Can. J. Math., Vol. XXXIII, No. 2, 1981, pp. 421–436

ON THE NUMBER AND DISTRIBUTION OF SIMULTANEOUS SOLUTIONS TO DIAGONAL CONGRUENCES

KENNETH W. SPACKMAN

1. Introduction. Two aspects, and their connections, of the problem of enumerating solutions to certain systems of congruences are explored in this paper. Although slightly more general cases are mentioned, the basic object of study is a system of diagonal equations

$$
\begin{array}{c}
a_{11}x_1^{d_1} + a_{12}x_2^{d_2} + \ldots + a_{1t}x_t^{d_t} = 0 \\
\vdots \\
a_{n1}x_1^{d_1} + a_{n2}x_2^{d_2} + \ldots + a_{nt}x_t^{d_t} = 0,
\end{array} \tag{1}
$$

where $d_1, d_2, \ldots, d_t$ are positive integers and the coefficient matrix $[a_{ij}]$ has entries from $F_p = GF(p)$, and for which solutions $\mathbf{x} = (x_1, x_2, \ldots, x_t) \in F_p^{\,t}$ are sought. Speaking loosely, such a system usually has approximately p^{t-n} solutions in the sense that the difference between p^{t-n} and the correct value becomes small in comparison with p^{t-n} as p becomes large. A parameter is introduced which measures the extent to which the matrix $[a_{ij}]$ is non-singular over F_p. The effect of this parameter on the size of the error term (the difference between the number of solutions to (1) and p^{t-n}) is the first aspect of the enumeration problem to be treated. The reader may wish to glance at Theorem 3.2 for the precise formulation of the main result in this direction. Secondly, it is of interest to determine the finest possible partition of the set of elements of $F_p^{\,t}$ having no zero component into (for example) hypercubes of the type

$$
\begin{aligned}
H(\mathbf{a}, b) = \{\mathbf{x} \in F_p^{\,t}\colon a_i \leqq x_i \leqq a_i + b - 1;& \\
0 < a_i < a_i + b - 1 < p; &\quad i = 1, 2, \ldots, t\}
\end{aligned}
$$

so that every such hypercube contains at least one solution to the given system (1). It happens that under modest restrictions there is an approximate regularity of distribution of the solutions to (1) into such hypercubes. The effect of the parametrized non-singularity of $[a_{ij}]$, as well as of p, n and t, on this regularity of distribution is the subject of Theorem

Received August 21, 1979. This research was supported in part by a University of Kentucky Summer Faculty Research Fellowship.

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

4.3. Since a statement of this theorem without many preliminary conventions would be too lengthy, the reader is invited to turn to Theorem 4.3 for a first impression.

The above-mentioned parameter μ of non-singularity serves to generalize a commonly used non-singularity condition:

(i) every non-trivial solution to (1) in the algebraic closure of F_p is non-singular (the matrix $\left[\left(\frac{\partial f_i}{\partial x_j}\right)(\mathbf{x})\right]$ has rank n for every solution $\mathbf{x}$ in the algebraic closure), or equivalently,

(ii) every $n \times n$ submatrix of $[a_{ij}]$ is non-singular, since each of these is equivalent to the condition

(iii) $\mu = 1$.

However, separate arguments are often required to treat the cases $\mu = 1$ and $\mu > 1$. Furthermore, the formulas which are derived to describe the behavior of solutions to (1) when $\mu > 1$ unfortunately do not reduce to the formulas obtained for $\mu = 1$ upon simple substitution of 1 for μ.

The results obtained herein generalize some of the work of J. H. H. Chalk [**1**] who confined his attention to the case $n = 1$. Moreover, the results of this paper demonstrate the effect of the nature of generalization (measured by n and μ) on the behavior of solutions. As an application of the established regularity of distribution, an estimate is made for the size of small solutions to systems of diagonal congruences modulo a prime.

We shall adopt the classical elementary approach to problems of this kind. This involves the expression of the number of solutions to a particular system of equations by an exact formula consisting of exponential (character) sums. That is the easy part; the work consists in selective manipulation and efficient estimation of the relevant sums. In the case at hand a famous inequality of Vinogradov together with a synthesis and extension of existing results including an estimate of Weil for a particular exponential sum, an estimate previously obtained by the author using elementary techniques, and some results of Chalk in the case of a single equation also obtained by classical methods form the structure of the methods employed. In fact, it was the paper of Chalk [**1**] which inspired the distribution aspect of this study. The idea of parametrizing the non-singularity condition arose naturally in an effort to facilitate the estimate of certain error terms, normally occurring as sums themselves, by making each summand approximately the same size. The author first found the value $\mu = 2$ of the non-singularity parameter useful in his Ph.D. dissertation for a crude estimate for the total number of solutions to a system of diagonal equations, and owes the originality of that idea to his thesis director, Professor Wolfgang Schmidt.

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

We shall begin by providing some brief background information in § 2. In § 3 is treated the problem of estimating the total number of solutions to a system of diagonal equations and the dependence of the accuracy of this estimate on the non-singularity parameter already discussed. An example to show that the error term is in a certain sense best possible is included. The regularity of distribution of such solutions is established in § 4 along with some consequences concerning the size of small solutions.

2. Preliminaries. To prevent ambiguities and the necessity of cumbersome parenthetical explanations later, this section has been included as a collection of miscellaneous information and notation.

The cardinality of a finite set S shall be denoted $|S|$. For n-vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \cdot \mathbf{y}$ is the usual dot product $x_1y_1 + \ldots + x_ny_n$ in the field where the x_i and y_i lie. For $x > 0$, $\log x$ shall always mean the natural logarithm of x. Square brackets [ ] are used for two (always obviously distinguishable) concepts; the greatest integer function (e.g. $[\pi] = 3$) and a matrix ($[a_{ij}]$ is the matrix whose entry in the ith row and jth column is a_{ij}). If c is a complex number, $\bar{c}$ always denotes the complex conjugate.

