ACTA ARITHMETICA
V (1959)

Note on a problem of Chowla

by

B. J. Birch and H. P. F. Swinnerton-Dyer (Cambridge)

1. In a lecture given to the American Mathematical Society some years ago [2], Chowla listed a number of problems connected with the zeta-function; one of these was to estimate the number of values taken by a polynomial in a finite field. Precisely, let $f(x)$ be a polynomial of degree d defined over the finite field k with q elements, and let $N(f)$ be the number of distinct values of y in k for which at least one of the roots of

$$f(x) = y \tag{1}$$

is in k; then we must estimate $N(f)$, at least for "general" polynomials $f(x)$.

It turns out that we can express $N(f)$ in terms of the numbers of points on a certain finite set of curves, with an error which is bounded independently of q; and we can estimate these numbers by using the well-known results of Weil [5]. We thus obtain

Theorem 1. *For "general" polynomials $f(x)$ we have*

$$N(f) = q\left\{1-\frac{1}{2}+\frac{1}{3!}-\ldots-(-1)^d\frac{1}{d!}\right\}+O(q^{1/2}), \tag{2}$$

where the constant implied depends only on d.

The principal term in (2) had already been conjectured by Chowla and others.

We have now to explain what we mean by "general". Write $G(f)$ for the Galois group of the equation (1) over $k(y)$, and $G^+(f)$ for its Galois group over $k^+(y)$, where k^+ is the algebraic closure of k. We can identify $G^+(f)$ as a subgroup of $G(f)$. It will appear from the proof that Theorem 1 holds for all polynomials $f(x)$ such that $G^+(f)$ is the symmetric group on d elements: more generally we have

THEOREM 2.

$$N(f) = \lambda q + O(q^{1/2}),$$

where λ depends only on $G(f)$, $G^+(f)$ and d.

The proof of Theorem 2 provides an explicit method of calculating λ.

It is well known to anyone who has tried it that finding the Galois group of an equation is by no means straightforward. We therefore obtain simple sufficient conditions for $G^+(f)$ to be the full symmetric group. Though not necessary, these are in fact satisfied by almost all polynomials. We give also some illustrative examples.

The remarks above require the polynomial $f(x)-y$ to be separable over $k(y)$. This is so unless $f(x)$ is a polynomial in x^p, where p is the characteristic of k. But in this case $f(x) = [g(x)]^p$, where g is a polynomial in k of degree d/p; and $N(g) = N(f)$. Our results may therefore be applied to inseparable polynomials as well; however from now on we shall always assume $f(x)-y$ separable.

We are indebted to Prof. H. Davenport and Dr. J. W. S. Cassels for their helpful comments on an earlier draft of this note.

2. It is convenient to rephrase the problem in geometric terms. For $r = 2, 3, \ldots, d$, take $x_1, \ldots, x_r, y$ as coordinates in $(r+1)$-dimensional affine space; we shall need to consider the set of points given by

$$f(x_1) = \ldots = f(x_r) = y. \tag{3}$$

This is the union of a finite number of irreducible curves (each defined over k^+). We denote by T_r the sum of those curves which do not lie entirely in any hyperplane $x_i = x_j$; then it is clear that T_r is defined over k. If we write n_r for the number of points of T_r rational over k, then Weil's results show that

$$n_r = q\nu_r + O(q^{1/2}) \tag{4}$$

where ν_r is the number of those components of T_r over k^+ which are defined over k. (The error introduced by the possible singularities or intersections of components of T_r is $O(1)$.) In particular, if T_r is an irreducible curve then $\nu_r = 1$ and $n_r = q + O(q^{1/2})$.

We now show that the decomposition of T_r is determined by the Galois groups $G(f)$ and $G^+(f)$. Suppose that η is generic over k. Then $P = (\xi_1, \ldots, \xi_r, \eta)$ is on T_r if and only if the ξ_i are distinct roots of $f(x) = \eta$. Clearly, P is simple on T_r and so T_r has a unique component C containing P. By the definition of the Galois groups, the points in which C meets $y = \eta$ are just the $\sigma^+ P$ with σ^+ in G^+; and the points in which some conjugate of C over k meets $y = \eta$ are the σP with σ in G. In particular, C is defined over k if and only if every σP is a $\sigma^+ P$.

