ON A PROBLEM OF ARNOLD ON UNIFORM DISTRIBUTION

Mei-Chu Chang

Abstract. In this note we consider some quantitative versions of conjectures made by Arnold related to Galois dynamics in finite fields. We refine some results by Shparlinski using exponential sum results.

The present note is a refinement of work of I. Shparlinski [S] on the ergodic properties of certain dynamical systems associated to multiplication in finite fields and originating from some problems posed by I. V. Arnold (see [A]). As shown in [S] the issue (explained below) turns out to be ultimately connected to questions on incomplete exponential sums of Gauss type. Using the 'standard' bounds on such sums, (combined with discrepancy estimates) Arnold's question was settled affirmatively in [S]. In fact, the result obtained in [S] shows uniform distribution of even much shorter orbits than considered in [A] and raises the natural question: *what is the true condition to establish this phenomenon?* Our first and main aim here is to show how recently obtained exponential sum bounds in fields $\mathbb{F}_p^n$ (see [BC]) and going well beyond Gauss' estimates, permit to prove uniform distribution of orbits of length $M \geq p^\delta$ for any fixed $\delta > 0$, while in [S] the condition $M \gg p^{n/2}(\log p)^2$ is required. Next, a self-contained account is given of how to derive directly from the exponential sum bound the discrepancy estimates for smooth domains (without first 'passing through boxes'). No effort has been made however to optimize the error term.

Next, we describe our problem in details.

Let p be a large prime, and let $n \in \mathbb{Z}^+$ be fixed. A finite field $\mathbb{F}_{p^n} \cong \mathbb{F}_p[\xi]$ can be viewed as a n-dimensional vector space over $\mathbb{F}_p$ via the correspondence

$$\sum_{j=0}^{n-1} x_j \xi^j \longleftrightarrow (x_0, \dots, x_{n-1}).$$

Let $\overline{a}_m = (a_{m,0}, \dots, a_{m,n-1})$ be the vector corresponding to $\xi^m = \sum_{j=0}^{n-1} a_{m,j}\, \xi^j$. After identifying $\mathbb{F}_p$ with $\{0, 1, \dots, p-1\}$, we have

$$\frac{1}{p}\, \overline{a}_m \in [0,1]^n \subset \mathbb{Q}^n, \quad \text{where } m = 1, \dots, M.$$

We are interested in the distribution of $\{\frac{1}{p}\,\overline{a}_m : m = 1, \dots, M\}$ in $[0,1]^n$. For example, if $M = p^n - 1$, we get regular lattice with only one point $(0, \dots, 0)$ missing. Arnold Conjectured [A] that there is uniform distribution even for small values of M. More precisely, let $\Omega \subset [0,1]^n$ be a region with smooth boundary and denote

$$N_\xi(M, \Omega) = \big|\{1 \le m \le M : \frac{1}{p}\,\overline{a}_m \in \Omega\}\big|. \tag{1}$$

Conjecture. (*Arnold*)

$$N_\xi(M, \Omega) = M \ vol\, \Omega + o(M)$$

even for small values of M.

By 'small' Arnold refers to $M = o(p^n)$. Shparlinski proved that $M \gg p^{\frac{n}{2}} (\log p)^2$. In this note we will improve the lower bound on M obtained by Shparlinski.

Theorem 1. *Let* $N_\xi(M, \Omega)$ *be defined as in (1). Then*

$$N_\xi(M, \Omega) = M \ vol\, \Omega + o(M)$$

for $M > p^{\delta n}$ *for any fixed* $\delta > 0$.

In [S], the author also considers the following general problem. Let $f(x) \in \mathbb{Z}[x]$ be a fixed nonconstant polynomial and write

$$\xi^{f(m)} = \sum_{j=0}^{n-1} a_{m,j} \xi^j, \tag{2}$$

where $\bar{a}_m = (a_{m,0}, \cdots, a_{m,n-1}) \in \{0, \cdots, p-1\}^n$.

Similarly, one can then study the distribution of orbits $\{\frac{1}{p}\,\overline{a}_m : m = 1, \dots, M\}$ in $[0,1]^n$. In this context, two results are proven in [S]. The first (see [S], Theorem 5) establishes for general $f(x)$ as above a uniform distribution property

$$N_\xi(M, \Omega) = M \ \text{vol}\, \Omega + o(M) \tag{3}$$

for 'most' primitive roots $\xi \in \mathbb{F}_{p^n}^*$ and $M > p^{\frac{n}{2}+\epsilon}$, where $N_\xi(M, \Omega)$ is defined as in (1). In the second result, the special case of a monomial $f(x) = x^k$, $k \ge 2$ is considered, for which it is shown that for any generator ξ, the full orbit satisfies

$$N_\xi(p^n, \Omega) = p^n \ \text{vol}\, \Omega + O(p^{n-\delta(k)})$$

for some $\delta(k) > 0$. (See [S], Theorem 6.)

