References

[1] E. Bombieri, *Le grand crible dans la théorie analytique des nombres*, Asterisque 18 (1974), Societé Mathématique de France, page 22.
[2] J. H. Conway and A. J. Jones, *Trigonometric diophantine equations (on vanishing sums of roots of unity)*, Acta Arith. 30 (1976), 229–240.
[3] J. H. Loxton, *On two problems of R. M. Robinson about sums of roots of unity*, ibid. 26 (1974), 159–174.
[4] H. B. Mann, *On linear relations between roots of unity*, Mathematika 12 (1965), 107–117.
[5] A. Schinzel, *Reducibility of lacunary polynomials VIII*, Acta Arith. 50 (1988), 91–106.

Received on 7.8.1987 (1743)

ACTA ARITHMETICA
LII (1989)

Some number-theoretical properties of generalized sum-of-digit functions

by

GERHARD LARCHER (Salzburg) and ROBERT F. TICHY (Wien)

1. Introduction. There has been a great deal of work in investigating the number-theoretical properties of the sum of digits of positive integers in a given number system. In the special case of a q-ary number system ($q \geqslant 2$) write n in the digit representation

$$n = \sum_{i=0}^{\infty} \varepsilon_i q^i \tag{1.1}$$

with $\varepsilon_i = \varepsilon_i(q, n) \in \{0, \ldots, q-1\}$ and $\varepsilon_i = 0$ for $i > [\log n/\log q]$; $[x]$ denotes, as usual, the greatest integer $\leqslant x$. Then by a famous result of Delange [3]

$$\frac{1}{n}\sum_{k=0}^{n-1} s(q, k) = \frac{q-1}{2}\frac{\log n}{\log q} + nF\left(\frac{\log n}{\log q}\right), \tag{1.2}$$

where $s(q, k) = \sum_{j=0}^{\infty} \varepsilon_j(q, k)$ denotes the sum of q-ary digits and F is a suitable continuous and nowhere differentiable function with period 1. Exact bounds of the error term $F(\log n/\log q)$ have been given by Drazin and Griffiths [4]. A further precise information on the average value of the sum of q-ary digits is given in a recent paper of Foster [6]. In the case $q = 2$ he proved

$$-\frac{2}{13} < \frac{2}{n}\sum_{k=0}^{n-1} s(2, k) - \left[\frac{\log n}{\log q}\right] < 1, \tag{1.3}$$

where both bounds are best possible. A paper of Stolarsky [12] contains a brief survey of the history of such problems and cites many references.

Other authors, expecially French mathematicians investigated certain exponential sums, e.g.

$$\sum_{k=0}^{n-1} e^{2\pi i h s(q,k) x} \quad (h \text{ integral, } x \text{ irrational}) \tag{1.4}$$

in connection with the uniform distribution of the sequence $(s(q, n)x)_{n=0}^{\infty}$.

These investigations were initiated by Mendès-France [10] and continued in several articles by Coquet, e.g. [2].

In recent time more general digit depending sums and sequences turned out to be of some importance in various fields of applications; for instance the Rudin–Shapiro sequence (cf. Allouche and Mendès-France [1]) with applications in harmonic analysis and in the theory of automata and the Gray code representation with applications in computer science (cf. Sedgewick [11], Flajolet and Ramshaw [5]). The digits $\gamma_i(n)$ in Gray code representation are given by

$$\gamma_i(n) = \varepsilon_i(2, n)+\varepsilon_{i+1}(2, n) \mod 2, \tag{1.5}$$

and $G(k) = \sum_{j=0}^{\infty} \gamma_j(k)$ denotes the sum of Gray code digits. Obviously $G(k)$ is the number of maximal 0-blocks and 1-blocks in the binary representation of k. In an appendix we prove an explicit formula for

$$\frac{1}{n}\sum_{k=0}^{n-1} G(k)-\frac{1}{2}\left[\frac{\log n}{\log q}\right] \tag{1.6}$$

(cf. Foster [6]). From this formula it is possible to derive lower and upper bounds for the expression (1.6) which can be used to give estimates for the average case complexity

$$\frac{n}{2}+2\sum_{k\geqslant 1} G(k)\left(\binom{2n}{n-k}\Big/\binom{2n}{n}\right)$$

of Batcher's sorting algorithm with n files (cf. Sedgewick [11], Flajolet and Ramshaw [5]).