A *character* of a finite abelian group G is a group homomorphism from G to the multiplicative group of complex numbers. The finiteness of G requires that $\theta(g)$ be a root of unity for any character θ and any $g \in G$. The characters of G form a group G', under the operation $(\theta_1\theta_2)(g) = \theta_1(g)\theta_2(g)$, which is isomorphic to G. Denote by θ_0 the identity element of G' (i.e., $\theta_0(g) = 1$ for all $g \in G$). θ_0 is called the *principal character* of G. Two very useful facts are the following:

$$\sum_{g\in G} \theta(g) = \begin{cases} |G| & \text{if } \theta = \theta_0 \\ 0 & \text{if } \theta \neq \theta_0, \end{cases}$$

and (by duality)

$$\sum_{\theta\in G'} \theta(g) = \begin{cases} |G| & \text{if } g = 1 \\ 0 & \text{if } g \neq 1. \end{cases}$$

To a finite field are naturally associated two groups of characters. The group of characters of the additive group of F_q is denoted AF_q and the characters of the (cyclic) multiplicative group are denoted MF_q. Additive characters will always be denoted by ψ (possibly subscripted) and multiplicative characters by χ. We adopt the convention that $\chi(0) = 0$ for $\chi_0 \neq \chi \in MF_q$ and $\chi_0(0) = 1$. If $\psi_0 \neq \psi \in AF_q$, then $\{\psi_\lambda : \lambda \in F_q\} = AF_q$, where $\psi_\lambda(x) = \psi(\lambda x)$. A multiplicative character is said to be *of exponent d* if and only if the order of χ as an element of the group MF_q divides d; that is, $\chi^d(x) = \chi_0(x)$ for all $0 \neq x \in F_q$. When $d|(q - 1)$ there are precisely d multiplicative characters of exponent d given

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

explicitly by

$$\{\chi_\alpha : 0 \leqq \alpha < 1, d\alpha \equiv 0 \ (\mathrm{mod}\ 1)\},$$

where if ϵ is a generator of the multiplicative group of F_q then $\chi_\alpha(\epsilon) = e^{2\pi i\alpha}$. The subgroup of MF_q consisting of characters of exponent $d|(q-1)$ will be denoted M_dF_q. If m is a positive integer and x is real, the abbreviation $e_m(x) = \exp(2\pi ixm^{-1})$ will be useful.

Gaussian sums,

$$G(\chi, \psi) = \sum_{x \in F_q} \chi(x)\psi(x),$$

play a central role in the classical methods for estimating the number of solutions to equations in finite fields. The well-known observations that

$$\begin{aligned} G(\chi, \psi_0) &= 0 && \text{if } \chi \neq \chi_0, \\ G(\chi_0, \psi) &= 0 && \text{if } \psi \neq \psi_0, \quad \text{and} \\ |G(\chi, \psi)| &= q^{1/2} && \text{if } \chi \neq \chi_0 \text{ and } \psi \neq \psi_0 \end{aligned}$$

are presumed to be familiar to the reader. A proof of the following essential fact can be found in [**3**].

Fact. If $\psi_0 \neq \psi \in AF_q$, $0 \neq a \in F_q$ and $d|(q-1)$, then

$$\sum_{y \in F_q} \psi(ay^d) = \sum_{\chi \in M_dF_q} \overline{\chi(a)} G(\chi, \psi).$$

3. The number of unrestricted solutions. Let N denote the number of solutions $\mathbf{x} \in F_q{}^t$ to the system (1) of equations.

LEMMA 3.1.

$$q^nN = \sum_{\mathbf{x} \in F_q{}^t} \sum_{\psi \in (AF_q)^n} \prod_{i=1}^{n} \psi_i(a_{i1}x_1{}^{d_1} + \ldots + a_{it}x_t{}^{d_t}).$$

Proof. This is a routine initial step in the method of character sums and is a special case of Lemma 3.1 of [**4**].

For each natural number μ, call an $n \times t$ matrix over F_q *μ-weakly non-singular* if and only if for each natural number k satisfying

$$\mu(k-1) + 1 \leqq \min\{t, \mu(n-1) + 1\}$$

the matrix has the property that among any $\mu(k-1) + 1$ column vectors there are at least k F_q-linearly independent ones. Notice that an $n \times t$ μ-weakly non-singular matrix for which $t \leqq \mu(n-1)$ may have rank smaller than n, and it may not. In the case that the coefficient matrix of system (1) has rank less than n, a new system having the same solutions as the original one could be formed by deleting one or more of the equations. Since one cannot reasonably expect to obtain sharp

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

estimates for N (in terms of n) in such a situation, we shall henceforth assume t to be large in comparison with n; in particular, we suppose $t > \mu(n - 1)$. A μ-weakly non-singular matrix of dimensions satisfying $t > \mu(n - 1)$ clearly has rank n and, furthermore, it is obvious that if $\lambda \leqq \mu$ are natural numbers, then a λ-weakly non-singular matrix is also μ-weakly non-singular. If $t > \mu(n - 1)$, a 1-weakly non-singular matrix is in some sense as non-singular as it can be; every $n \times n$ submatrix is non-singular.

In the present context, three cases distinguish themselves:

(i) $n = 1, \mu = 1$;
(ii) $n \geqq 2, \mu = 1$;
(iii) $\mu \geqq 2$.

For each case suppose that n, t and μ are natural numbers with $t > \mu(n - 1)$ and that the $n \times t$ coefficient matrix of system (1) is μ-weakly non-singular. For case (i), Weil [**6**] proved

$$N = q^{t-1} + O(q^{t/2})$$

using methods involving character sums. For case (ii), the author [**4**] has proved

$$N = q^{t-n} + O(q^{(t-1)/2})$$

using elementary methods in the spirit of the work of Weil and others. The third case is the subject of the following theorem.