In this view we will establish here a result under an additional assumption on p and n. More precisely, assume the following

$(*)$ *$p^n - 1$ has a square factor q^2, such that $q > p^\varepsilon$ and $(p^\nu - 1, q) = 1$ for all $1 \le \nu < n, \nu | n$.*

(Here $\varepsilon > 0$ is arbitrary and fixed.)

Let us first point out that this condition may be fulfilled for infinitely many primes p. For example, for the case $n = 2$, the condition (*) amounts to $p + 1$ having a large square divisor q^2, $q > p^\varepsilon$. Fix q large. According to Linnik's theorem, there is a prime p such that

$$p \equiv q^2 - 1 \ (\mathrm{mod}\ q^2)$$

and

$$p \lesssim q^{2L},$$

where L is an absolute constant. (One may take $L = 5.5$ according to a result of Heath-Brown.) See [IK] for details. It follows that $p^2 - 1$ satisfies (*) with $q > cp^{\frac{1}{11}}$.

Theorem 2. *Assume that (*) holds. Given a nonconstant $f(x) \in \mathbb{Z}[x]$ and with notation as in (2) and (1), we have*

$$N_\xi(M, \Omega) = M \ vol\, \Omega + o(M)$$

holds, provided $M > p^{n - \varepsilon/2}$.

(The generator ξ is arbitrary.)

The remainder of the paper is organized as follows. We first prove Theorem 1, relying essentially on [BC], and making our treatment self-contained by providing a full argument for the discrepancy bound. At the end we state and prove the exponential sum bound that replaces the estimate in [BC] in order to derive Theorem 2.

We will follow the new convention to use $d \lesssim f$ meaning $d \le cf$, where c is a function of some parameters independent of d and f.

Let $\tau > 0$ be fixed. We construct two smooth functions $F_+, F_- : [0,1]^n \to [0,1]$ with the following properties:

(i) $\mathrm{supp}\, F_- \subset \Omega$

(ii) $\big|\Omega \backslash \{F_- = 1\}\big| < \tau |\partial\Omega|$

(iii) $\big|\partial_x^{(\alpha)} F_-\big|$, $\big|\partial_x^{(\alpha)} F_+\big| < C\tau^{-|\alpha|}$ for any multi-index $\alpha = (\alpha_0, \dots, \alpha_n)$

(i') $F_+ = 1$ on Ω

(ii') $\big|(\text{ supp}\, F_+)\backslash\Omega\big| < \tau|\partial\Omega|$

Here $\partial\Omega$ is the boundary of $|\Omega|$, and $\partial_x^{(\alpha)} = \partial_{x_0}^{(\alpha_0)} \cdots \partial_{x_{n-1}}^{(\partial_{n-1})}$ is the differential. Also, $|\Gamma|$ denote the measure of a region Γ.

Let $\mathcal{X}_\Omega =$ be the indicator function of Ω. Then (i), (ii), (i') and (ii') imply

$$F_- \le \mathcal{X}_\Omega \le F_+.$$

Hence

$$\sum_{m=1}^{M} F_-\Big(\frac{1}{p}\bar{a}_m\Big) \le N_\xi(M,\Omega) \le \sum_{m=1}^{M} F_+\Big(\frac{1}{p}\bar{a}_m\Big). \tag{4}$$

Claim. *For $F = F_+$ or F_-, we have*

$$\Big|\sum_{m=1}^{M} F\Big(\frac{1}{p}\bar{a}_m\Big) - |\Omega|\cdot M\Big| < cM(\tau|\partial\Omega| + \tau^{-n-1}p^{-\varepsilon}).$$

Proof. For $k \in \mathbb{Z}^n$, recall that the Fourier transform of F at k is

$$\hat{F}(k) = \int_0^1 \cdots \int_0^1 F(x)e^{-2\pi ik\cdot x}dx.$$

Hence we have

$$F(x) = \hat{F}(0) + \sum_{k\neq 0} \hat{F}(k)e^{2\pi ik\cdot x}.$$

Then

$$\sum_{m=1}^{M} F\Big(\frac{1}{p}\bar{a}_m\Big) = M\cdot\hat{F}(0) + \sum_{k\neq 0}\hat{F}(k)\Big[\sum_{m=1}^{M} e_p(k\cdot\bar{a}_m)\Big], \tag{5}$$

where $e_p(\theta) = e^{\frac{2\pi i}{p}\theta}$.

