THE NUMBER OF SOLUTIONS OF CONGRUENCES IN INCOMPLETE RESIDUE SYSTEMS

J. H. H. CHALK

To Professor L. J. Mordell
on the occasion of his 75th birthday

1. Let $\mathfrak{L}$ denote the set of points $\mathbf{x} = (x_1, \ldots, x_n)$ with integral co-ordinates in Euclidean n-space. For any fixed integer $m \geqslant 2$, let $C = C(m)$ be the set of such points in the cube $0 \leqslant x_i < m$ $(i = 1, 2, \ldots, n)$ and let $\mathfrak{S}$ be any subset of C. Suppose that $f(\mathbf{x})$ is any single-valued, integral-valued function, defined for all $\mathbf{x} \in \mathfrak{L}$. We consider solutions $\mathbf{x} \in \mathfrak{S}$ of the congruence

$$f(\mathbf{x}) \equiv 0 \pmod{m}. \tag{1}$$

When $\mathfrak{S} = C$, the problem is the familiar one in the theory of congruences, where $\mathbf{x}$ runs over complete residue systems, mod m; but when $\mathfrak{S} \neq C$, only special cases have been investigated (cf. Vinogradov **5**, Chap. V, problem 12a for the case $n = 1$ and Mordell **4**, for $n = 2$). So far, estimates for $N(\mathfrak{S}) = N(m, f, \mathfrak{S})$, the number of solutions $\mathbf{x} \in \mathfrak{S}$ of (1) take the form of asymptotic formulae, valid for large m. Vinogradov's method introduces a certain inequality; by expressing it in terms of n variables, as follows, we can deduce an inequality (see (7)) used by Mordell for the case $n = 2$:

THEOREM 1. *Let S be any subset of $\mathfrak{L}$. Suppose that $\phi(\mathbf{x})$ is a single-valued, complex-valued function defined on $\mathfrak{L}$ and satisfying*†

$$\left| \sum_{\mathbf{x} \in S} \phi(\mathbf{x}) e(-\mathbf{x} \cdot \mathbf{y}) \right| \leqslant \Phi \text{ for all non-zero } \mathbf{y} \in C, \tag{2}$$

where Φ is independent of $\mathbf{y}$. Let

$$S^* = \{\mathbf{x} | \mathbf{x} \in S, \mathbf{x} \equiv \mathbf{y} \pmod{m} \text{ for some } \mathbf{y} \in \mathfrak{S}\}. \tag{3}$$

If $M(\mathfrak{S})$ denotes the number of points in $\mathfrak{S}$, then

$$\sum_{\mathbf{x} \in S^*} \phi(\mathbf{x}) = m^{-n} M(\mathfrak{S}) \sum_{\mathbf{x} \in S} \phi(\mathbf{x}) + \theta m^{-n} \Phi E(\mathfrak{S}), \tag{4}$$

where

$$E(\mathfrak{S}) = \sum_{\mathbf{0} \neq \mathbf{y} \in C} \left| \sum_{\mathbf{z} \in \mathfrak{S}} e(-\mathbf{y} \cdot \mathbf{z}) \right|, \tag{5}$$

for some θ with $|\theta| \leqslant 1$.

As I can find no references to this, I give a proof in § 2. Mordell's inequality is the special case of (4),

Received March 16, 1962.

†For any real t, we write $e(t)$ for $\exp(2\pi i t m^{-1})$.

$$S = C, \qquad \phi(\mathbf{x}) = \sum_{t=0}^{m-1} e(tf(\mathbf{x})); \tag{6}$$

it then takes the shape

$$N(\mathfrak{S}) = m^{-n}M(\mathfrak{S})N(C) + \theta m^{-n-1}\Phi E(\mathfrak{S}), \tag{7}$$

where Φ now satisfies

$$\left| \sum_{\mathbf{x}\in C} \sum_{t=0}^{m-1} e\{tf(\mathbf{x}) - \mathbf{x}.\mathbf{y}\} \right| \leqslant \Phi, \tag{8}$$

by (2) and (6). In applications, where $f(\mathbf{x})$ is a given polynomial with integral coefficients, $\mathfrak{S}$ is usually restricted to be some fixed parallelopiped in C; thus in §§ 4, 5, 6 we take $\mathfrak{S} = \mathfrak{C}$, where

$$\mathfrak{C} : \nu_i \leqslant x_i \leqslant \nu_i + h_i - 1 \qquad (i = 1, 2, \ldots, n) \tag{9}$$

