DIAGONAL FORMS OF ODD DEGREE OVER A FINITE FIELD

James F. Gray

1. A PROBLEM

Throughout this paper, k is a finite field of q^f elements, k^* is the multiplicative group of nonzero elements of k, and k^p the set of p-th powers in k^*.

The literature shows the existence of nontrivial zeros in k of each of the following forms (here p denotes an odd prime):

$$a_1 x_1^3 + a_2 x_2^3 + a_3 x_3^3 \qquad (a_i \in k), \tag{1}$$

$$a_1 x_1^p + a_2 x_2^p + \cdots + a_p x_p^p \qquad (a_i \in k;\ p \geq 3), \tag{2}$$

$$a_1 x_1^p + a_2 x_2^p + \cdots + a_{p-1} x_{p-1}^p \qquad (a_i \in k;\ p \geq 5). \tag{3}$$

In particular, Lewis [2] established the existence of zeros for (1), and the author [1, Theorems 5, 8] for (2) and (3).

Without change, the proofs for (2) and (3) extend in addition to all odd positive integers p relatively prime to $q^f - 1$. The question naturally arises whether or not a restriction to higher values of p would permit further improvements. More precisely, for a fixed odd positive integer p, either itself prime or relatively prime to $q^f - 1$, what is the maximum value of t for which

$$a_1 x_1^p + a_2 x_2^p + \cdots + a_{p-t} x_{p-t}^p \qquad (a_i \in k) \tag{4}$$

has a nontrivial zero in k? Since (4) is solvable with t = 0 by (2), and since t is obviously bounded above by p - 2, such a maximum value exists.

This paper proposes the following estimate of t (notation: [x] is the greatest integer not greater than x).

THEOREM A. *If* k *is a finite field of* q^f *elements, and* p *is an odd positive integer, either prime or relatively prime to* $q^f - 1$, *then* (4) *has a nontrivial zero in* k *for* $t = t(p) = [2\sqrt{p+2}] - 4$.

Note that t(3) = 0 and t(5) = 1, in agreement with results (2) and (3) above. Further, for p = 1, although t(p) = -1, the theorem is true as stated, by inspection. Henceforth, then, we shall consider only $p \geq 3$.

2. A REFORMULATION OF THE PROBLEM

A few simple observations will suffice to show that Theorem A is a consequence of

THEOREM B. *If* k *is a finite field of* q^f *elements, if* p *is an odd prime such that* $p \mid q^f - 1$, *and if* dk^p *is a generator of* k^*/k^p (k^* the multiplicative group of k;

Received February 13, 1960.

k^p the p-th powers in k^*) *so chosen that* $1 + d \notin k^p$ *and* $1 + d \notin dk^p$, *then for* $t = [2\sqrt{p+2}] - 4$ *the form*

$$e_0 d^0 x_0^p + e_1 d^1 x_1^p + \cdots + e_{p-1} d^{p-1} x_{p-1}^p \tag{5}$$

has a nontrivial zero in k, *where*

$$e_i = \begin{cases} 0 & (i \in T = \{i_1, i_2, \cdots, i_t\},\ 2 \leq i_1 < i_2 < \cdots < i_t = p - 1), \\ 1 & (i \in N_0 = P - T,\ P = \{0, 1, 2, \cdots, p - 1\}), \end{cases}$$

and where the initial block $\{e_0, e_1, \cdots, e_{i_1-1}\}$ *of nonzero elements has a maximal length among the blocks of consecutive nonzero elements in* $\{e_0, e_1, \cdots, e_{p-1}\}$.

First of all, it may be assumed that none of the coefficients in (4) is zero, for $a_i = 0$ implies an obvious solution $x_i = 1$, $x_j = 0$ $(j \neq i)$. Further, it may be assumed that no two coefficients a_i and a_j belong to the same coset of k^*, modulo k^p, for then $a_i^{-1} a_j = a^p$ for some nonzero element a in k, and $x_i = a$, $x_j = -1$, $x_h = 0$ $(i \neq h \neq j)$ provide the obvious solution. Third, it may be assumed that p is an odd prime and that $p | q^f - 1$; for otherwise $(p, q^f - 1) = 1$ and $k^p = k^*$, all coefficients lie in the same coset modulo k^p, and the above solution applies.

The factor group k^*/k^p is cyclic of order p, and it is generated by any coset other than k^p itself. Thus for any element $d \in k^*$, $d \notin k^p$, the cosets of k^*, modulo k^p, are $\{d^i k^p\}$ $(0 \leq i \leq p - 1)$, and (4) may be rewritten in the form of (5), with exactly t zero coefficients.

Now consider the coefficients $e_i d^i$ of (5), written counterclockwise in cyclic order; there is a maximal block of consecutive nonzero coefficients, of length i_1, say, and beginning with d^r, say. Multiplication by d^{p-r} preserves the cyclic order of the coset representatives, preserves the relative location of the zeros, and after renumbering of the e's, yields

$$e_0 d^0, e_1 d^1, \cdots, e_{i_1-1} d^{i_1-1}$$

as a maximal block of consecutive nonzero coefficients, while $e_{p-1} d^{p-1} = 0$. Hence multiplication of (4) by d^{p-r} (which does not affect the existence or value of its zeros) enables us to rewrite it in the desired form.

