THE REPRESENTATION OF INTEGERS BY POSITIVE TERNARY QUADRATIC FORMS

G. L. Watson

1. *Introduction and definitions.* Let $f=f(x, y, z)$ be a positive definite form of the type

$$ax^2+by^2+cz^2+ryz+szx+txy,$$

where x, y, z are integral valued variables, and the coefficients $a, \ldots, t$ are integers whose highest common factor is 1. As the determinant of such a form may be fractional, I define

$$d=d(f)=\tfrac{1}{2}\begin{vmatrix} 2a & t & s \\ t & 2b & r \\ s & r & 2c \end{vmatrix}=4abc+rst-ar^2-bs^2-ct^2,$$

and

$$C=C(f)=4ab-t^2;$$

thus $-C$ is the discriminant of the binary form $f(x, y, 0)$, and the necessary and sufficient condition for f to be positive definite is that $a>0$, $C>0$, and $d>0$.

In order that f should represent a positive integer n it is necessary, but not sufficient, that the congruence

$$f\equiv n \pmod m$$

should be soluble for every positive integer m. I shall call n (an) *exceptional* (integer of f) if this necessary condition holds and yet f does not represent n. If, further, n is not of the form $n_1^2 n_2$, where $n_1>1$ and n_2 is exceptional, then I shall say that n is *primitively* exceptional. I denote by $E(f)$ and $E_0(f)$ the numbers (possibly ∞) of exceptional and primitively exceptional integers of f. Clearly these numbers are arithmetical invariants of f, that is to say, they are unaltered by an integral unimodular transformation of the variables.

It is easy to see that there cannot be too many exceptional integers, or more precisely, that they must have zero density. Linnik* has shown how to find arithmetic progressions all of whose positive integers are representable by f. The special case when $E(f)=E_0(f)=0$ has been considered by Dickson† and by Jones and Pall‡, and I have carried these investigations further in an unpublished thesis.

* *Izvestia Akad. Nauk S.S.S.R.* 4, (1940) 363–402.

† *Annals of Math.*, 28 (1927), 333–341.

‡ *Acta Math.*, 70 (1939), 165–191.

[Mathematika, 1 (1954), 104–110]

As far as I know there are no results in the opposite direction in the literature. I prove here the

Theorem. *For any positive* δ, *and sufficiently large* $d=d(f)$, *we have*

$$E_0(f)>d^{1-\delta}.$$

In another paper under the same title I shall show that $E(f)$ is infinite except in certain very special cases.

There is a very voluminous literature, both classical and modern, on the arithmetical properties of quadratic forms. For the convenience of readers not familiar with it, I sketch the proof of a very elementary and imperfect result (Lemma 1), which is all that I require.

2. *Outline of proof.* Our main object is to prove

$$\sum_{\substack{0<n<d^{1-\delta/3} \\ n \text{ exceptional}}} 1>d^{1-\frac{1}{2}\delta-\epsilon}; \qquad (1)$$

we then show by an elementary argument that, among the integers n enumerated in the sum on the left, there cannot be too few that are primitively exceptional.

The proof of (1) depends essentially on the facts that the integers n, for which $f\equiv n$ is always soluble, have a lower density $>d^{-\epsilon}$, and that the number of such integers below a given bound can be estimated with an error term $O(d^{\frac{1}{2}+\epsilon})$. On the other hand, the proportion of integers up to $d^{1-\frac{1}{3}\delta}$ that can be represented by f is usually much smaller than $d^{-\epsilon}$; if this is so, (1) follows. If not, we consider integers, up to $d^{1-\frac{1}{3}\delta}$, satisfying the foregoing condition and also another, which (i) does not reduce their density too much and (ii) necessarily makes them unrepresentable by f.

The latter case arises only when the first two minima of the form are small in relation to d, and the situation may be illustrated by the particular case

$$f=x^2+y^2+cz^2, \quad d=4c, \quad c \text{ large}.$$

Here any integer below $d^{1-\delta/3}<c$ must, if representable by f, be a sum of two squares. This can be made impossible by taking, as the supplementary condition above referred to,

$$q\,\|\,n \ \ (i.e.\ q\,|\,n,\ q^2 \nmid n), \text{ for } q \text{ prime}, \equiv 3 \pmod 4).$$

The density of the integers considered is thereby reduced by a factor about q^{-1}, which is not important if we can choose an otherwise suitable q sufficiently small in relation to d.