and $0 < \nu_i < \nu_i + h_i - 1 < m$ $(i = 1, 2, \ldots, n)$. If we choose $\nu_i = 1$, say $(i = 1, 2, \ldots, n)$ the problem is essentially that of finding *small* solutions $\mathbf{x}$ of (1); that is, seeking small values of $h_1, \ldots, h_n$ compatible with $N(\mathfrak{C}) > 0$. So far as (7) is concerned we require estimates for Φ and $E(\mathfrak{S})$, assuming that the classical question for $N(C)$ has been settled. With $\mathfrak{S} = \mathfrak{C}$, the ranges of summation in (5) are independent of one another and knowledge of the elementary result for $n = 1$ (cf. **5**, Chap. III, problem 11c) provides a uniform estimate for $E(\mathfrak{C})$:

$$E(\mathfrak{C}) < \{m \log m - \phi m + d\}^n, \tag{10}$$

where $d = \max h_i$ and $\phi > 0$ is a constant (see § 3). Determination of a useful bound Φ is more difficult, even though (8) involves only sums over complete sets of residues, mod m. It is possible for certain classes of functions $f(\mathbf{x})$, and I illustrate the procedure in the case

$$f(\mathbf{x}) = a_1x_1^{k_1} + a_2x_2^{k_2} + \ldots + a_nx_n^{k_n} + c, \tag{11}$$

where $(a_i, m) = 1$, $(c, m) = 1$, $2 \leqslant k_i \leqslant p - 2$ $(i = 1, 2, \ldots, n)$ and m is prime (see §§ 4, 5, 6), obtaining an asymptotic formula for $N(\mathfrak{C})$ valid for $n \geqslant 3$, together with partial results for $n = 2$. Since $x^{m-1} \equiv 1 \pmod m$ if $x \not\equiv 0 \pmod m$ and m is prime, it is clear that we may assume that $1 \leqslant k_i \leqslant p-2$ $(i = 1, 2, \ldots, n)$. The case when one or more of the k_i's is 1 could also be treated by means of (7), giving a slightly weaker result. Presumably, it is then more effective to adopt Vinogradov's procedure, using (4) with $\phi(\mathbf{x}) = 1$ (cf. **5**, Chap. VI, problem 15b, β for the case $n = 2$, $k_2 = 1$).

2. Proof of the theorem. The inequality (4) is deduced from the identity:

$$m^n \sum_{\mathbf{x}\in S^*} \phi(\mathbf{x}) = \sum_{\mathbf{z}\in\mathfrak{S}} \sum_{\mathbf{x}\in S} \sum_{\mathbf{y}\in C} \phi(\mathbf{x})e\{\mathbf{y}\cdot(\mathbf{z} - \mathbf{x})\}, \tag{12}$$

which is an immediate consequence of

https://doi.org/10.4153/CJM-1963-032-5 Published online by Cambridge University Press

$$\sum_{\mathbf{w}_1 \in C} e(\mathbf{w}_1 \cdot \mathbf{w}_2) = \begin{cases} m^n \text{ if } \mathbf{w}_2 \equiv \mathbf{0} \pmod{m}, \\ 0 \text{ otherwise;} \end{cases} \tag{13}$$

for the sum in $\mathbf{y}$ on the right of (12) is 0 unless $\mathbf{z} \equiv \mathbf{x} \pmod{m}$ and m^n otherwise. Picking out the term $\mathbf{y} = \mathbf{0}$ in (12), we have

$$\begin{aligned} m^n \sum_{\mathbf{x} \in S^*} \phi(\mathbf{x}) &= \sum_{\mathbf{x} \in S} \sum_{\mathbf{z} \in \mathfrak{S}} \phi(\mathbf{x}) + \sum_{\mathbf{0} \neq \mathbf{y} \in C} \sum_{\mathbf{z} \in \mathfrak{S}} \sum_{\mathbf{x} \in S} \phi(\mathbf{x})\, e\{\mathbf{y} \cdot (\mathbf{z} - \mathbf{x})\} \\ &= M(\mathfrak{S}) \sum_{\mathbf{x} \in S} \phi(\mathbf{x}) + \sum_{\mathbf{0} \neq \mathbf{y} \in C} \left\{ \sum_{\mathbf{x} \in S} \phi(\mathbf{x})\, e(-\mathbf{x} \cdot \mathbf{y}) \right\} \sum_{\mathbf{z} \in \mathfrak{S}} e(\mathbf{y} \cdot \mathbf{z}) \\ &= M(\mathfrak{S}) \sum_{\mathbf{x} \in S} \phi(\mathbf{x}) + \theta \Phi E(\mathfrak{S}) \end{aligned}$$

for some θ with $|\theta| \leqslant 1$.