Our main results are concerned with estimates for exponential sums of type (1.4) with respect to Gray code and some extensions. The sequence $(G(n))_{n=0}^{\infty}$ is a special case of a more general class of sequences. Let $q, \varkappa \in N$ (positive integers), $q \geqslant 2$ and for $i_1, \ldots, i_\varkappa \in \{0, \ldots, q-1\}$ let $a(i_1, \ldots, i_\varkappa)$ be a real number and assume $a(0, \ldots, 0) = 0$. For n in q-ary digit representation (1.1) we define

$$t(n) := \sum_{i=0}^{\infty} a(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+\varkappa-1}). \tag{1.7}$$

In the case $\varkappa = 1$ and $a(i) = i$, $t(n) = s(q, n)$ is the usual sum of q-ary digits. For $q = \varkappa = 2$ and $a(0, 0) = a(1, 1) = 0$, $a(0, 1) = a(1, 0) = 1$ we have $t(n) = G(n)$. In the case $q = \varkappa = 2$ and $a(0, 0) = a(0, 1) = a(1, 0) = 0$, $a(1, 1) = 1$, $t(n)$ is the number of (11)-blocks in the binary representation of n (Rudin–Shapiro-sequence).

For a pleasant formulation of the main theorems we need the following

quantity ω. Let b, l be integers with $0 \leqslant b, l < q^\varkappa$ in q-ary representation

$$b = \sum_{i=0}^{\varkappa-1} b_i q^i, \quad l = \sum_{i=0}^{\varkappa-1} l_i q^i.$$

Then for real x we define

$$\sigma = \sigma(x, l, b) = x\cdot\Big\{t(l)+\sum_{m=1}^{\varkappa-1}\big(a(b_m, \ldots, b_{\varkappa-1}, l_0, \ldots, l_{m-1})- a(b_m, \ldots, b_{\varkappa-1}, 0, \ldots, 0)\big)\Big\}. \tag{1.8}$$

Furthermore we set

$$\omega(x) = \max_{b,l}\|\sigma(x, l, b)\| \quad (\|y\| = \min(y-[y], 1-y+[y])),$$
$$S(N, x) = \sum_{k=0}^{N-1} e^{2\pi i t(k)\cdot x}. \tag{1.9}$$

THEOREM 1. *For all positive integers N and real x*

$$|S(N, x)| \leqslant C_1 N^{\frac{\log(q-C_2\omega^2(x))}{\log q}}$$

with some positive constants C_1, C_2 only depending on q and $\varkappa$.

Our second result is concerned with a special lower bound for $S(N, x)$.

THEOREM 2. *There is a constant $\delta > 0$ such that for all reals x with $\omega(x) \leqslant \delta$ and for all $N = q^{j\varkappa}$ $(j \in N)$ we have*

$$|S(N, x)| \geqslant C_3 N^{\frac{\log(q-C_4\omega^2(x))}{\log q}}$$

with some positive constants C_3, C_4 only depending on q and $\varkappa$.

We can use the above theorems to generalize results concerning the discrepancy of the sequence $(s(q, n)x)_{n=0}^{\infty}$ (cf. [9], [13]) by giving best possible estimates for the discrepancy of the sequence $\tau = (t(n)x)_{n=0}^{\infty}$. Let us recall that the discrepancy of a sequence $\xi = (x_n)_{n=0}^{\infty}$ of real numbers is defined by

$$D_N(\xi) = \sup_{0\leqslant\alpha<\beta\leqslant 1}\left|\frac{A(N; \alpha, \beta, x_n)}{N}-(\beta-\alpha)\right| \tag{1.10}$$

with $A(N; \alpha, \beta, x_n) = \operatorname{card}\{0 \leqslant n < N: \alpha \leqslant x_n-[x_n] < \beta\}$. The sequence $\xi = (x_n)_{n=0}^{\infty}$ is called *uniformly distributed modulo* 1 if $D_N(\xi)$ tends to 0 (for $N \to \infty$); cf. the monographs [7], [8].