We can therefore put the curves C in one-one correspondence with the sets of r-tuples of roots of $f(x) = y$ equivalent under $G^+(f)$; and a component C is defined over k if and only if the corresponding set of r-tuples is still complete for equivalence under $G(f)$. This proves that ν_r depends only on $G(f)$ and $G^+(f)$; in particular, if $G^+(f)$ is the full symmetric group then all r-tuples are equivalent under $G^+(f)$ and so T_r is an irreducible curve.

Now let us return to the original problem. Let n_r' $(r = 2, 3, \ldots, d)$ be the number of solutions in k of

$$f(x_1) = \ldots = f(x_r)$$

for which $x_1, \ldots, x_r$ are all different; then

$$n_r' = n_r + O(1), \tag{5}$$

the error arising from those points of T_r which lie on some hyperplane $x_i = x_j$. Again, let m_r $(r = 0, 1, \ldots, d)$ be the number of y in k for which the equation $f(x) = y$ has exactly r distinct roots in k. Clearly

$$N(f) = m_1 + m_2 + \ldots + m_d; \tag{6}$$

and since to each x in k corresponds just one y,

$$q = m_1 + 2m_2 + \ldots + dm_d. \tag{7}$$

Also, from the definition of n_r',

$$n_r' = r!\, m_r + \frac{(r+1)!}{1!} m_{r+1} + \ldots + \frac{d!}{(d-r)!} m_d;$$

that is

$$\frac{n_r'}{r!} = m_r + \binom{r+1}{1} m_{r+1} + \ldots + \binom{d}{d-r} m_d.$$

Hence

$$\begin{aligned} &\frac{n_2'}{2!} - \frac{n_3'}{3!} + \ldots + (-1)^d \frac{n_d'}{d!} \\ &\quad = m_2 + \left\{\binom{3}{1} - 1\right\} m_3 + \ldots + \left\{\binom{d}{d-2} - \binom{d}{d-3} + \ldots \pm 1\right\} m_d \\ &\quad = m_2 + 2m_3 + \ldots + (d-1) m_d = q - N(f), \end{aligned}$$

by (6) and (7); for the coefficient of m_r in curly brackets is the binomial expansion of $\{(1-1)^r - 1 + r\}$. Substituting for the n_r' from (4) and (5), we find

$$N(f) = q\left\{1 - \frac{\nu_2}{2!} + \frac{\nu_3}{3!} - \ldots - (-1)^d \frac{\nu_d}{d!}\right\} + O(q^{1/2}).$$

Since the ν_r depend only on the Galois groups, this proves Theorem 2; and since when $G^+(f)$ is the full symmetric group $\nu_r = 1$ for all r, Theorem 1 follows at once.

In the particular case $d = 3$, it is easy to show that T_2 and T_3 are in general irreducible curves of genus zero. Now the error term in (4) and so also in Theorem 1 is only $O(1)$; and since we know how the error arises it would be easy to give an exact result.

3. There is in principle no difficulty in finding the Galois group of a given equation. Corresponding to any subgroup Γ of the symmetric group there is a finite set of polynomials in the roots x_i of the given equation with the property that Γ contains the Galois group of the equation if and only if these polynomials have values in the field over which we are considering the equation. (For example, the Galois group is contained in the alternating group if and only if the discriminant of the equation is a perfect square; the corresponding polynomial is $\prod(x_i - x_j)$ taken over all pairs i, j with $i < j$.) Thus the determination of the Galois group is merely a matter of examining a certain finite set of equations; see for instance van der Waerden [4], § 61. However, as a practical method this is not attractive.

Our primary problem is the calculation of $G^+(f)$; in particular we want sufficient conditions for it to be the full symmetric group. For this we have been forced to a rather inelegant device: we lift $f(x)$ to a polynomial $F(x)$ in characteristic zero.