3. *Preliminary; supplementary definitions.* Let

$$P=p_1\cdots p_{\nu-2}$$

be the product of the odd primes whose squares divide d, and let Q be the product of the odd primes that are simple factors of d. Write

$$\left.\begin{aligned}\chi_1(n)&=(2\,|\,n)\\ \chi_2(n)&=(-2\,|\,n)\end{aligned}\right\}\ \text{(Jacobi symbol) for } n \text{ odd, and } 0 \text{ for } n \text{ even,}$$

$$\chi_i(n)=(n\,|\,p_{i-2})\quad \text{(Legendre symbol), for } i=3,\ \ldots,\ \nu.$$

Consider the conditions

$$\chi_i(n)=\eta_i,\quad i=1,\ 2,\ \ldots,\ \nu, \tag{2}$$

$$(n,\ Q)=1, \tag{3}$$

where each $\eta_i=+1$ or -1.

Lemma 1. *$\eta_1,\ \ldots,\ \eta_\nu$ may be so chosen that, for every n satisfying* (2) *and* (3), *the congruence $f\equiv n$ is soluble to every modulus.*

Proof. It suffices to consider the congruence

$$f\equiv n \pmod{p^k}, \tag{4}$$

p^k being any prime power. Now f being primitive must, for any p, represent properly some integer prime to p, and may be transformed so as to have that integer as its leading coefficient. That is, we may suppose $p\nmid a$. Then if $p=2$, or if $p\,|\,P$, (4) is soluble, with $y=z=0$, for any n with $n\equiv a \pmod 8$, or $(n\,|\,p)=(a\,|\,p)$. Thus we have only to take $\eta_1,\ \eta_2=(\pm 2\,|\,a)$, or $\eta_{i+2}=(a\,|\,p_i)$, for $p=2,\ p_i$.

Supposing therefore that p is odd and does not divide P, we can transform f further so that p divides s and t. Then a transformation on y, z only makes $p\,|\,r$; thus we have

$$f\equiv ax^2+by^2+cz^2 \pmod p,$$

$$d\equiv 4abc\not\equiv 0 \pmod{p^2},$$

and so we may suppose $\qquad p\nmid ab.$

In case $p\,|\,Q$, we have $p\,|\,c$; we put $z=0$, and (4) reduces, for $k=1$, to

$$ax^2+by^2\equiv n \pmod p. \tag{5}$$

It is elementary that (5) is always soluble, and since by (3) we have $p\nmid n$, there cannot be a solution with $\partial f/\partial x\equiv 2ax$ and $\partial f/\partial y\equiv 2by$ both $\equiv 0 \pmod p$. Hence a solution of (4) for any k is easily deduced by induction*.

If p does not divide Q, we have $p\nmid c$ and we argue as above if $p\nmid n$, but put $z=1$ if $p\,|\,n$, and so replace n in (5) by $n-c\not\equiv 0 \pmod p$.

For the rest of the proof we may assume that f is a reduced form, that is, one which has a, b, c as its successive minima; for it is well known that every f is equivalent to such a form. This implies

$$a\leqslant b\leqslant c, \tag{6}$$

$$2abc\leqslant d=Cc-f(r,\ -s,\ 0)\leqslant Cc\leqslant 4abc, \tag{7}$$

$$f\geqslant c,\ \text{unless}\ z=0. \tag{8}$$

From these inequalities we deduce:

Lemma 2. *If $c<d^{1-\frac{1}{3}\delta}$, there are at most $O(d^{1-\frac{1}{2}\delta})$ values of $n<d^{1-\frac{1}{3}\delta}$ for which $f=n$ is soluble.*

Proof. We have identically

$$4aCf=C(2ax+ty+sz)^2+\{Cy+(2ar-st)z\}^2+4adz^2,$$

whence the number of z for which $f<d^{1-\frac{1}{3}\delta}$ is soluble (for x, y) is at most

$$O(C^{1/2}d^{-\delta/6})+O(1)=O(a^{1/2}b^{1/2}d^{-\delta/6}),$$

since by (7) and the hypothesis of the Lemma $d^{\delta/3}=O(C)$. From similar estimates for the number of possible x, y (in which by (6) the term $O(1)$ is relatively less important) the result follows.