Since $t = [2\sqrt{p+2}] - 4 < (p - 1)/2$ for $p > 1$, it follows that $p - t \geq (p + 1)/2 + 1$, and there must be at least two consecutive nonzero coefficients remaining in (5); hence $e_0 = e_1 = 1$ and $2 \leq i_1$.

Further, it is readily shown [1] that d may be chosen so that $1 + d \notin k^p$ and $1 + d \notin dk^p$.

Let W be the set of nonzero elements of k^* which are of the form $z^p - 1$ $(z \in k^*)$. Then the number of elements of W is $(q^f - 1)/p - 1$. Let W^{-1} be the set of inverses of elements of W, and let $V = k^p \cup W \cup W^{-1}$. Then, if $p \geq 3$, V has at most

$$\frac{q^f - 1}{p} + 2\left(\frac{q^f - 1}{p} - 1\right) = \frac{3q^f - 2p - 3}{p} < q^f - 1$$

elements. Now choose d to be one of the elements of k^* not in V. Clearly d and $1 + d$ are not in k^p and, since V is closed under the operation of taking inverses,

d^{-1} is not in V. Further $1 + d$ is not in dk^p; for if it were, $(1 + d)/d = 1 + d^{-1}$ would be in k^p, and d^{-1} would be in W and consequently in V, contrary to fact. Hence $1 + d = d^{i_0} a^p$ for some i_0 $(2 \leq i_0 \leq p - 1)$ and for some a in k which must be nonzero, for otherwise $d = -1 \in k^p$.

Now let $T = \{i \mid e_i = 0\}$. Then $T = \{i_1, i_2, \cdots, i_t\}$ $(2 \leq i_1 < i_2 < \cdots < i_t = p - 1)$. Hence (4) either has an evident solution, or it can be rewritten to satisfy the hypotheses of Theorem B. Hence, to establish Theorem A, we need only complete the proof of Theorem B.

3. A USEFUL NOTATION

The roles of the lengths and placements of blocks of zero and nonzero coefficients suggests the following terminology which, in addition to supplying the mechanism for proving Theorem B, may be useful in further extensions.

Our concern is now with naturally ordered subsequences of the ordered sequence $P = \{0, 1, 2, \cdots, p - 1\}$ —in particular, with the index set of nonzero coefficients in (5),

$$T = \{i_1, i_2, \cdots, i_t\} \qquad (2 \leq i_1 < i_2 < \cdots < i_t = p - 1),$$

and with $N_0 = P - T$.

By the *length* of a sequence is meant the number of elements in the sequence. If A is an ordered subsequence of P, a *distinguished sequence* in A is an ordered subsequence $\{j_1, j_2, j_3, \cdots, j_n\}$ of A such that i) $j_{r+1} - j_r = 1$ $(1 \leq r \leq n - 1)$, and ii) $j_n + 1 \notin A$.

Distinguished sequences in A will be called A-*sequences*. It has been noted that the initial N_0-sequence $\{0, 1, 2, \cdots, i_1 - 1\}$ has maximal length among the N_0-sequences.

Let

$$N_j = \{i \in T \mid i \text{ initiates a T-sequence of length } j\},$$

and let us, by convention, say that if $i \in N_0$ (that is, $i \notin T$), then i initiates a T-sequence of length zero.

Since p is finite, there exists a unique integer c such that $N_c \neq \emptyset$, $N_{c+1} = \emptyset$. If T itself is void, our convention yields $c = 0$. The sets $N_0, N_1, \cdots, N_c$ partition P.

A *terminal* T-sequence is one which contains $i_t = p - 1$.

[Example. If $p = 19$ and $T = \{4, 5, 6, 7, 13, 17, 18\}$, the family of T-sequences consists of $\{7\}$, $\{13\}$, $\{18\}$ (of length one), $\{6, 7\}$, $\{17, 18\}$ (of length two), $\{5, 6, 7\}$ (of length three), $\{4, 5, 6, 7\}$ (of length four); $\{18\}$ and $\{17, 18\}$ are terminal T-sequences; $N_0 = \{0, 1, 2, 3, 8, 9, 10, 11, 12, 14, 15, 16\}$, $N_1 = \{7, 13, 18\}$, $N_2 = \{6, 17\}$, $N_3 = \{5\}$, $N_4 = \{4\}$, $N_5 = \emptyset$ and $c = 4$. The initial N_0-sequence is $\{0, 1, 2, 3\}$ and has length four.

We say that b can appear effectively at location i in (5) if $b = d^i a^p$ for some $a \in k^*$ and some $i \in N_0$.