Write Q^+ for the algebraic closure of the rationals, and write $\mathcal{C}$ for the field of complex numbers. Once and for all, we pick a definite specialisation of Q^+ to k^+, and a definite embedding of Q^+ into $\mathcal{C}$. There is a polynomial $F(x)$ over Q^+ which specialises into $f(x)$; denote by $G^+(F)$ the Galois group of $F(x) - y$ over $Q^+(y)$. Then it is clear that $G^+(F)$ is just the Galois group of $F(x) - y$ over $\mathcal{C}(y)$; and to find this we may use the properties of Riemann surfaces.

Let $\mathcal{S}$ be the Riemann surface of $F(x) = y$ over the y-plane; it has branch points at infinity and at the roots of $F'(x) = 0$. Let $\eta_0^* = \infty$, $\eta_1^*, \ldots, \eta_r^*$ be the points of the y-plane lying under these branch points. The effect of making a small closed circuit in the y-plane about one of these branch points is to permute the sheets of $\mathcal{S}$; we will denote the corresponding permutation of the roots of $F(x) = y$ by γ_i; thus, γ_i is an element of $G^+(F)$. Let Γ be the subgroup generated by these $r+1$ permutations; any one of them is redundant since the sum of circuits round each η_i^* (in the right order) vanishes, and so there is an expression for any one generator γ_i of Γ in terms of the rest. Γ is transitive, since $\mathcal{S}$ is connected.

LEMMA 1. *$G^+(f)$ may be identified as a subgroup of $G^+(F)$; moreover if F is given then $G^+(f)$ is isomorphic to $G^+(F)$ for almost all characteristics p.*

This may be proved by the method of van der Waerden; however, it is more illuminating to return to the ideas of the previous section. Write T_r^* for the sum of curves obtained from $F(x)$ in the way that T_r is from $f(x)$, let $P_0^* = (\xi_1^*, \ldots, \xi_d^*, \eta^*)$ be one intersection of T_d^* with the generic hyperplane $y = \eta^*$, and let P^* be any other such intersection. When we specialise to characteristic p, P_0^* specialises into a specific point P_0, an intersection of T_d with $y = \eta$; and once P_0 is chosen the specialisations P of all the P^* are determined. Now $G^+(F)$ consists of those permutations of the first d coordinates which take P_0^* into a P^* on the same irreducible component of T_d^*. The statements of the Lemma now follow from the facts that T_d splits at least as much as T_d^*, and for almost all p splits no more than T_d^* (Shimura [3]).

The identification of the lemma is not unique, since it depends on the choice of P_0; but for our purposes we may regard it as natural.

To state the next lemma, we need a definition. The point η_i^* on the y-plane under one or more branch points is *simple for p* if distinct roots of the equation $F(x) = \eta_i^*$ specialise into distinct roots of $f(x) = \eta_i$.

LEMMA 2. *Let γ_i be the generator of Γ corresponding to the point η_i^*; then if the order of γ_i is prime to p and η_i^* is simple for p, $G^+(f)$ contains γ_i.*

Let l be the order of γ_i. We know that if $y - \eta_i^* = t^l$ then t is a uniformising parameter for the branch points of $\mathcal{S}$ above η_i^*; thus we can expand the roots of $F(x) = y$ as

$$x_j = \xi_j^* + \sum_{\nu=1}^{\infty} a_{\nu j} t^\nu, \tag{8}$$

where the ξ_j^* are the roots of $F(x) = \eta_i^*$ and the $a_{\nu j}$ are algebraic numbers. From the algorithm which gives the expansion (8) we see that in finding the $a_{\nu j}$ we divide only by factors of l and roots of non-vanishing $\xi_m^* - \xi_n^*$; and since none of these specialise into zero, all the $a_{\nu j}$ remain finite under specialisation. Moreover, if ξ_j^* is an r-fold root of $F(x) = \eta_i^*$, then the first non-vanishing $a_{\nu j}$ is that with $\nu = l/r$; and then $a_{\nu j}$ is an lth root of a product of non-zero $\xi_m^* - \xi_n^*$ and so does not specialise into zero.