THEOREM 3.2. *Let n, t and μ be natural numbers satisfying $\mu \geqq 2$ and $t > \mu(n - 1)$. The number N of solutions to the system* (1) *whose coefficient matrix is μ-weakly non-singular over F_q satisfies*

$$|N - q^{t-n}| \leqq (d_1 - 1) \dots (d_t - 1)(2^t - 1)q^{(t+(\mu-2)(n-1))/2}.$$

Proof. There is no loss of generality in assuming that each exponent d_i is a divisor of $q - 1$, since the number of solutions in F_q to the equation $x^d = y$ for fixed $y \in F_q$ and $d \in \mathbf{N}$ is precisely the same as the number of solutions to $x^h = y$ where $h = \mathrm{GCD}(d, q - 1)$. Moreover, if the estimate of the theorem holds with the greatest common divisors in the constant factor of the error term, it certainly holds with the original d's.

If ψ is any fixed non-principal additive character on F_q, it follows from Lemma 3.1 and some remarks of § 2 that

$$(2) \quad q^n N = \sum_{\mathbf{x} \in F_q{}^t} \sum_{\lambda \in F_q{}^n} \prod_{j=1}^{n} \psi(\lambda_i a_{i1} x_1{}^{d_1} + \dots + \lambda_i a_{it} x_t{}^{d_t})$$

$$= \sum_{\lambda \in F_q{}^n} \prod_{j=1}^{t} \sum_{x_j \in F_q} \psi(L_j(\lambda) x_j{}^{d_j}),$$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

where

$$L_j(\lambda) = \sum_{i=1}^{n} \lambda_i a_{ij}, \quad (1 \leqq j \leqq t).$$

For each u-subset $(0 \leqq u \leqq t)$ U of $T = \{1, 2, \ldots, t\}$, set

$$\Lambda_U = \{\lambda \in F_q^n : \lambda \neq \mathbf{0} \text{ and } L_j(\lambda) \neq 0 \text{ if and only if } j \in U\}.$$

Then by separating the summand of expression (2) for which $\lambda = \mathbf{0}$ and associating the remaining terms according to the parameter u, $0 \leqq u \leqq t$, which measures the number of indices j having $L_j(\lambda) \neq 0$, we obtain

$$q^n N = q^t + \sum_{u=0}^{t} q^{t-u} \sum_{\substack{U \subseteq T \\ |U|=u}} \sum_{\lambda \in \Lambda_U} \prod_{j \in U} \sum_{x_j \in F_q} \psi(L_j(\lambda) x_j^{d_j}).$$

The idea is to subtract q^t from both sides of this equation, divide through by q^n, take absolute values and use the triangle inequality. We are therefore interested in the size and number of summands in the triple sum above. To determine their sizes we introduce Gaussian sums according to the preliminary remarks of the previous section. Namely, for a fixed u-subset U, $\lambda \in \Lambda_U$ and $j \in U$,

$$\sum_{x_j \in F_q} \psi(L_j(\lambda) x_j^{d\cdot}) = \sum_{\chi_j \in M_{d_j} F_q} \overline{\chi_j}(L_j(\lambda)) G(\chi_j, \psi).$$

The principal multiplicative character has exponent d_j but contributes zero to this sum, so the right hand side is a sum of $d_j - 1$ terms each of modulus $q^{1/2}$. Hence

$$|N - q^{t-n}| \leqq \sum_{u=0}^{t} q^{t-n-u} \sum_{\substack{U \subseteq T \\ |U|=u}} \sum_{\lambda \in \Lambda_U} (d_1 - 1) \cdots (d_t - 1) q^{u/2}.$$

Now observe that if $0 \leqq u < t - \mu(n-1)$ (recall the hypothesis that $t > \mu(n-1)$), Λ_U is empty for all u-subsets U. For otherwise there would exist $\lambda \neq \mathbf{0}$ and at least $\mu(n-1) + 1$ zero column sums ($L_j(\lambda) = 0$ for at least $\mu(n-1) + 1$ indices), contrary to the hypothesis that among these columns occur n linearly independent ones. Hence $t - u \leqq \mu(n-1)$. Fix u. Then if k is the largest integer such that $\mu(k-1) + 1 \leqq t - u$, one has $t - u \leqq \mu k$ so that among any $t - u$ columns there are always at least $(t-u)/\mu$ linearly independent ones. It follows that for a fixed u-subset U, the number of $\lambda \in \Lambda_U$ is at most $q^{n-(t-u)/\mu}$. Finally since the number of u-subsets U of T is the binomial coefficient $\binom{t}{u}$,

$$\begin{aligned} |N - q^{t-n}| &\leqq (d_1 - 1) \cdots (d_t - 1) \sum_{u=t-\mu(n-1)}^{t} q^{t-n-u} \binom{t}{u} q^{n-(t-u)/\mu} q^{u/2} \\ &= (d_1 - 1) \cdots (d_t - 1) q^{t(\mu-1)/\mu} \sum_{u=t-\mu(n-1)}^{t} \binom{t}{u} q^{u(2-\mu)/2\mu}. \end{aligned}$$

But since $\mu \geqq 2$, $(2-\mu)/2\mu \leqq 0$, so for each u under consideration,

$$q^{u(2-\mu)/2\mu} \leqq q^{(t-\mu(n-1))(2-\mu)/2\mu}.$$

Hence

$$|N - q^{t-n}| \leqq (d_1 - 1)\cdots(d_t - 1)q^{t(\mu-1)/\mu}q^{(t-\mu(n-1))(2-\mu)/2\mu}$$
$$\times \sum_{u=t-\mu(n-1)}^{t} \binom{t}{u} \leqq (d_1 - 1)\cdots(d_t - 1)(2^t - 1)q^{(t+(\mu-2)(n-1))/2}.$$

It is often convenient to suppress the details of the constant error factor in the above formula for N; one simply writes

$$N = q^{t-n} + O(q^{(t+(\mu-2)(n-1))/2}).$$

The exponent in the error term is best possible without any further hypothesis as the following example shows.