Recall that $1 + d = d^{i_0} a^p$ $(2 \leq i_0 \leq p - 1,\ a \in k^*)$.

If $i_0 \in N_j$, then $i_0 + j$ is in N_0 and $i_0 + j \leq p$ with equality holding only if i_0 belongs to a terminal sequence. Then

$$d^j + d^{j+1} = d^j(1 + d) = d^j d^{i_0} a^p = d^{i_0+j} a^p$$

and $d^j + d^{j+1}$ appears effectively at location $i_0 + j$ when $i_0 + j < p$ and at location 0 when $i_0 + j = p$, since then $d^{i_0+j} a^p = 1\,(da)^p$.

One additional assumption will now provide a solution. Let it be assumed that the initial N_0-sequence has length at least $c + 2$, in other words, that it is $\{0, 1, 2, \cdots, c + 1, \cdots\}$. Then, since $j \leq c$, we see that j and $j + 1$, as well as $i_0 + j$, belong to N_0 and hence x_j, x_{j+1}, and x_{i_0+j} appear effectively in (5). Therefore $- E_j - E_{j+1} + aE_{i_0+j}$ is a solution of (5), where $E_r = (x_0, x_1, \cdots, x_{p-1})$ with $x_r = 1$ and $x_s = 0$ for $s \neq r$.

Thus we have established

LEMMA 1. *If* p *is an odd prime, and if* c *is the maximum number of consecutive zero coefficients in* (5), *then* (5) *has a solution in* k, *nontrivial in the* $p - t$ *effectively appearing variables, provided that* $c + 1 < i_1$ (*in other words, provided the first* $c + 2$ *coefficients are nonzero*).

The next lemma is readily obtained.

LEMMA 2. *If* t *is chosen so that* $t + 4 < 2\sqrt{p + 3}$, *and if* c *is the maximum number of consecutive zero coefficients in* (5), *then at least the first* $c + 2$ *coefficients in* (5) *are nonzero* (*that is*, $c + 1 < i_1$).

We note that $e_p - 1 = e_{i_t} = 0$, so that the presence of t zeros among the coefficients of (5), with c of these zeros consecutive, leaves at most $t - c + 1$ separated blocks of consecutive nonzero coefficients. Suppose that the maximal length of these is less than $c + 2$. Then we have, as the maximum possible number of nonzero coefficients, $(t - c + 1)(c + 1) \geq p - t$. But simplification of this inequality yields

$$c^2 - tc - 1 - 2t + p \leq 0. \tag{7}$$

In (7) the quadratic form in c has discriminant

$$t^2 + 4 + 8t - 4p \tag{8}$$

and has a real zero if and only if $t^2 + 4 + 8t - 4p \geq 0$, that is, $(t + 4)^2 \geq 4(p + 3)$. Since this inequality contradicts the hypothesis on t, we conclude that for $t + 4 < 2\sqrt{p + 3}$ the initial (maximal) block of consecutive nonzero coefficients has length at least $c + 2$ (that is, $c + 1 < i_1$).

Theorem B now follows, since, for $t = [2\sqrt{p + 2}] - 4$, we have $t + 4 < 2\sqrt{p + 3}$, and Lemma 2 applies, guaranteeing the hypotheses of Lemma 1, which in turn guarantees the desired solution of (5).

There is no weakening of Lemma 2 in replacing the condition $t + 4 < 2\sqrt{p + 3}$ by $t + 4 = [2\sqrt{p + 2}]$, since it is an elementary fact that the maximum integral value of t less than $2\sqrt{p + 3}$ is precisely $[2\sqrt{p + 2}] - 4$.

4. A MORE GENERAL FORM OF THEOREM A FOR PRIMES

THEOREM A*. *If* k *is a finite field and* p *and* p_0 *are odd primes* ($p \geq p_0$), *then* (4) *has a nontrivial zero in* k *for* $t = t(p_0) = [2\sqrt{p_0 + 2}] - 4$.

This follows immediately from Theorem A, since t(p) is a nondecreasing function and $p - t(p) \leq p - t(p_0)$ for $p \geq p_0$. But Theorem A establishes the desired solution for (4) in $p - t(p)$ variables and so, *a fortiori*, for $p - t(p_0)$ variables.

5. COUNTEREXAMPLES

Since there are no nontrivial zeros for $x_1^3 + 2x_2^3$ in k = GF(7) and for $x_1^5 + 2x_2^5 + 4x_3^5$ in k = GF(11), the values t(3) = 0 and t(5) = 1 are best possible. However, it seems improbable that the given value of t is best possible for forms of prime degree greater than 7.

In fact, it still remains an open question whether t(7) = 2 is best possible; for the simple type of counterexample given above fails to apply to four seventh powers.

REFERENCES

1. J. F. Gray, *Diagonal forms of prime degree.* Doctoral Dissertation, University of Notre Dame (1958).

2. D. J. Lewis, *Cubic congruences*, Michigan Math. J. 4 (1957), 85-95.

St. Mary's University
San Antonio, Texas