Let $Tr(x) = x + x^p + \ldots + x^{p^{n-1}}$ be the trace of $x \in \mathbb{F}_{p^n}$ and let $\omega_0, \ldots, \omega_{n-1} \in \mathbb{F}_{p^n}$ be the dual basis to $1, \xi, \ldots, \xi^{n-1}$. Hence

$$Tr(\omega_i\xi^j) = \delta_{i,j}, \text{ for } 0 \le i,j < n$$

and

$$\bar{a}_m = \big(Tr(\omega_0\xi^m), \dots, Tr(\omega_{n-1}\,\xi^m)\big).$$

Therefore

$$\begin{aligned}\sum_{m=1}^{M} e_p(k\cdot\bar{a}_m) &= \sum_{m=1}^{M} e_p\big(Tr\big((k_0\omega_0+\cdots+k_{n-1}\omega_{n-1})\xi^m\big)\big)\\ &= \sum_{m=1}^{M} e_p\big(Tr\big((k\cdot\omega)\,\xi^m\big)\big),\end{aligned} \tag{6}$$

where $\omega = (\omega_0, \cdots, \omega_{n-1})$.

We will use the following estimates on incomplete Gauss sums in $\mathbb{F}_{p^n}$.

Theorem BC. *[BC] Let $g \in \mathbb{F}_{p^n}$ be a unit with ord$(g) = t$, and let $t \geq t_1 > p^{\varepsilon}$. Suppose*

$$\max_{\substack{1\leq\nu<n\\ \nu|n}} \gcd(p^{\nu}-1, t) < p^{-\varepsilon}t$$

for some $\varepsilon > 0$. Then

$$\max_{a\in\mathbb{F}_{p^n}^*} \Big|\sum_{j\leq t_1} e\big(Tr(ag^j)\big)\Big| < cp^{-\delta}t_1$$

where $\delta = \delta(\varepsilon) > 0$.

Remark BC. *The assumption $t \geq t_1$ is vacuous. One only needs to assume that $t, t_1 > p^{\varepsilon}$. In fact, we can write $t_1 = tq + r$, with $q \in \mathbb{N}$ and $0 \leq r < t$. If $r < p^{\varepsilon/2}$, then $r < p^{-\varepsilon/2}t_1$ and Theorem BC gives*

$$\begin{aligned}\Big|\sum_{j\leq t_1} e\big(Tr(ag^j)\big)\Big| &\leq \sum_{i=0}^{q}\Big|\sum_{j=ti+1}^{t(i+1)} e(Tr(ag^j)\Big| + r\\ &< cqp^{-\delta}t + p^{-\varepsilon/2}t_1\\ &\lesssim t_1p^{-\varepsilon/2}t_1.\end{aligned}$$

If $r \geq p^{\varepsilon/2}$, we apply Theorem BC to $t \geq r \geq p^{\varepsilon/2}$.

For those k satisfying

$$0 < |k| = \max|k_i| < p, \tag{7}$$

we use Theorem BC to bound (6). Since (7) implies $k \cdot \omega \neq 0$, we have

$$\Big| \sum_{m=1}^{M} e_p(k \cdot \bar{a}_m) \Big| < p^{-\varepsilon} M, \text{ if } 0 < |k| = \max |k_i| < p.$$

Therefore, the second term in the right-hand-side of (5) is bounded by

$$\Big| \sum_{k \neq 0} \hat{F}(k) \Big[\sum_{m=1}^{M} e_p(k \cdot \bar{a}_m) \Big] \Big| < p^{-\varepsilon} M \sum_{0<|k|<p} |\hat{F}(k)| + M \sum_{|k| \geq p} |\hat{F}(k)|.$$