Example. Let $\mu > 2$ be even, $n \geqq 1$, $t = n\mu$, and $d_1 = d_2 = \ldots = d_t = 2$. Define the matrix $[a_{ij}]$ of coefficients over $GF(p)$, $p \equiv 1 \pmod 4$, by

$$a_{ij} = \begin{cases} 1 & \text{if } (i-1)\mu + 1 \leqq j \leqq i\mu \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leqq i \leqq n$ and $1 \leqq j \leqq t$. The coefficient matrix looks as follows.

$$\begin{bmatrix} \underbrace{1\,1\ldots 1}_{\mu} & & & 0 \\ & \underbrace{1\,1\ldots 1}_{\mu} & \ldots & \\ 0 & & & \underbrace{1\,1\ldots 1}_{\mu} \end{bmatrix}$$

That is, we ask for the number N of t-tuples over $GF(p)$ making n diagonal quadratic forms, each in μ variables, each variable appearing in exactly one form, simultaneously zero. A pigeonhole argument shows that the matrix $[a_{ij}]$ is μ-weakly non-singular. Further, it is not difficult to show by the method of Gaussian sums (or by other means; e.g., see [**3**, p. 145–147]) that the number of solutions to the equation

$$x_1^2 + x_2^2 + \ldots + x_\mu^2 = 0$$

is, for μ even and $p \equiv 1 \pmod 4$, exactly

$$p^{\mu-1} + p^{\mu/2} - p^{\mu/2-1}.$$

Hence the number of solutions to the system (1) under these circumstances is

$$N = (p^{\mu-1} + p^{\mu/2} - p^{\mu/2-1})^n.$$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

Since $\mu > 2$,

$$N = p^{n\mu - n} + np^{(n-1)(\mu-1)+\mu/2} + \text{(terms involving } p \text{ to a lower exponent)}$$
$$= p^{t-n} + np^{(t+(\mu-2)(n-1))/2} + \text{(terms involving } p \text{ to a lower exponent).}$$

Therefore in this example,

$$N - p^{t-n} \sim np^{(t+(\mu-2)(n-1))/2}$$

as $p \rightarrow \infty$ through primes congruent to 1 modulo 4.

Under the same hypotheses on μ, n, t and the matrix $[a_{ij}]$ as were made in Theorem 3.2, the non-homogeneous system

$$\begin{array}{ccc} a_{11}x_1^{d_1} + \ldots + a_{1t}x_t^{d_t} & = & b_1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{n1}x_1^{d_1} + \ldots + a_{nt}x_t^{d_t} & = & b_n \end{array} \tag{3}$$

(having, say, N_1 solutions) may now be treated in a routine way outlined as follows. If no assumptions whatsoever are made on the elements $b_1, \ldots, b_n$ (so that the homogeneous case is included), then inspection of the system

$$\begin{array}{ccc} a_{11}x_1^{d_1} + \ldots + a_{1t}x_t^{d_t} - b_1x_{t+1}^{q-1} & = & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{n1}x_1^{d_1} + \ldots + a_{nt}x_t^{d_t} - b_nx_{t+1}^{q-1} & = & 0 \end{array} \tag{4}$$

(having M solutions, let's say) yields the relation

$$M = N + (q-1)N_1.$$

But application of the present method to system (4) also produces

$$M = q^{t+1-n} + O(q^{(t+2+(\mu-2)(n-1))/2}).$$

The details are omitted because they so closely resemble those of the homogeneous case, but the idea is to consider separately the cases when U contains the column index $t+1$ and when it does not. The arguments of the proof of Theorem 3.2 are then amended appropriately to give the stated result. Together with

$$N = q^{t-n} + O(q^{(t+(\mu-2)(n-1))/2})$$

the previous two equations imply that

$$N_1 = q^{t-n} + O(q^{(t+(\mu-2)(n-1))/2}).$$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

If, on the other hand, it happens that the augmented matrix

$$\begin{bmatrix} a_{11} \dots a_{1t} & b_1 \\ \cdot \quad\quad \cdot & \cdot \\ \cdot \quad\quad \cdot & \cdot \\ \cdot \quad\quad \cdot & \cdot \\ a_{n1} \dots a_{nt} & b_n \end{bmatrix}$$

is also μ-weakly non-singular, then by consideration of the system

$$\begin{array}{l} a_{11}y_1^{d_1} + \dots + a_{1t}y_t^{d_t} - b_1 y_{t+1}^{d} = 0 \\ \cdot \qquad\qquad\quad \cdot \qquad\qquad\quad \cdot \\ \cdot \qquad\qquad\quad \cdot \qquad\qquad\quad \cdot \\ \cdot \qquad\qquad\quad \cdot \qquad\qquad\quad \cdot \\ a_{n1}y_1^{d_1} + \dots + a_{nt}y_t^{d_t} - b_n y_{t+1}^{d} = 0 \end{array}$$

(having M solutions), where $d = \mathrm{LCM}[d_1, d_2, \dots, d_t]$, for which

$$M = N + (q - 1)N_1,$$

we have by Theorem 3.2,

$$M = q^{t+1-n} + O(q^{(t+1+(\mu-2)(n-1))/2}).$$

Since also

$$N = q^{t-n} + O(q^{(t+(\mu-2)(n-1))/2}),$$

we conclude that

$$N_1 = q^{t-n} + O(q^{(t-1+(\mu-2)(n-1))/2}).$$

4. The regularity of solution distribution.

The estimates of §3 can be used to obtain asymptotic formulas for the distribution of simultaneous solutions to congruences modulo a prime and to deduce bounds for the size of small simultaneous solutions. For this purpose a theorem of [**1**], reproduced here for the convenience of the reader, is fundamental. Actually, Chalk's theorem is slightly more general than the following version. For the very short proof, consult [**1**].