THEOREM 3. *The sequence $\tau = (t(n)x)_{n=0}^{\infty}$ is uniformly distributed* mod 1 *if* *and only if* $\omega(hx) \neq 0$ *for all positive integers h. If*

$$\omega(hx) \geqslant C_5/h^\eta$$

for all $h \in N$ *and fixed constants* C_5, $\eta > 0$, *then we have for all* $N \in N$

$$D_N(\tau) \leqslant C_6/(\log N)^{1/2\eta} \tag{1.11}$$

with a constant C_6 *only depending on* q, $\varkappa$, C_5, η. *Conversely we have*

$$D_N(\tau) \geqslant C_7/(\log N)^{1/2\eta} \tag{1.12}$$

for infinitely many $N \in N$ *provided that*

$$\omega(hx) \leqslant C_8/h^\eta \quad \textit{for infinitely many } h \in N.$$

Coquet [2] was heavily interested in the distribution behaviour of the sequences $(s(q, n+k)x)_{n=0}^{\infty}$ uniformly in k. In this special case a (not best-possible) estimate for the uniform discrepancy

$$\tilde{D}_N(\xi) = \sup_{k=0,1,2,\ldots} D_N(\xi^{(k)}) \quad (\xi^{(k)} = (x_{n+k})_{n=0}^{\infty}) \tag{1.13}$$

is known (cf. [14]).

In the following theorem a best possible estimate for $\tilde{D}_N(\tau)$ is established.

THEOREM 4. *For all positive integers N and* $\tau = (t(n)x)_{n=0}^{\infty}$ *the estimate*

$$\tilde{D}_N(\tau) \leqslant 4 \cdot q^\varkappa \max_{1 \leqslant j \leqslant N} D_j(\tau)$$

holds.

Remark 1. As an immediate consequence of Theorem 4 we get the estimates (1.11) and (1.12) even for the uniform discrepancy $\tilde{D}_N(\tau)$.

Remark 2. In the case $\varkappa = 1$ we have

$$\omega(x) = \max_{i=1,\ldots,q-1} \|a(i)x\|;$$

hence the theorem in [9] is a special case of our results.

Remark 3. In the case of Gray code we obtain $\omega(x) = \max(\|x\|, \|2x\|)$. Hence $\chi = (G(h)x)_{n=0}^{\infty}$ is uniformly distributed mod 1 for all irrationals x; χ is even well distributed in this case. If x is of approximation type η then we have

$$\tilde{D}_N(\tau) \leqslant \frac{C_9}{(\log N)^{1/2\eta}} \tag{1.14}$$

with a constant C_9 only depending on η, and this estimate is best possible.

Remark 4. In the case of the Rudin–Shapiro sequence we also obtain $\omega(x) = \max(\|x\|, \|2x\|)$. Hence the same conclusions as in Remark 3 are valid.

2. Proof of Theorem 1. We make use of the following auxiliary results.

LEMMA 1. *There are constants* c_1, $c_2 > 0$ *such that for all* $\mu > 0$ *and all* $z_1, \ldots, z_s \in C$ $(s \geqslant 2)$ *we have*

$$|z_1 + \ldots + z_s| \geqslant (s - c_1\mu^2) \min_{i=1,\ldots,s} |z_i| \tag{2.1}$$

provided that $\max_{i,j=1,\ldots,s} \|\arg z_i - \arg z_j\| \leqslant \mu$;

$$|z_1 + \ldots + z_s| \leqslant (s - c_2\mu^2) \max_{i=1,\ldots,s} |z_i| \tag{2.2}$$

provided that $\max_{i,j=1,\ldots,s} \|\arg z_i - \arg z_j\| \geqslant \mu$. $(\arg r \cdot e^{2\pi i\gamma} := \gamma$ *for* $-1/2 < \gamma \leqslant 1/2$.)

We omit the easy proof.