Property (iii) implies

$$|\hat{F}(k)| < c(\tau|k|)^{-n-1}, \quad k \in \mathbb{R}^n \backslash \{0\}.$$

Therefore,

$$\begin{aligned} \Big| \sum_{k \neq 0} \hat{F}(k) \Big[\sum_{m=1}^{M} e_p(k \cdot \bar{a}_m) \Big] \Big| &\lesssim p^{-\varepsilon} M \tau^{-n-1} \sum_{0<|k|<p} |k|^{-n-1} + M \tau^{-n-1} \sum_{|k| \geq p} |k|^{-n-1} \\ &\lesssim M \tau^{-n-1} (p^{-\varepsilon} + p^{-1}) \\ &\leq c M \tau^{-n-1} p^{-\varepsilon}. \end{aligned} \tag{8}$$

Also, properties (i), (ii), (i') and (ii') imply

$$\hat{F}(0) = \int F(x) dx = |\Omega| + O(\tau |\partial\Omega|). \tag{9}$$

Now the claim follows from (5), (8) and (9). □

Claim and (1) imply

$$\big| N_\xi(M, \Omega) - M \cdot |\Omega| \big| < cM(\tau |\partial\Omega| + \tau^{-n-1} p^{-\varepsilon}). \tag{10}$$

Taking

$$\tau = \Big(\frac{p^{-\varepsilon}}{|\partial\Omega|} \Big)^{\frac{1}{n+2}} \tag{11}$$

gives

$$\big| N_\xi(M, \Omega) - M \cdot |\Omega| \big| < cM p^{-\frac{\varepsilon}{n+2}} |\partial\Omega|^{\frac{n+1}{n+2}}. \tag{12}$$

Recall that Ω has a smooth boundary. (In particular $|\partial\Omega| < \infty$). Therefore inequality (12) gives that

$$N_\xi(M,\Omega) = M|\Omega| + o(Mp^{-\frac{\varepsilon}{n+2}}) = M|\Omega| + o(M), \tag{13}$$

and Theorem 1 is proved.

Next, we will prove Theorem 2. Following the same argument, it is clear that the only additional input needed are nontrivial bounds on sums of the form

$$\sum_{m=1}^{M} e\big(Tr(a\xi^{f(m)})\big). \tag{14}$$

Assuming $p^n - 1$ satisfies (*). Then Theorem 2 follows from the following

Proposition 3. *Assume (*) holds. Then*

$$\max_{a\in\mathbb{F}_{p^n}^*} \left| \sum_{m=1}^{M} e\big(Tr(a\xi^{f(m)})\big) \right| < cM^{1-\delta}, \tag{15}$$

provided $M > p^{n-\frac{\varepsilon}{2}}$. *Here* $\delta = \delta(\varepsilon, f) > 0$.

Proof. Let

$$f(x) = \sum_{s=0}^{d} c_s x^s,$$

where $c_s \in \mathbb{Z}$ and $c_d \neq 0$.

By assumption

$$p^n - 1 = q^2 A \text{ with } A \in \mathbb{Z}^+. \tag{16}$$

For $m \in \{1, \cdots, M\}$ we write m in the form

$$m = qAj + r \text{ with } r \in \{0, 1, \cdots, qA-1\} \text{ and } j \leq \frac{M}{qA}. \tag{17}$$

Recall that $qA < p^{n-\varepsilon}$ by the assumption on q. Hence $\frac{M}{qA} > p^{\varepsilon/2}$. Next, by (17) and (16),

$$\begin{aligned} f(m) &= \sum_{s=0}^{d} c_s(qAj + r)^s \\ &\in \sum_{s=0}^{d} c_s r^s + qAj\Big(\sum_{s=1}^{d} s c_s r^{s-1}\Big) + (p^n - 1)\mathbb{Z}. \end{aligned}$$

Therefore,

$$\Big|\sum_{m=1}^{M} e\big(Tr(a\xi^{f(m)})\big)\Big| \le \sum_{r=0}^{qA-1}\Big|\sum_{j=0}^{[\frac{M}{qA}]} e\big(Tr(a_r\xi_r^j)\big)\Big| + qA, \tag{18}$$

where

$$a_r = a\xi^{\sum_{s=0}^{d} c_s r^s} \neq 0$$

and

$$\xi_r = \xi^{qA(\sum_{s=1}^{d} sc_s r^{s-1})}.$$

Fixing r, we apply Remark BC to the inner sum in (18). Thus

$$\Big[\frac{M}{qA}\Big] \sim \frac{M}{qA} > p^{\frac{\varepsilon}{2}},$$

and by (16), ξ_r satisfies

$$t_r = ord(\xi_r) = \frac{p^n-1}{\gcd(p^n-1, qA\sum_{s=1}^{d} sc_s r^{s-1})} = \frac{q}{\gcd(q, \sum_{s=1}^{d} sc_s r^{s-1})}. \tag{19}$$