It follows that we can specialise the expansion (8) into formal power series with coefficients in k^+; and that when we do so the transformation (corresponding to γ_i)

$$t \to \omega t,$$

where ω is a primitive lth root of unity, will have order l as a permutation of the formal power series. This therefore induces an element

of order l in the Galois group of $f(x) = \eta_i + t^l$ over the field of formal power series in t^l with coefficients in k^+, and so a fortiori in $G^+(f)$; and this proves the Lemma.

LEMMA 3. *$G^+(f)$ is the symmetric group on d elements if either*
(i) *$f'(x) = 0$ has $d-1$ distinct roots giving rise to distinct values of $y = f(x)$; or*
(ii) *d is prime, there is a value of y for which $f(x) = y$ has $d-1$ distinct roots, one of them with multiplicity two, and $p \neq 2, d$.*

In case (i), $p \neq 2$, since otherwise $f'(x)$ could not have distinct roots. All the finite branch points of $\mathcal{O}$ are points at which only two sheets meet, and they are all simple for p. Hence on the one hand $G^+(f)$ contains Γ, and on the other hand Γ is a transitive group generated by transpositions and therefore is the full symmetric group.

In case (ii) we have a finite branch point at which just two of the x_j are interchanged, and this is simple for p; thus $G^+(f)$ contains a transposition. Further, since the point at infinity is simple for p, $G^+(f)$ contains, after suitable renumbering, the permutation $(12\ldots d)$. Hence, d being prime, $G^+(f)$ is primitive as well as transitive. Since it contains a transposition, it must be the full symmetric group ([1], Theorem I, p. 207). This proves the Lemma.

We may rephrase (i) in a more euphonious form, suggested by Davenport:

$G^+(f)$ is the symmetric group if the discriminant of the discriminant of $f(x)-y$ does not vanish.

Finally, we give some illustrative examples.

(i) $f_1(x) = x^d - ax$, where $d \geqslant 2$ and $a \neq 0$.

$G^+(f_1)$ is the full symmetric group by Lemma 3(i), so long as $p \nmid 2d(d-1)$.

(ii) $f_2(x) = x^d$, where $p \nmid d$.

$G^+(f_2)$ is cyclic with d elements, while $G(f_2)$ depends on g. c. d. $(p, d-1)$.

(iii) $f_3(x) = (x^2-1)^2$, $p \neq 2$.

$G^+(f_3)$ is the quaternion group. In this case all the branch points join just two sheets, but two of them lie above one another.

(iv) $f_4(x) = \frac{1}{5}x^5 - \frac{2}{3}x^3 + x$, $p \geqslant 7$.

$f_4'(x) = (x^2-1)^2$, so the discriminant of $f_4(x)-y$ is a perfect square in y. We deduce that $G^+(f_4)$ is the alternating group.

(v) $f_5(x) = x^d - ax^{d-1}$, where $d \geqslant 2$, $p \nmid 2d(d-1)$, and $a \neq 0$.

If d is prime, $G^+(f_5)$ is the full symmetric group by Lemma 3(ii); if d is not prime, this criterion is no longer applicable, but $G^+(f_5)$ is the symmetric group none the less.

References

[1] W. Burnside, *Theory of groups*, 2nd edition, Cambridge 1911. Republished by Dover, 1955.

[2] S. Chowla, *The Riemann zeta and allied functions*, Bull. Amer. Math. Soc. 58 (1952), p. 287-303.

[3] G. Shimura, *Reduction of algebraic varieties with respect to a discrete valuation of the basic field*, Amer. Journ. of Math. 77 (1955), p. 134-176.

[4] B. L. van der Waerden, *Moderne Algebra*, Springer 1931. English edition, Ungar 1949.

[5] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Actualités Sci. et Ind. 1041, Paris 1948.

TRINITY COLLEGE, CAMBRIDGE

Reçu par la Rédaction le 7. 4. 1959