LEMMA 2. *Let* $a_1(0), \ldots, a_s(0)$ *and* α_{ik} *be complex numbers with* $|a_i(0)| = |\alpha_{ik}| = 1$ *and* $\alpha_{1i} = 1$ $(i, k = 1, \ldots, s;\ s \geqslant 2)$. *Furthermore assume that*

$$a_l(j) = \sum_{i=1}^{s} \alpha_{li} a_i(j-1) \quad \textit{for } l = 1, \ldots, s \textit{ and } j \in N \tag{2.3}$$

and set

$$\nu = \max_{i,j=1,\ldots,s} |\arg \alpha_{ik}|.$$

Then there are constants c_3, $c_4 > 0$ *(only depending on s) such that*

$$|a_l(j)| \leqslant c_3(s - c_4\nu^2)^j \quad \textit{for all } l = 1, \ldots, s \textit{ and } j \in N. \tag{2.4}$$

Proof. We proceed by induction. For $j = 0, 1, 2$ the assertion is trivially true and we assume that it is proved for $j \leqslant n$ $(n \geqslant 2)$. Then by Lemma 1

$$|a_1(n-1)| \leqslant c_3(s - c_5\nu^2)(s - c_4\nu^2)^{n-2}$$

provided that

$$\max_{l,m}(\|\arg a_l(n-2) - \arg a_m(n-2)\|) \geqslant \nu/8.$$

Hence

$$\begin{aligned} |a_l(n)| &\leqslant c_3(s-1)(s-c_4\nu^2)^{n-1} + c_3(s-c_5\nu^2)(s-c_4\nu^2)^{n-2} \\ &\leqslant c_3\left(s - \frac{c_5 - c_4}{s}\nu^2\right)(s-c_4\nu^2)^{n-1}. \end{aligned} \tag{2.5}$$

Now we consider the case

$$\max_{l,m}\left(\|\arg a_l(n-2)-\arg a_m(n-2)\|\right)<v/8$$

and without loss of generality we may assume $\arg\alpha_{s1}=v$. Then

$$\left\|\arg(\alpha_{si}\cdot a_i(n-2))-\arg(\alpha_{s1}\cdot a_1(n-2))\right\|\geqslant v/4-v/8=v/8$$

provided that $\arg\alpha_{si}\leqslant 3v/4$ for some i.

Therefore we obtain as above

$$|a_s(n)|\leqslant c_3\left(s-\frac{c_6-c_4}{s}v^2\right)(s-c_4v^2)^{n-1}. \tag{2.6}$$

In the case $\arg\alpha_{si}>3v/4$ (for all $i=1,\ldots,s$) we have

$$\left\|\arg(a_s(n-1))-\arg(a_1(n-1))\right\|\geqslant v-v/4-2v/8=v/2.$$

Hence we obtain again

$$|a_1(n)|\leqslant c_3(s-c_7v^2)(s-c_4v^2)^{n-1}. \tag{2.7}$$

Combining suitable constants c_1, c_8 yields

$$\begin{aligned}|a_1(n+1)|&\leqslant c_3(s-1)(s-c_4v^2)^n+c_3(s-c_8v^2)(s-c_4v^2)^{n-1}\\&\leqslant c_3\left(s-\frac{c_8-c_4}{s}v^2\right)(s-c_4v^2)^n\leqslant c_3(s-c_4v^2)^{n+1}.\end{aligned}$$

Thus the proof of Lemma 2 is complete.

Now we continue with the proof of Theorem 1. Let N be a positive integer with $q^\varkappa$-ary digit representation

$$N=\sum_{j=0}^{r}N_jq^{\varkappa j},\quad N_j\in\{0,1,\ldots,q^\varkappa-1\},\ N_r\neq 0. \tag{2.8}$$

We have (using the notation $L_j=N_{j+1}q^{\varkappa(j+1)}+\ldots+N_rq^{\varkappa r}$ for $j=0,\ldots,r-1$ and $L_r=0$)

$$\begin{aligned}S(N,x)&=\sum_{j=0}^{r}\sum_{\varepsilon=0}^{N_j-1}\sum_{k=L_j+\varepsilon q^{\varkappa j}}^{L_j+(\varepsilon+1)q^{\varkappa j}-1}e^{2\pi it(k)x}\\&=\sum_{j=0}^{r}\sum_{\varepsilon=0}^{N_j-1}\sum_{l=0}^{q^\varkappa-1}\sum_{k=lq^{\varkappa(j-1)}}^{(l+1)q^{\varkappa(j-1)}-1}\exp_l\left(2\pi it(k+L_j+\varepsilon q^{\varkappa j})x\right)\end{aligned} \tag{2.9}$$

($\exp(t)=e^t$).