Let $m \geqq 2$ be an integer. Let $C = C(m)$ denote the cube consisting of all integral t-tuples $(x_1, x_2, \dots, x_t)$ satisfying $0 \leqq x_i < m$ $(i = 1, 2, \dots t)$.

THEOREM. (Chalk). *Let S be any subset of C and let $\varphi(\mathbf{x})$ be a complex-valued function defined on C which satisfies*

$$\left| \sum_{\mathbf{x} \in C} \varphi(\mathbf{x}) e_m(-\mathbf{x} \cdot \mathbf{y}) \right| \leqq \Phi \quad (\mathbf{0} \neq \mathbf{y} \in C) \tag{5}$$

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

$$\sum_{\mathbf{x} \in S}{}' \varphi(\mathbf{x}) = m^{-t}|S| \sum_{\mathbf{x} \in C} \varphi(\mathbf{x}) + \theta m^{-t} \Phi E_m(S) \tag{6}$$

where $|\theta| \leqq 1$ *and where*

$$E_m(S) = \sum_{0 \neq \mathbf{y} \in C} \left| \sum_{\mathbf{z} \in S} e_m(\mathbf{y} \cdot \mathbf{z}) \right| .$$

It will shortly become apparent that with an appropriate choice for the function φ, Eq. (6) takes the form of an asymptotic formula for the number of solutions to a system of equations over $GF(p)$. If S is of a certain type, then an estimate of Vinogradov provides information about $E_m(S)$. To complete an estimation of the size of the error term, information concerning the size of Φ is required; in fact, this will be our main concern and it will require the estimates derived in § 3.

Our interest lies in systems of congruences

$$(7) \quad \begin{array}{lll} f_1(\mathbf{x}) & \equiv 0 & \pmod{m} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ f_n(\mathbf{x}) & \equiv 0 & \pmod{m}. \end{array}$$

Let $S \subseteq C = C(m)$ and let $\Lambda = \Lambda(m)$ consist of all integral n-tuples $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ having $0 \leqq \lambda_i < m$ $(i = 1, 2, \ldots, n)$. Define

$$\varphi(\mathbf{x}) = \sum_{\lambda \in \Lambda} \prod_{i=1}^{n} e_m(\lambda_i f_i(\mathbf{x})) = \sum_{\lambda \in \Lambda} e_m(\lambda \cdot \mathbf{f}(\mathbf{x})).$$

The following lemma provides the link between Chalk's theorem and the number $N(S)$ of $\mathbf{x} \in S$ which are solutions to the system (7).

Lemma 4.1. *If S is any subset of C, then*

$$\sum_{\mathbf{x} \in S} \varphi(\mathbf{x}) = m^n N(S).$$

Proof. Since among the points $\mathbf{x} \in S$, $\varphi(\mathbf{x}) = 0$ unless $\mathbf{x}$ is a solution for (7) in which case $\varphi(\mathbf{x}) = |\Lambda| = m^n$, the result follows.

Hence if condition (5) is satisfied with this choice of φ, then by Chalk's theorem

$$m^n N(S) = m^{-t}|S| m^n N(C) + \theta m^{-t} \Phi E_m(S),$$

or more simply,

$$(8) \quad N(S) = m^{-t}|S| N(C) + \theta m^{-t-n} \Phi E_m(S),$$

where $|\theta| \leqq 1$. We shall make assumptions to ensure the second term on the right is "small" so that, roughly, the ratio $N(S)/N(C)$ will be approximately the same as the ratio $|S|/|C|$; certainly a desirable situation. It is perhaps worth noting that when $S = C$, Eq. (8) is a tautology since $|C| = m^t$ and $E_m(C) = 0$. Also $E_m(\emptyset) = 0$ so Eq. (8) gives the

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

correct statement $N(\emptyset) = 0$ when $S = \emptyset$. Of course, in the special case of diagonal equations we have estimates for $N(C)$ when $m = p$ by § 3; in fact, we will need them to estimate $N(S)$ too.

To permit a good estimate for $E_m(S)$ when S is reasonably large ($|S| > (\log m)^t$), we assume S is an arbitrary, but fixed, box in C of the form

$$(9)\quad S = S(\mathbf{a}, \mathbf{b}) = \{\mathbf{x} \in C\colon 0 \leqq a_i \leqq x_i \leqq a_i + b_i - 1 < m;\ i = 1, 2, \ldots, t\},$$

where $b_i > 1$ for $i = 1, 2, \ldots, t$. Then evidently $|S| = b_1 b_2 \ldots b_t$ and a trivial estimate for $E_m(S)$ is

$$E_m(S) < b_1 b_2 \ldots b_t m^t.$$

When $|S| > (\log m)^t$ and m is sufficiently large, a better estimate is available by use of a famous inequality of Vinogradov (see [**5**]):

$$\sum_{\lambda=1}^{m-1} \left| \sum_{x=a}^{a+b-1} e_m(\lambda x) \right| < m \log m - \delta m.$$

where δ is an absolute constant. Chalk [**1**] used Vinogradov's inequality to obtain the bound

$$E_m(S) < (m \log m)^t,$$

provided $m \geqq 60$. Notice the uniformity of this bound in $b_1, b_2, \ldots, b_t$ which may each be considerably larger than $\log m$.