COROLLARY. *If* (6) *holds, then* $S^* = \mathfrak{S}$,

$$\sum_{\mathbf{x} \in S^*} \phi(\mathbf{x}) = mN(\mathfrak{S}),$$

$$\sum_{\mathbf{x} \in S} \phi(\mathbf{x}) = mN(C),$$

and (7) *holds.*

Proof. From the definition of $\phi(\mathbf{x})$ in (6), we see that $\phi(\mathbf{x}) = 0$ unless $f(\mathbf{x}) \equiv 0 \pmod{m}$ and $\phi(\mathbf{x}) = m$ wherever $f(\mathbf{x}) \equiv 0 \pmod{m}$, on summing the G.P. with respect to t.

3. A bound for $E(\mathfrak{C})$. For $t = 1, 2, \ldots, m-1$ we have the well-known estimate

$$\sum_{t=1}^{m} \left| \sum_{x=\nu}^{\nu+h-1} e(tx) \right| < m \log m - \phi m, \tag{14}$$

where ϕ is an absolute constant ($=1$ for $m \geqslant 60$), which is uniform in h (see **5**, Chap. III, problem 11c). Consider

$$E(\mathfrak{C}) = \sum_{\mathbf{y} \in C} \left| \sum_{\mathbf{z} \in \mathfrak{C}} e(\mathbf{y} \cdot \mathbf{z}) \right|$$

and take, firstly, the terms with $\mathbf{y} = (y_1, \ldots, y_n)$, $y_i > 0$ $(i = 1, 2, \ldots, n)$. These may be written as

$$\sum_{y_1=1}^{m-1} \cdots \sum_{y_n=1}^{m-1} \left| \sum_{z_n=\nu_n}^{\nu_n+h_n-1} e(y_n z_n) \right| \cdots \left| \sum_{z_1=\nu_1}^{\nu_1+h_1-1} e(y_1 z_1) \right|,$$

and so, applying (14) successively, we see that their contribution to $E(\mathfrak{C})$ is less than $(m \log m - \phi m)^n$. Now take the terms with $\mathbf{y} = (y_1, \ldots, y_n)$, which have just r of the co-ordinates of $\mathbf{y}$ vanishing. There are $\binom{n}{r}$ such terms, each of which is less than

$$d^r (m \log m - \phi m)^{n-r},$$

https://doi.org/10.4153/CJM-1963-032-5 Published online by Cambridge University Press

where d is the largest of the numbers $h_1, \ldots, h_n$. Hence

$$\begin{aligned} E(\mathfrak{C}) &< \sum_{r=0}^{n} \binom{n}{r} d^r (m \log m - \phi m)^{n-r} \\ &= (m \log m - \phi m + d)^n \\ &< (m \log m)^n \quad \text{if } m \geqslant 60. \end{aligned}$$

4. Let m be a prime p, say, and consider the case when

$$f(\mathbf{x}) = a_1 x_1^{k_1} + \ldots + a_n x_n^{k_n} + c \equiv 0 \pmod{p}. \tag{15}$$

This has received considerable attention in the literature. Special cases of it ($n = 2$, $(k_1, k_2) = (3, 3), (4, 4), (2, 4)$) were investigated by Gauss (cf. *Werke*, Vol. I, pp. 445–449) and others. Later, Hardy and Littlewood required the total number of solutions of the congruence

$$\sum_{i=1}^{n} a_i x_i^m + c \equiv 0 \pmod{p}$$

in connection with their work on Waring's problem. In 1933, Mordell **(3)** obtained for (15) the asymptotic formula

$$N(C) = p^{n-1} + O(p^{\frac{1}{2}(n-1)}) \tag{16}$$

when $(a_1 a_2 \ldots a_n c, p) = 1$. Since then, A. Weil **(7)** and Hua and Vandiver **(2)** have developed exact expressions for N in terms of generalized Gaussian sums, giving (16) with fairly precise information about the size of the implied constant in the O-symbol. Thus, quoting Weil, we know that

$$|N - p^{n-1}| \leqslant M_0 (p-1) p^{\frac{1}{2}(n-2)} \quad \text{if } c \equiv 0 \pmod{p}, \tag{17}$$

where M_0 is the number of systems of rational numbers α_i satisfying

$$k_i \alpha_i \equiv 0, \quad \sum \alpha_i \equiv 0 \pmod{1}, \quad 0 < \alpha_i < 1$$