Let $\varepsilon=\sum_{j=0}^{\varkappa-1}\varepsilon_jq^j$, $l=\sum_{j=0}^{\varkappa-1}l_jq^j$ be given in q-ary digit representation and

let k be given such that $l\cdot q^{\varkappa(j-1)}\leqslant k<(l+1)q^{\varkappa(j-1)}$, then

$$\begin{aligned}t(k+L_j+\varepsilon q^{\varkappa j})&=t(k)+t(L_j+\varepsilon q^{\varkappa j})\\&+\sum_{m=1}^{\varkappa-1}\left(a(l_m,\ldots,l_{\varkappa-1},\varepsilon_0,\ldots,\varepsilon_{m-1})-a(l_m,\ldots,l_{\varkappa-1},0,\ldots,0)\right).\end{aligned} \tag{2.10}$$

We use the notation

$$\varphi(l,\varepsilon,j,x)=\exp\left(\left(t(k+L_j+\varepsilon q^{\varkappa j})-t(k)\right)x\right),$$

and derive from (2.9)

$$S(N,x)=\sum_{j,\varepsilon,l}\varphi(l,\varepsilon,j,x)\sum_{k=lq^{\varkappa(j-1)}}^{(l+1)q^{\varkappa(j-1)}-1}e^{2\pi it(k)x}. \tag{2.11}$$

Now we consider the sum

$$S_l(j,x)=\sum_{k=lq^{\varkappa j}}^{(l+1)q^{\varkappa j}-1}e^{2\pi it(k)x}\quad\text{for } 0\leqslant l\leqslant q^\varkappa-1$$

and obtain

$$\begin{aligned}S_l(j,x)&=\sum_{b=0}^{q^\varkappa-1}\sum_{k=bq^{\varkappa(j-1)}}^{(b+1)q^{\varkappa(j-1)}-1}e^{2\pi it(k+lq^{\varkappa j})x}\\&=\sum_{b=0}^{q^r-1}\sigma(x,l,b)S_b(j-1,x)\end{aligned} \tag{2.12}$$

(in the notation (1.8)). By Lemma 2 we have for $0\leqslant l\leqslant q^\varkappa-1$

$$|S_l(j,x)|\leqslant c_3\left(q^\varkappa-c_4\omega^2(x)\right)^j,$$

and therefore

$$\begin{aligned}|S(N,x)|&\leqslant c_9\sum_{j=0}^{r}\left(q^\varkappa-c_4\omega^2(x)\right)^{j-1}\\&=c_9\frac{\left(q^\varkappa-c_4\omega^2(x)\right)^r-1}{q^\varkappa-c_4\omega^2(x)-1}\leqslant c_{10}\left(q^\varkappa-c_4\omega^2(x)\right)^{r\varkappa}\\&\leqslant c_{10}\left(q-\frac{c_4}{\varkappa q^\varkappa}\omega^2(x)\right)^{\varkappa r}=c_1N^{\frac{\log(q-c_2\omega^2(x))}{\log q}}.\end{aligned}$$

Thus the proof of Theorem 1 is complete.

3. Proof of Theorem 2. We make use of the following auxiliary results.

LEMMA 3. *For* $z_j=r_je^{2\pi i\alpha_j}$, $\varrho_j=c^{2\pi i\beta_j}$ $(j=1,\ldots,s;\ s\geqslant 2)$ *with* $|\alpha_j|$,

$|\beta_j| \leqslant \pi/8$ let $B = z_1 + \ldots + z_s$ and $B_\varrho = \varrho_1 z_1 + \ldots + \varrho_s z_s$. Then

$$|\arg B_\varrho - \arg B| \leqslant 2 \max_{j=1,\ldots,s} |\beta_j|.$$

Proof. Without loss of generality we assume

$$\tan(\arg B) = \frac{\sum_{i=1}^{s} r_i \sin\alpha_i}{\sum_{i=1}^{s} r_i \cos\alpha_i}, \qquad \tan(\arg B_\varrho) = \frac{\sum_{i=1}^{s} r_i \sin(\alpha_i+\beta_i)}{\sum_{i=1}^{s} r_i \cos(\alpha_i+\beta_i)}$$

and

$$\arg B_\varrho - \arg B \leqslant \tan(\arg B_\varrho) - \tan(\arg B) = \frac{\sum_{i,j=1}^{s} r_i r_j \sin(\alpha_i - \alpha_j + \beta_j)}{\sum_{i,j=1}^{s} r_i r_j \cos\alpha_i \cos(\alpha_j+\beta_j)}.$$