Put

$$F(\mathbf{y}) = \sum_{\mathbf{x} \in C} \varphi(\mathbf{x}) e_m(-\mathbf{x} \cdot \mathbf{y}).$$

In order to apply Chalk's theorem, we require a bound for $|F(\mathbf{y})|$ which is uniform in $\mathbf{0} \neq \mathbf{y} \in C$. Using the definition for $\varphi(\mathbf{x})$ one obtains

$$F(\mathbf{y}) = \sum_{\mathbf{x} \in C} \sum_{\lambda \in \Lambda} e_m(\lambda \cdot \mathbf{f}(\mathbf{x})) e_m(-\mathbf{x} \cdot \mathbf{y}).$$

If we set

$$S_\lambda(\mathbf{y}) = \sum_{\mathbf{x} \in C} e_m(\lambda \cdot \mathbf{f}(\mathbf{x}) - \mathbf{x} \cdot \mathbf{y}),$$

then by noting that $S_{\mathbf{0}}(\mathbf{y}) = 0$ if $\mathbf{y} \neq \mathbf{0}$, one sees that

$$F(\mathbf{y}) = \sum_{\mathbf{0} \neq \lambda \in \Lambda} S_\lambda(\mathbf{y}).$$

It is implicit that $F(\mathbf{y})$ depends upon the integer m, the set of functions $f_1, f_2, \ldots, f_n$, and upon the set S. Henceforth, we shall assume that $m = p$ is a rational prime, that

$$f_i(\mathbf{x}) = a_{i1} x_1^{d_1} + \ldots + a_{it} x_t^{d_t}$$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

where $a_{ij} \in GF(p)$ and $d_j | (p - 1)$ $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, t)$, and that S is the box given by (9) with $|S| > (\log m)^t$. Should it become necessary or convenient to vary p while holding the d_j fixed, then it shall be understood that p is to vary among primes $p \equiv 1 \pmod{\mathrm{LCM}[d_1, d_2, \ldots, d_t]}$. Whenever $F(\mathbf{y})$ is now written, these conventions are understood.

LEMMA 4.2. *Let n, t and μ be natural numbers satisfying $t > \mu(n - 1)$. If the coefficient matrix $[a_{ij}]$ is μ-weakly non-singular, $d_j \geqq 2$ $(j = 1, 2, \ldots, t)$, and if $\mathbf{0} \neq \mathbf{y} \in C = C(p)$, then*

$$F(\mathbf{y}) = O(p^{t/2+n}) \qquad \text{if } \mu = 1, \text{ and,}$$
$$F(\mathbf{y}) = O(p^{(t+\mu(n-1))/2+1}) \qquad \text{if } \mu \geqq 2.$$

The implied constants depend only on n, t, $d_1, d_2, \ldots, d_t$ (and not on p, $\mathbf{y}$, $\mathbf{b}$, or μ).

Proof. Put $\Lambda = \Lambda(p)$ and for $\lambda \in \Lambda$ put (just as in § 3)

$$L_j(\lambda) = \sum_{i=1}^{n} \lambda_i a_{ij} \quad (j = 1, 2, \ldots, t).$$

Then if $\mathbf{0} \neq \mathbf{y} \in C$,

$$F(\mathbf{y}) = \sum_{\mathbf{0} \neq \lambda \in \Lambda} S_\lambda(\mathbf{y})$$
$$= \sum_{\mathbf{0} \neq \lambda \in \Lambda} \sum_{\mathbf{x} \in C} \prod_{j=1}^{t} e_p(L_j(\lambda) x_j^{d_j} - x_j y_j)$$
$$= \sum_{\mathbf{0} \neq \lambda \in \Lambda} \prod_{j=1}^{t} \sum_{x_j=0}^{p-1} e_p(L_j(\lambda) x_j^{d_j} - x_j y_j).$$

Again as in § 3 for each subset $U \subseteq T = \{1, 2, \ldots, t\}$, set

$$\Lambda_U = \{\lambda \in \Lambda : L_j(\lambda) \not\equiv 0 \pmod{p} \text{ if and only if } j \in U\}.$$

Also for each choice of $\mathbf{y}$ and λ and any subset V of T, define

$$P_V(\lambda, \mathbf{y}) = \prod_{j \in V} \sum_{x_j=0}^{p-1} e_p(L_j(\lambda) x_j^{d_j} - x_j y_j).$$

Then

$$F(\mathbf{y}) = \sum_{u=1}^{t} \sum_{\substack{U \subseteq T \\ |U| = u}} \sum_{\lambda \in \Lambda_U} P_U(\lambda, \mathbf{y}) P_{T-U}(\lambda, \mathbf{y}).$$

First observe that the number of subsets U of T having cardinality u $(1 \leqq u \leqq t)$ is the binomial coefficient $\binom{t}{u}$. Next, notice that for a fixed u-subset U, the number of $\lambda \in \Lambda_U$ is at most $p^{n-(t-u)/\mu}$, for we have seen that the assumptions that $[a_{ij}]$ is μ-weakly non-singular and $t > \mu(n - 1)$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

imply that among any $t - u$ linear conditions (from the columns of the matrix $[a_{ij}]$) on λ there must be at least $(t - u)\mu$ linearly independent ones. The product $|P_U(\lambda, \mathbf{y})|$ may be bounded above independently of $\mathbf{y}$, if $\lambda \in \Lambda_U$, by an estimate of Weil (for a proof of Weil's estimate, see [**3**, Corollary 2F, p. 45]); namely, if p is prime, $g(X) = a_m X^m + \ldots + a_0 \in \mathbf{Z}[X]$ with $0 < m < p$ and $p \nmid a_m$, then

$$\left|\sum_{x=0}^{p-1} e_p(g(x))\right| \leqq (m - 1)p^{1/2}.$$

Specifically, since $2 \leqq d_j < p$ (recall $d_j|(p - 1)$) and since for a u-subset $U \subseteq T$ and $\lambda \in \Lambda_U$, $L_j(\lambda) \not\equiv 0 \pmod{p}$,

$$|P_U(\lambda, \mathbf{y})| \leqq \prod_{j\in U} (d_j - 1)p^{1/2} \leqq p^{u/2} \prod_{j=1}^{t} (d_j - 1).$$