Also, by assumption (*) that $(q, p^\nu - 1) = 1$ for any $1 \le \nu < n, \nu | n$, hence we have

$$(t_r, p^\nu - 1) = 1, \text{ if } 1 \le \nu < n, \nu | n.$$

For Remark BC to be applicable, it therefore suffices to assume that $t_r > p^{\frac{\varepsilon}{2}}$. Since $q > p^\varepsilon$, we require that

$$\gcd\,(q, \sum_{s=1}^{d} sc_s r^{s-1}) < p^{\frac{\varepsilon}{2}}, \tag{20}$$

by excluding a set of exceptional values of r. We will analyze condition (20). Denoting

$$\mathcal{D} = \{k \in \mathbb{Z} : k \text{ divides } q \text{ and } k \ge p^{\frac{\varepsilon}{2}}\}.$$

For $k \in \mathcal{D}$, denote

$$\mathcal{R}_k = \big\{0 \le r < qA : \sum_{s=1}^{d} sc_s r^{s-1} \equiv 0 \ (\mathrm{mod}\ k)\big\}.$$

Note that r fails to satisfy (20) if and only if r is in the set

$$\bigcup_{k\in\mathcal{D}} \mathcal{R}_k = \bigcup_{k\in\mathcal{D}} \big\{0 \le r < qA : \sum_{s=1}^{d} sc_s r^{s-1} \equiv 0 \ (\mathrm{mod}\ k)\big\}. \tag{21}$$

Hence we want to bound the cardinality of the set (21).

To bound $|\mathcal{R}_k|$, we divide the interval $[0, qA]$ into subintervals of length k^{1/d^2} each. Observe that each subinterval cannot contain d distinct values of r. In fact, assume by contradiction that $r_1 < r_2 < \cdots < r_d$ are in the same subinterval of $\mathcal{R}_k$ such that $\sum_{s=1}^{d} sc_s r_i^{s-1} \equiv 0 \ (\mathrm{m}od\ k)$. By dividing $\gcd(k, dc_d)$ if necessary, we may assume dc_d is invertible in $\mathbb{Z}_k$. Hence the Van der Monde determinate

$$\det_{\substack{1\le i\le d\\ 0\le s<d}} (r_i^s) \equiv 0 \ (\mathrm{m}od\ k).$$

This is clearly impossible since $|r_i - r_j| < k^{1/d^2}$ and $|\det(r_i^s)| < k$. From this observation, we conclude that the cardinality of (21) is bounded by

$$\begin{aligned}\sum_{k\in\mathcal{D}} d\ \frac{qA}{k^{1/d^2}} &\lesssim qA\ p^{-\varepsilon/2d^2}|\mathcal{D}| \\ &\lesssim qA\ p^{-\varepsilon/2d^2} \exp\left(\frac{n\log p}{\log\log p}\right) \\ &< qA\ p^{-\varepsilon/4d^2}. \end{aligned} \tag{22}$$

(since p is large.)

Returning to (18), it follows from the proceeding and Remark BC that

$$\begin{aligned}\left|\sum_{m=1}^{M} e\big(Tr(a\xi^{f(m)})\big)\right| &\le \sum_{\substack{0\le r<qA\\ r\notin \mathrm{set}(21)}} \left|e\big(Tr(a\xi^{f(m)})\big)\right| + qA\ p^{-\varepsilon/4d^2}\frac{M}{qA} + qA \\ &< qAp^{-\delta}\frac{M}{qA} + p^{-\varepsilon/4d^2}M + p^{-\varepsilon/2}M \\ &< M^{1-\delta'}. \qquad \square\end{aligned}$$

References

[A]. V.I. Arnold, *Geometry and dynamics of Galois fields*, Russian Math. Surveys, 59 (2004), 1029-1046.

[BC]. J. Bourgain, M-C. Chang, *A Gauss sum estimate in arbitrary finite fields*, Comptes-Rendus (to appear).

[IK]. H. Iwaniec, E. Kowalski, *Analytic number theory, AMS Coll 53 (2004).*

[S]. I. Shparlinski, *On some dynamical systems in finite fields and residue rings*, preprint (2005).

Mathematics Department, UCR Riverside, CA 92521 U.S.A.

E-mail address: mcc@math.ucr.edu