Summed over $a \leq q$ and over $q \leq P^\delta$, it becomes $P^{s-k-1+5\delta}$, and since δ is small this is of the form given in (4.8). $\square$

Definition. Let

$$\mathfrak{S}(N) = \sum_{q=1}^{\infty} \sum_{\substack{a\,=\,1 \\ (a,\,q)\,=\,1}}^{q} (q^{-1} S_{a,q})^s e(-Na/q). \tag{4.12}$$

This is called the *singular series* for the problem of representing N as a sum of s positive integral kth powers. If $s \geq 2^k + 1$, the series is absolutely convergent, and uniformly with respect to N, for by the Corollary to Lemma 3.1 we have (with $K = 2^{k-1}$):

$$|(q^{-1} S_{a,q})^s e(-Na/q)| \ll q^{-s/K+\varepsilon} \ll q^{-2-1/K+\varepsilon}.$$

Later we shall prove that the same is true under the less restrictive condition that $s \geq 2k + 1$.

Theorem 4.1. *If $s \geq 2^k + 1$, the number $r(N)$ of representations of N as a sum of s positive integral kth powers satisfies*

$$r(N) = C_{k,s} N^{s/k-1} \mathfrak{S}(N) + O(N^{s/k-1-\delta'}) \tag{4.13}$$

for some fixed $\delta' > 0$, where

$$C_{k,s} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)} > 0. \tag{4.14}$$

Proof. By (4.1) and Lemmas 4.1 and 4.3,

$$\begin{aligned} r(N) &= \left\{ \int_{\mathfrak{M}} + \int_{\mathfrak{m}} \right\} (T(\alpha))^s e(-N\alpha) d\alpha \\ &= P^{s-k} \mathfrak{S}(P^\delta, N) J(P^\delta) + O(P^{s-k-\delta'}). \end{aligned} \tag{4.15}$$

We first investigate the integral $J(P^\delta)$, defined in (4.10). The inner integral there can be expressed, by obvious changes of variable, in three ways:

$$\int_0^1 e(\gamma\boldsymbol{\xi}^k)d\boldsymbol{\xi} = k^{-1}\int_0^1 \zeta^{-1+1/k}e(\gamma\zeta)d\zeta = k^{-1}\gamma^{-1/k}\int_0^\gamma \zeta^{-1+1/k}e(\zeta)d\zeta,$$

where in the last expression we have supposed for simplicity that γ is positive. The integral in the last expression is a bounded function of γ, by Dirichlet's test for the convergence of an infinite integral together with the fact that the integral is absolutely convergent at 0. Hence

$$\left|\int_0^1 e(\gamma\boldsymbol{\xi}^k)d\boldsymbol{\xi}\right| \ll |\gamma|^{-1/k}.$$

This enables us to extend to infinity the integration over γ in (4.10); we obtain

$$J(P^\delta) = J + O(P^{-(s/k-1)\delta}),$$

where

$$J = \int_{-\infty}^{\infty}\left(k^{-1}\int_0^1 \zeta^{-1+1/k}e(\gamma\zeta)d\zeta\right)^s e(-\gamma)d\gamma. \tag{4.16}$$

Plainly J depends only on k and s, and we shall prove in a moment that $J = C_{k,s}$. We shall call J the *singular integral* for the problem of representing N as a sum of s positive kth powers.

By the absolute convergence of the series $\mathfrak{S}(N)$ and the result just proved for $J(P^\delta)$, we can replace $\mathfrak{S}(P^\delta, N)$ in (4.15) by $\mathfrak{S}(N)$ and we can replace $J(P^\delta)$ by J, with errors which are permissible. We can also replace P by $N^{1/k}$ with permissible error, and this gives (4.13), except for the proof that $J = C_{k,s}$. The exact value of J is perhaps unimportant, but we need to know that $J > 0$.

To evaluate J we start from the fact that

$$\int_{-\lambda}^{\lambda} e(\mu\gamma)d\gamma = \frac{\sin 2\pi\lambda\mu}{\pi\mu}.$$

Hence

$$\begin{aligned} k^s J &= \lim_{\lambda\to\infty}\int_0^1\cdots\int_0^1 (\zeta_1\cdots\zeta_s)^{-1+1/k}\frac{\sin 2\pi\lambda(\zeta_1+\cdots+\zeta_s-1)}{\pi(\zeta_1+\cdots+\zeta_s-1)}d\zeta_1\cdots d\zeta_s \\ &= \lim_{\lambda\to\infty}\int_0^s \phi(u)\frac{\sin 2\pi\lambda(u-1)}{\pi(u-1)}du, \end{aligned}$$

where

$$\phi(u) = \int_0^1 \cdots \int_0^1 \{\zeta_1 \cdots \zeta_{s-1}(u - \zeta_1 - \cdots - \zeta_{s-1})\}^{-1+1/k} \, d\zeta_1 \cdots d\zeta_{s-1},$$

and is taken over $\zeta_1, \ldots, \zeta_{s-1}$ for which $u-1 < \zeta_1 + \cdots + \zeta_{s-1} < u$. Here we have made a change of variable from ζ_s to u, where $\zeta_1 + \cdots + \zeta_s = u$.