4. *Estimation of character sums.* We now come to the main part of the proof. We assume that δ is positive, fixed throughout the argument, and sufficiently small, and we denote by ϵ a positive real number, not necessarily the same at each occurrence, which is always small compared with δ. We assume that d is large, and denote by ξ an arbitrary positive real number. The constants implied by the O-notation depend on ϵ and δ in formulae in which ϵ and δ occur, but are otherwise absolute.

Lemma 3. *For any $\eta_1,\ \ldots,\ \eta_\nu$, each $=+1$ or -1,*

$$\sum_{0<n<\xi,\ (2)} 1=2^{-\nu}\sum_{0<n<\xi,\ (n,\ 2P)=1} 1+O(P^{\frac{1}{2}+\epsilon}).$$

Proof. The sum on the left, multiplied by 2^ν, is equal to

$$\sum_{\rho_1=1,\ 2}\ \ldots\ \sum_{\rho_\nu=1,\ 2}\ \sum_{0<n<\xi}\ \prod_{i=1}^{\nu}\{\eta_i\chi_i(n)\}^{\rho_i}.$$

If we put $\rho_1=\ldots=\rho_\nu=2$, we obtain the sum on the right. Any other set of $\rho_1,\ \ldots,\ \rho_\nu$ gives an error term of the type

$$\pm\sum_{0<n<\xi}\chi(n),$$

where $\chi(n)$ is a non-principal character mod $8P$. By an inequality due to Pólya* such a sum is $O(P^{\frac{1}{2}}\log P)$. Adding $2^\nu-1$ such error terms, and dividing by 2^ν, we obtain the result.

* See *e.g.*, G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers* Oxford, 1938), 96, Theorem 123.

* *Nachrichten K. Ges. Wiss. Göttingen*, Math.-Phys. Klasse, 1918, 21–29. Pólya states the result for a proper character, but this restriction is easily removed.

For future reference we note that $\eta_1, \dots, \eta_\nu$ have not here been assumed to have the values of Lemma 1.

LEMMA 4. *For any $\eta_1, \dots, \eta_\nu$, each $= +1$ or -1,*

$$\sum_{0<n<\xi,\ (2),\ (3)} 1 = 2^{-\nu} \frac{\phi(2PQ)}{2PQ}\, \xi + O(P^{\frac{1}{2}+\epsilon} Q^\epsilon).$$

Proof. In the following summations, let u, v run through the (positive) divisors of $2P$, Q. Then, since by definition these integers are coprime, each divisor of $2PQ$ occurs just once among the values of $uv = w$.

The sum on the left is

$$\sum_v \mu(v) \sum_{0<n<\xi,\ (2),\ v|n} 1.$$

In the inner sum, $v^{-1}n$ satisfies conditions (2) with a different set of η_i. Hence by Lemma 3 the sum to be estimated is equal to

$$\sum_v \mu(v) \left\{ 2^{-\nu} \sum_{\substack{0<n<\xi,\ v|n \\ (2P,\,n)=1}} 1 + O(P^{\frac{1}{2}+\epsilon}) \right\}.$$

Since $\sum_u 1$, $\sum_v 1$, $\sum_w 1$ are respectively $O(P^\epsilon)$, $O(Q^\epsilon)$, $O(P^\epsilon Q^\epsilon)$, the last expression is

$$\begin{aligned} 2^{-\nu} \sum_v \mu(v) \sum_u \mu(u) \sum_{0<n<\xi,\ v|n,\ u|n} 1 &+ O(P^{\frac{1}{2}+\epsilon} Q^\epsilon) \\ &= 2^{-\nu} \sum_w \mu(w) \sum_{0<n<\xi,\ w|n} 1 + O(P^{\frac{1}{2}+\epsilon} Q^\epsilon) \\ &= 2^{-\nu}\, \xi \sum_w \mu(w)\, w^{-1} + O(P^{\frac{1}{2}+\epsilon} Q^\epsilon), \end{aligned}$$

whence the result.