(and is therefore an integer depending only on $k_1, \ldots, k_n$) and

$$|N - p^{n-1}| \leqslant M p^{\frac{1}{2}(n-1)} \quad \text{if } c \not\equiv 0 \pmod{p}, \tag{18}$$

where, for example, $M = (d_1 - 1)(d_2 - 1) \ldots (d_n - 1) < k_1 \ldots k_n$ and $a_i = (k_i, p-1)$; provided that in each case $(a_i, p) = 1$ $(i = 1, 2, \ldots, n)$.

The derivation of these results may be termed elementary in the sense that they do not appeal to an analogue of the Riemann hypothesis. But for our problem, where we require an asymptotic formula for $N(\mathfrak{C})$, as opposed to $N(C)$, I have been unable to avoid employing Weil's estimate for the exponential sum

$$\sum_{x=0}^{p-1} e(ax^k + bx), \tag{19}$$

which, so far as I know, is derivable only as a consequence of his proof of the Riemann hypothesis for algebraic function-fields over a finite field (cf. **8**). Explicitly, we assume that the sum (19) is $O(p^{\frac{1}{2}})$ as $p \to \infty$, when $2 \leqslant k \leqslant p-2$

and either $a \not\equiv 0$ or $b \not\equiv 0 \pmod{p}$. For a proof of this and of more general exponential sums, see A. Weil **(6)** or L. Carlitz and S. Ucheyama **(1)**. Less precise estimates for (19) are known (see, for example, **5**, Chap. VI, problem 15a), but they are not sharp enough for some of the cases arising.

5. THEOREM 2. *If* $(a_1 a_2 \ldots a_n c, p) = 1$, $2 \leqslant k_i \leqslant p-2$ $(i = 1, 2, \ldots, n)$, *then*

$$N(\mathfrak{C}) = p^{-1} M(\mathfrak{C}) + O(p^{\frac{1}{2}n} (\log p)^n). \tag{20}$$

Proof. By (2) and (6), we put

$$F(\mathbf{y}) = \sum_{\mathbf{x} \in C} \sum_{t=0}^{p-1} e\{tf(\mathbf{x}) - \mathbf{x}\cdot\mathbf{y}\} \tag{21}$$

and seek an upper bound for $|F(\mathbf{y})|$ over all non-zero $\mathbf{y} \in C$. Picking out the term with $t = 0$, we get

$$\begin{cases} F(\mathbf{y}) = \sum_{\mathbf{x} \in C} e(-\mathbf{x}\cdot\mathbf{y}) + \sum_{t=1}^{p-1} \sum_{\mathbf{x} \in C} e\{tf(\mathbf{x}) - \mathbf{x}\cdot\mathbf{y}\} \\ \qquad = S_0(\mathbf{y}) + \sum_{t=1}^{p-1} S_t(\mathbf{y}), \text{ say,} \end{cases} \tag{22}$$

where

$$S_0(\mathbf{y}) = \sum_{\mathbf{x} \in C} e(-\mathbf{x}\cdot\mathbf{y}) = \begin{cases} p^n \text{ if } \mathbf{y} \equiv \mathbf{0} \pmod{p}, \\ 0 \text{ otherwise,} \end{cases} \tag{23}$$

and

$$S_t(\mathbf{y}) = \sum_{\mathbf{x} \in C} e\{tf(\mathbf{x}) - \mathbf{x}\cdot\mathbf{y}\}. \tag{24}$$

By (15),

$$\begin{aligned} \left| \sum_{t=1}^{p-1} S_t(\mathbf{y}) \right| &= \left| \sum_{t=1}^{p-1} e(ct) \left\{ \prod_{i=1}^{n} \sum_{x_i=0}^{p-1} e\{ta_i x_i^{k_i} - x_i y_i\} \right\} \right| \\ &\leqq \sum_{t=1}^{p-1} \prod_{i=1}^{n} \left| \sum_{x_i=0}^{p-1} e\{ta_i x_i^{k_i} - x_i y_i\} \right| \qquad (25) \\ &= O(p^{\frac{1}{2}n+1}), \end{aligned}$$