For the remaining product we make the trivial estimate,

$$|P_{T-U}(\lambda, \mathbf{y})| \leqq p^{t-u}.$$

These observations permit one to write for all $\mathbf{y} \in C$,

$$\begin{aligned} |F(\mathbf{y})| &\leqq \sum_{u=t-\mu(n-1)}^{t} \binom{t}{u} p^{n-(t-u)/\mu} \left(\prod_{j=1}^{t} (d_j - 1)\right) p^{u/2} p^{t-u} \\ &= \left(\prod_{j=1}^{t} (d_j - 1)\right) \sum_{u=t-\mu(n-1)}^{t} \binom{t}{u} p^{n+((\mu-1)/\mu)t+((2-\mu)/2\mu)u}. \end{aligned}$$

Two cases distinguish themselves according to the sign of the parameter $(2 - \mu)/2\mu$. If $\mu = 1$, $(2 - \mu)/2\mu = 1/2$ whence

$$F(\mathbf{y}) = O(p^{t/2+n}),$$

because $u \leqq t$. If $\mu \geqq 2$, then $(2 - \mu)/2\mu \leqq 0$ and so

$$F(\mathbf{y}) = O(p^{n+((\mu-1)/\mu)t+((2-\mu)/2\mu)(t-\mu(n-1))})$$

since $u \geqq t - \mu(n - 1)$. Simplification of the above exponent gives the result stated in the lemma.

THEOREM 4.3. *Let n, t, μ, $d_1, d_2, \ldots, d_t$ be natural numbers with $t > \mu(n - 1)$ and $d_j \geqq 2$ $(j = 1, 2, \ldots, t)$. If $[a_{ij}]$ is μ-weakly non-singular over F_p, then the number $N(S)$ of solutions to the system* (1) *which lie in the box S given by* (9) *satisfies*

$$N(S) = p^{-n}|S| + O(p^{t/2}(\log p)^t) \text{ if } \mu = 1, \text{ and}$$

$$N(S) = p^{-n}|S| + O(p^{(t+(\mu-2)(n-1))/2}(\log p)^t) \text{ if } \mu \geqq 2.$$

Proof. The proof consists of the replacement of $N(C)$ and Φ in Eq. (8) by estimates obtained in Theorem 3.2 (or by Weil or Spackman in case $\mu = 1$) and Lemma 4.2, respectively, using the Chalk–Vinogradov bound

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

for $E_p(S)$. If $\mu = 1$, one obtains

$$\begin{aligned} N(S) &= p^{-t}|S|(p^{t-n} + O(p^{t/2})) + O(p^{-t-n}p^{t/2+n}(p \log p)^t) \\ &= p^{-n}|S| + O(|S|p^{-t/2}) + O(p^{t/2}(\log p)^t). \end{aligned}$$

Since $|S| \leqq p^t$, the second error term dominates the first for large p, so the first result follows. Next suppose $\mu \geqq 2$. Then

$$\begin{aligned} N(S) &= p^{-t}|S|(p^{t-n} + O(p^{(t+(\mu-2)(n-1))/2})) \\ &\qquad\qquad + O(p^{-t-n}(p \log p)^t p^{(t+\mu(n-1))/2+1}) \\ &= p^{-n}|S| + O(|S|p^{(-t+(\mu-2)(n-1))/2}) + O(p^{(t+(\mu-2)(n-1))/2}(\log p)^t). \end{aligned}$$

Again the trivial bound $|S| \leqq p^t$ makes the second error term the dominant one and the proof is complete.

Remarks. (i) The same arguments go through for an inhomogeneous system of diagonal equations (3) since by the remarks concluding § 3, we have a no less accurate estimate for $N(C)$ while neither the estimate for $E_p(S)$ nor $|F(\mathbf{y})|$ need be altered. Theorem 4.3 therefore generalizes a theorem of [**1**, Theorem 2] which treated the case $n = 1$ exclusively.

(ii) Observe that when $n = 1$, the property that the $1 \times t$ coefficient matrix is μ-weakly non-singular coincides for every μ with the property that each coefficient is non-zero modulo p. In this light, it is natural that the estimates of Theorem 4.3 should agree and be independent of μ when $n = 1$.

(iii) Although in general the property "2-weakly non-singular" is weaker than the property "1-weakly non-singular," the (asymptotic) magnitude of the error terms in Theorem 4.3 coincide for all n in the cases $\mu = 1$ and $\mu = 2$. It appears that any improvement should arise in the estimation of $F(\mathbf{y})$ in the case $\mu = 1$.

(iv) Finally, observe that for fixed n and μ and for sufficiently large p and t, the estimates of Theorem 4.3 are genuine asymptotic formulas in that the "main term" dominates the "error term". The question in the case $\mu = 1$ is one of existence of a box $S \subseteq C$ with

$$|S| \geqq c_1 p^{t/2+n}(\log p)^t,$$

for if such an S exists then with a suitable choice for the constant c_1, $p^{-n}|S|$ dominates the error term. But if $t/2 + n < t$ (i.e., $t > 2n$), then

$$c_1 p^{t/2+n}(\log p)^t = o(p^t) \quad \text{as } p \rightarrow +\infty,$$

and the choice of such an S is possible. Similarly, if $\mu \geqq 2$ and $t > \mu(n-1) + 2$ (exactly the condition required to make the estimate for $N(C)$ of Theorem 3.2 meaningful) then for any constant c_2,

$$c_2 p^{(t+(\mu-2)(n-1))/2+n}(\log p)^t = o(p^t) \quad \text{as } p \rightarrow +\infty.$$

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

In particular, if c_2 is larger than the implied constant in the case $\mu \geqq 2$ of Theorem 4.3, one may choose (for large enough p) a box $S \subseteq C$ so that $p^{-n}|S|$ dominates the error term provided $t > \mu(n-1)+2 \geqq 2n$.