LEMMA 5. *There exists a (least possible) prime q such that*

$$(2CPQ, q) = 1, \tag{9}$$

$$\text{if } q \| n, \text{ then } f(x, y, 0) = n \text{ is insoluble,} \tag{10}$$

$$q = O(C^{\frac{1}{2}+\epsilon} P^\epsilon Q^\epsilon). \tag{11}$$

Proof. We consider the sum

$$\sum_{0<n<\xi,\ (2CPQ,\,n)=1,\ (-C|n)=-1} 1.$$

It may be estimated as in Lemmas 3, 4; the only difference is that instead of fixing the value of each separate character, we fix only their product, and so replace $2^{-\nu}$ by $\frac{1}{2}$. Plainly, the least n satisfying the summation conditions must be a prime, and it is the required q. For when $(-C|q) = -1$, then $q \nmid a$, and the congruence

$$f(x, y, 0) \equiv 0 \pmod{q},$$

that is

$$(2ax + ty)^2 + Cy^2 \equiv 0 \pmod{q},$$

is easily seen to imply $x \equiv y \equiv 0 \pmod{q}$, $f(x, y, 0) \equiv 0 \pmod{q^2}$, and (10) follows.

Now q must satisfy (11), since

$$0 = \sum_{0<n<q,\ (2CPQ,\,n)=1,\ (-C|n)=-1} 1 = \tfrac{1}{2} q \frac{\phi(2CPQ)}{2CPQ} + O(C^{\frac{1}{2}+\epsilon} P^\epsilon Q^\epsilon),$$

and

$$\frac{2CPQ}{\phi(2CPQ)} = O\{(2CPQ)^\epsilon\}.$$

LEMMA 6. *For q as defined in Lemma 5, we have*

$$\sum_{0<n<\xi,\ q\|n,\ (2),\ (3)} 1 = 2^{-\nu} \frac{q-1}{q^2} \frac{\phi(2PQ)}{2PQ}\, \xi + O(C^\epsilon P^{\frac{1}{2}+\epsilon} Q^\epsilon).$$

Proof. We note that $q^{-1}n$ satisfies (2), with $\eta_i \chi_i(q)$ for η_i, and (3), with qQ for Q.

5. *Proof of Theorem.* First suppose $c \geqslant d^{1-\delta/3}$. In Lemma 6, take $\xi = d^{1-\delta/3}$. Then for each of the integers n enumerated in that Lemma, $f = n$ is insoluble with $z \neq 0$, by (8), and with $z = 0$, by (10); yet by Lemma 1 (giving the η_i the values of that Lemma) the congruence $f \equiv n$ is soluble to any modulus. Thus all these integers are exceptional, and so (1) follows on noting that ($2^{\nu-2}$ being the number of divisors of P)

$$2^\nu = O(P^\epsilon), \quad C = O(c^{-1}d) = O(d^{\delta/3}), \quad P^2 Q \leqslant d.$$

If however $c < d^{1-\delta/3}$, an inequality stronger than (1) may be obtained by subtracting the estimate of Lemma 2 from that of Lemma 4 (with $\xi = d^{1-\delta/3}$). Thus (1) holds in every case.

Now, supposing that the Theorem is false, write

$$R = \sum_{\substack{0<n<d^{1-\delta/3} \\ n \text{ primitively exceptional}}} 1,$$

and let $n_1, \dots, n_R$ be the integers enumerated in this sum. Plainly

$$R \leqslant E_0(f) \leqslant d^{1-\delta}.$$

We have therefore

$$\begin{aligned} \sum_{\substack{0<n<d^{1-\delta/3} \\ n \text{ exceptional}}} 1 \leqslant \sum_{m=1}^{R} \sum_{\substack{n' \\ 0<n_m n'^2<d^{1-\delta/3}}} 1 \leqslant \sum_{m=1}^{R} d^{\frac{1}{2}-\delta/6} n_m^{-\frac{1}{2}} \leqslant \sum_{m=1}^{R} d^{\frac{1}{2}-\delta/6} m^{-\frac{1}{2}} \\ = O(d^{\frac{1}{2}-\delta/6} R^{\frac{1}{2}}) = O(d^{1-2\delta/3}). \end{aligned}$$

This contradicts (1), and so the proof is complete.