by our assumptions concerning (19). Combining (22), (23), and (25), we have

$$F(\mathbf{y}) = O(p^{\frac{1}{2}n+1}) \qquad \text{if } \mathbf{y} \not\equiv \mathbf{0} \pmod{p}, \tag{26}$$

whence, by (8), we may take $\Phi = O(p^{\frac{1}{2}n+1})$. From (7), (8), (16), (21), (26), and our bound for $E(\mathfrak{C})$, we have

$$N(\mathfrak{C}) = M(\mathfrak{C})\{p^{n-1} + O(p^{\frac{1}{2}(n-1)})\}p^{-n} + O\{p^{-n-1} \cdot p^{\frac{1}{2}n+1} \cdot (p \log p)^n\} \tag{27}$$

$$\begin{aligned} &= p^{-1} M(\mathfrak{C}) + O\{p^{\frac{1}{2}n}(\log^n p + p^{-n-\frac{1}{2}} M(\mathfrak{C}))\} \qquad (28) \\ &= p^{-1} M(\mathfrak{C}) + O\{p^{\frac{1}{2}n} \log^n p\}, \end{aligned}$$

since $M(\mathfrak{C}) \leqslant p^n$.

https://doi.org/10.4153/CJM-1963-032-5 Published online by Cambridge University Press

6. Conclusions. We remark that (20) is a *bona fide* asymptotic formula when $n \geqslant 3$, in the sense that we can choose $\mathfrak{C} \subset C$ large enough to satisfy

$$M(\mathfrak{C}) > Cp^{\frac{1}{2}n+1} \log^n p = o(p^n), \qquad \text{as } p \to \infty$$

for a suitably large constant C. But, for $n = 2$, the estimate is vacuous and I have been unable to find a useful result, except when $f(\mathbf{x})$ is quadratic. Returning to the question of finding small non-trivial solutions of (15), observe that on taking

$$\nu_i = 1, \qquad h_i = [kp^{1/2+1/n} \log p]$$

with a suitably large constant k (depending, in fact, only on $k_1, \ldots, k_n$) we see that

$$p^{-1}M(\mathfrak{C}) = p^{-1}[kp^{1/2+1/n} \log p]^n$$

exceeds the error term in (20). Thus, there is a solution $(x_1, \ldots, x_n)$ of (20), with

$$1 \leqslant x_i \leqslant [kp^{1/2+1/n} \log p] \qquad (i = 1, 2, \ldots, n), \tag{29}$$

and we emphasize that the term on the right of (29) is $o(p)$, when $n \geqslant 3$. Mordell's method for the case $n = 2$, $k_1 = k_2 = 2$ uses the estimate for $E(\mathfrak{C})$ in § 3 and a *refinement* of the argument in § 5 for the estimation of $F(\mathbf{y})$ (and so, of Φ). Omitting details, it is straightforward to show that this refinement goes through for $k_1 = k_2 = \ldots = k_n = 2$ $(n \geqslant 2)$, giving

$$1 \leqslant x_i \leqslant [k'p^{1/2+1/(2n)} \log p]. \tag{30}$$

It would be of interest to find some estimate for the case of two variables, say

$$f(\mathbf{x}) = ax^l + by^m + c, \qquad abc \not\equiv 0 \pmod{p},$$

where $l \geqslant 2$, $m \geqslant 3$.

References

1. L. Carlitz and S. Ucheyama, *Bounds for exponential sums*, Duke Math. J., *34* (1957), 37–41.
2. L. K. Hua and H. S. Vandiver, *Characters over certain types of rings with applications to the theory of equations in a finite field*, Proc. Natl. Acad. Sci. U.S.A., *35* (1949), 94–99.
3. L. J. Mordell, *The number of solutions of some congruences in two variables*, Math. Zeitsch., *37* (1933), 193–209.
4. ——— *On the number of solutions in incomplete residue sets of quadratic congruences*, Archiv der Math., *8* (1957), 153–157.
5. I. M. Vinogradov, *Elements of number theory* (Dover, 1954).
6. A. Weil, *On some exponential sums*, Proc. Natl. Acad. Sci. U.S.A., *34* (1948), 205–207.
7. ——— *Number of solutions of equations in finite fields*, Bull. Amer. Math. Soc., *55* (1949), 497–508.
8. ——— *On the Riemann hypothesis in function fields*, Proc. Natl. Acad. Sci. U.S.A., *27* (1941), 345–347.

University of Toronto