An immediate consequence of Theorem 4.3 is the existence of small solutions to (1) having non-zero components.

COROLLARY 4.4. *If* $[a_{ij}]$ *is* μ*-weakly non-singular over* F_p, $d_j \geqq 2$, *and* $t > \mu(n-1)+2 \geqq 2n$, *then there exists for sufficiently large* p *a solution* $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ *for the system* (1) *having*

$$1 \leqq x_i \leqq kp^{1/2+(\mu(n-1)+2)/2t} \log p \quad (1 \leqq i \leqq t),$$

for some constant k *depending only on* $d_1, d_2, \ldots, d_t, t, n$.

Proof. The condition $\mu(n-1)+2 \geqq 2n$ is equivalent to $(\mu-2) \times (n-1) \geqq 0$. Chalk [**1**] has proved this result for $n = 1$, so we may assume $\mu \geqq 2$. Choose S to be the box defined by (9) with $a_i = 1$ and

$$b_i = [kp^{(t+(\mu-2)(n-1))/2t+n/t} \log p]$$

for each $i = 1, 2, \ldots, t$, where k is a constant to be determined. Then

$$\begin{aligned}|S| &\geqq (kp^{(t+(\mu-2)(n-1))/2t+n/t} \log p - 1)^t \\ &\geqq (k-1)^t p^{(t+(\mu-2)(n-1))/2+n} (\log p)^t.\end{aligned}$$

Choose k so large that $(k-1)^t$ is larger than the implied constant in the estimate for the case $\mu \geqq 2$ of Theorem 4.3. Then by Theorem 4.3, $N(S) > 0$. That is, there exists a solution $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ satisfying

$$1 \leqq x_i \leqq b_i \leqq kp^{1/2+(\mu(n-1)+2)/2t} \log p < p,$$

the last inequality holding for sufficiently large p since $\mu(n-1)+2 < t$.

The case $\mu = 1$ is treated analogously.

COROLLARY 4.5. *If every* $n \times n$ *submatrix of* $[a_{ij}]$ *is non-singular over* F_p, $d_j \geqq 2$ *and* $t > 2n$, *then for sufficiently large* p *there exists a solution* $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ *to the system* (1) *having*

$$1 \leqq x_i \leqq kp^{1/2+n/t} \log p \qquad (1 \leqq i \leqq t),$$

for some constant k *depending only on* $d_1, d_2, \ldots, d_t, t, n$.

Proof. Let S be of the form (9) with $a_i = 1$ and

$$b_i = [kp^{1/2+n/t} \log p],$$

for each $i = 1, 2, \ldots, t$, where k is to be determined. Then

$$|S| \geqq (k-1)^t p^{t/2+n} (\log p)^t.$$

Hence if k and p are sufficiently large, Theorem 4.3 guarantees the existence of a solution to (1) having

$$1 \leqq x_i \leqq b_i \leqq kp^{1/2+n/t} \log p < p.$$

provided $t > 2n$.

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press

Both Corollaries 4.4 and 4.5 generalize the result obtained by Chalk [**1**] for the case $n = 1$ regardless of the value of μ (recall Remark (ii) above). Furthermore, these two corollaries can be combined and reformulated in a weaker but slightly more elegant version as follows.

COROLLARY 4.6. *Fix natural numbers n and μ and let $\epsilon > 0$ be given. Let $d_j \geqq 2$ and suppose the $n \times t$ matrix $[a_{ij}]$ is μ-weakly non-singular over F_p. If p and t are sufficiently large then there exists a solution $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ for the system* (1) *having*

$$1 \leqq x_i \leqq p^{1/2+\epsilon} \quad (1 \leqq i \leqq t).$$

Proof. Take $t > 3\epsilon^{-1}(\mu(n - 1) + 2)$ so that in both Corollaries 4.4 and 4.5 there exists a constant k such that at least one solution lies in the box

$$1 \leqq x_i \leqq kp^{1/2+\epsilon/3} \log p \quad (1 \leqq i \leqq t)$$

for sufficiently large p. (The conditions $t > 2n$ and $t > \mu(n - 1) + 2$ are satisfied in the respective cases $\mu = 1$ and $\mu \geqq 2$ since we may assume $\epsilon \leqq 1/2$.) Now by taking p so large that $\log p \leqq p^{\epsilon/3}$ and $k \leqq p^{\epsilon/3}$, the result follows.

The ideas of the present section can be readily extended to include systems of diagonal equations over arbitrary finite fields. In view of the fact that § 3 was carried out in this generality and that the derivation of the estimate for $F(\mathbf{y})$ remains valid in any finite field $GF(q)$ (as long as $d_j|(q - 1)$), the only possible obstacles might arise with an attempt to generalize the concept of a "box" or with the Chalk–Vinogradov estimate for $E_m(S)$. Chalk and Williams [**2**] have resolved these difficulties so that it becomes routine to verify the counterparts of these estimates in the more general case.

REFERENCES

1. J. H. H. Chalk, *The number of solutions of congruences in incomplete residue systems*, Can. J. Math. *15* (1963), 291–296.

2. J. H. H. Chalk and K. S. Williams, *The distribution of solutions to congruences*, Mathematika *12* (1965), 176–192.

3. W. M. Schmidt, *Equations over finite fields. An elementary approach*, Lecture Notes in Mathematics *536* (Springer-Verlag, Berlin, 1976).

4. K. W. Spackman, *Simultaneous solutions to diagonal equations over finite fields*, J. Number Theory *11* (1979), 100–115.

5. I. M. Vinogradov, *Elements of number theory* (Dover, New York, 1954).

6. A. Weil, *Numbers of solutions of equations in finite fields*, Bull. Amer. Math. Soc. *55* (1949), 497–508.

University of Kentucky,
Lexington, Kentucky

https://doi.org/10.4153/CJM-1981-037-x Published online by Cambridge University Press