We now recall Fourier's integral theorem for a finite interval, which states[1] that under certain conditions,

$$\lim_{\lambda \to \infty} \int_A^B \phi(u) \frac{\sin 2\pi\lambda(u - C)}{\pi(u - C)} du = \phi(C),$$

provided $A < C < B$. Assuming that this is applicable, we deduce that

$$\begin{aligned} k^s J &= \phi(1) \\ &= \int_0^1 \cdots \int_0^1 \{\zeta_1 \cdots \zeta_{s-1}(1 - \zeta_1 - \cdots - \zeta_{s-1})\}^{-1+1/k} \, d\zeta_1 \cdots d\zeta_{s-1}, \end{aligned}$$

where the integral is taken over $\zeta_1, \ldots, \zeta_{s-1}$ for which $0 < \zeta_1 + \cdots + \zeta_{s-1} < 1$. The last definite integral, over $s - 1$ variables, is an instance of an integral evaluated by Dirichlet; it is indeed an immediate extension of Euler's integral

$$B(p, q) = \int_0^1 x^{p-1}(1 - x)^{q-1} dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p + q)}.$$

We have[2]

$$\phi(1) = \frac{\Gamma(1/k)^s}{\Gamma(s/k)},$$

whence

$$J = \left(\frac{1}{k}\right)^s \frac{\Gamma(1/k)^s}{\Gamma(s/k)} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)}.$$

A sufficient condition for the validity of Fourier's integral theorem is that $\phi(u)$ should be of bounded variation. To verify this, put $\zeta_j = ut_j$. Then $\phi(u)$ is equal to

$$u^{s/k-1} \int_0^{1/u} \cdots \int_0^{1/u} \{t_1 \cdots t_{s-1}(1 - t_1 - \cdots - t_{s-1})\}^{-1+1/k} \, dt_1 \cdots dt_{s-1},$$

where the integral is over $t_1, \ldots, t_{s-1}$ for which $1 - 1/u < t_1 + \cdots + t_{s-1} < 1$. The region of integration contracts as u increases, and the

[1] See [97, §9.43] for example.
[2] See [97, §12.5].

integrand does not involve u. Hence $\phi(u)$ is the product of $u^{-1+s/k}$ and a positive monotonic decreasing function of u, and is therefore a function of bounded variation. This completes the proof. $\square$

Note. In our treatment of the singular integral, we have followed a paper of Landau [54]. For a slightly more general treatment, see a paper of Kestelman [52]. There are various devices by which the use of Fourier's integral theorem can be avoided; for example one can replace $I(\beta)$ by the finite sum

$$k^{-1} \sum_{0<m<P^k} m^{-1+1/k} e(\beta m),$$

or one can evaluate J indirectly as in Vinogradov [93, Chapter 3]. But on the whole the reference to Fourier's integral theorem seems natural and appropriate.

In the asymptotic formula (4.13) one can regard the first factor, $C_{k,s}N^{s/k-1}$, as measuring the 'density' of the solutions of

$$x_1^k + \cdots + x_s^k = N, \quad x_1 > 0, \ldots, x_s > 0$$

in real numbers; it is the $(s-1)$-dimensional content of this portion of a hypersurface. Otherwise expressed, it is (with a negligible error) the s-dimensional volume of the region

$$N - \frac{1}{2} < x_1^k + \cdots + x_s^k < N + \frac{1}{2}, \quad x_1 > 0, \ldots, x_s > 0.$$

The second factor, $\mathfrak{S}(N)$, can be regarded as a compensating factor to allow for the fact that kth powers of integers are not as uniformly distributed as are kth powers of real numbers, in that they are constrained by congruence restrictions. (The relation between $\mathfrak{S}(N)$ and congruences will emerge in the next section.) Thus the conclusion we draw from the asymptotic formula, expressed in somewhat vague terms, is that asymptotically the representations of a large number as a sum of s positive integral kth powers are actually dominated by these two influences, provided s is greater than some function of k.