6. *Conclusion.* There is no reason to suppose that the estimate is anywhere near best possible; for forms of a certain type, I can substantially increase it by an arithmetical argument based on the results of the companion paper referred to above. On the other hand, I would conjecture that $E_0(f)$ is always finite. If this is so, the fact that $E(f)$ is usually infinite is due to the existence, for at least one primitively exceptional n, of infinitely many N such that $N^2 n$ is exceptional.

I am indebted to Professor Davenport, Dr. Estermann, and Dr. Dolciani for reading earlier drafts of this paper, and making a number of suggestions.

University College,
London.

(*Received 24th October*, 1953.)

THE MINKOWSKI-HLAWKA THEOREM

C. A. ROGERS

1. Minkowski's fundamental theorem[†] and the Minkowski-Hlawka theorem[‡] play basic complementary roles in the Geometry of Numbers. Blichfeldt[§] showed essentially that Minkowski's fundamental theorem was a simple consequence of a more general theorem, in which the convex body was replaced by any measurable set and the lattice was replaced by a discrete set of points having a definite asymptotic density. Hlawka[||] himself showed that the Minkowski-Hlawka theorem could be proved in a slightly modified form, when the star body was replaced by any measurable set, but he did not replace the lattice by a more general set of points.

Recently Prof. A. M. Macbeath suggested to me the possibility of proving a more general form of the Minkowski-Hlawka theorem, in which the lattice is replaced by a more general set of points. It soon became clear to me that one form of the method used by Mahler[¶], Weyl[††], Rogers[‡‡] and Cassels[§§] makes no essential use of the fact that the set of points is assumed to be a lattice, but depends only on the set of points being not too dense. To make precise this concept of a set, which is not too dense, we define the upper density of a discrete set Λ in the following way. If Σ is any sphere, we define the density $\delta(\Sigma)$ of Λ in Σ to be the number of points of Λ in Σ divided by the volume of Σ. We define[||||] the upper density of Λ to be

$$\delta_+(\Lambda) = \limsup_{r\to\infty} \sup_{\Sigma(r)} \delta\big(\Sigma(r)\big),$$

where we use $\Sigma(r)$ to denote any sphere of radius r.

Using this definition our main result can be stated in the following form.

[†] See, for example, G. H. Hardy and E. M. Wright, *Theory of Numbers*, 2nd Edit. (Oxford, 1945), Chapters 3 and 24.

[‡] E. Hlawka, *Math. Zeitschrift*, 49 (1944), 285–312; for a simple proof see, for example, J. W. S. Cassels, *Proc. Cambridge Phil. Soc.*, 49 (1953), 165–166.

[§] H. F. Blichfeldt, *Trans. American Math. Soc.*, 15 (1914), 227–235.

[||] *Loc. cit.*

[¶] K. Mahler, *Journal London Math. Soc.*, 19 (1944), 201–205.

[††] H. Weyl, in unpublished work; see *Notes of the Seminar on Geometry of Numbers*, The Institute for Advanced Study, Princeton, 1949, page 46.

[‡‡] C. A. Rogers, see *Notes of the Seminar on Geometry of Numbers*, The Institute for Advanced Study, Princeton, 1949, pages 46–50.

[§§] J. W. S. Cassels, *loc. cit.*

[||||] There are many different ways of giving an equivalent definition of this upper density $\delta_+(\Lambda)$: the same formula will, for example, serve to define $\delta_+(\Lambda)$, if $\Sigma(r)$ is used to denote any convex body containing a sphere of radius r.

[MATHEMATIKA 1 (1954), 111–124]