Because of

$$|\alpha_i - \alpha_j| \leqslant \pi/4, \quad |\beta_j| \leqslant \pi/4 \quad \text{and} \quad |\alpha_j + \beta_j| \leqslant \pi/4$$

we have

$$\sin(\alpha_i - \alpha_j + \beta_j) \leqslant \sin(\alpha_i - \alpha_j) + |\beta_j|$$

and

$$\cos\alpha_i \cos(\alpha_j + \beta_j) \geqslant 1/2.$$

Hence

$$\arg B_\varrho - \arg B \leqslant \frac{\sum_{i,j=1}^{s} r_i r_j \sin(\alpha_i - \alpha_j)}{\sum_{i,j=1}^{s} r_i r_j \cos\alpha_i \cos(\alpha_j+\beta_j)} + 2\frac{\sum_{i,j=1}^{s} r_i r_j |\beta_i|}{\sum_{i,j=1}^{s} r_i r_j}$$
$$\leqslant 2 \max_{i=1,\ldots,s} |\beta_i|,$$

and the proof of Lemma 3 is complete.

LEMMA 4. *There are constants δ, c_1, $c_2 > 0$ depending only on s such that wherever $|\arg a_i(0)| < \delta$ and $|\arg \alpha_{ik}| < \delta$ for all i, k then for all l and j*

$$|a_l(j)| \geqslant c_1 (s - c_2 v^2)^j \tag{3.1}$$

(in the notation of Lemma 2).

Proof. If δ is small enough the assertion is trivially true for $j = 0, 1$. We proceed by induction and assume that (3.1) is true for all $j \leqslant n$ $(n \geqslant 1)$.

We obtain by Lemma 3 for sufficiently small δ

$$\|\arg(a_l(m)) - \arg(a_1(m))\| \leqslant 2v$$

for all l and $m = 1, 2, 3, \ldots$ Hence by the induction hypothesis and Lemma 1 we have

$$|a_l(n+1)| \geqslant c_1 (s - c_2 v^2)(s - c_2 v^2)^n = c_1 (s - c_2 v^2)^{n+1}$$

which proves Lemma 4.

By the recurrence (2.12) and Lemma 4 we obtain for all reals x with $\omega(x) < \delta$ and all $N = q^{\varkappa j}$ that

$$|S(N, x)| \geqslant C_3 N^{\frac{\log(q - C_4 \omega^2(x))}{\log q}}$$

which completes the proof of Theorem 2.

4. Proof of Theorems 3 and 4. The first assertion of Theorem 3 follows immediately from Theorems 1, 2 and Weyl's criterion for uniform distribution of sequences (cf. [10]). The upper bound (1.11) can be established by means of the Erdős–Turán inequality

$$D_N(\xi) \leqslant 6\left(\frac{1}{H} + \sum_{h=1}^{H} \frac{1}{h}\left|\frac{1}{N}\sum_{n=0}^{N-1} e^{2\pi i h x_n}\right|\right) \quad (\xi = (x_n)) \tag{4.1}$$

(choosing a suitable positive integer H) by verbally the same calculations as in [9] and [13]. The lower bound (1.12) follows from Koksma's inequality

$$D_N(\xi) \geqslant \frac{1}{2\pi h N}\left|\sum_{n=0}^{N-1} e^{2\pi i h x_n}\right| \tag{4.2}$$

(cf. [8]) as in the special cases [9], [13].

Finally we establish a proof of Theorem 4. For $N, k \in \mathbf{N}$ let a, r be non-negative integers such that $q^{\varkappa(r-1)} \leqslant N < q^{\varkappa r}$ and $aq^{\varkappa r} \leqslant k < (a+1) q^{\varkappa r}$. First we consider the case

$$aq^{\varkappa r} \leqslant k < (k + N - 1) < (a+1) q^{\varkappa r}. \tag{4.3}$$

Let

$$b_0 q^{\varkappa(r-1)} + aq^{\varkappa r} \leqslant k < (b_0 + 1) q^{\varkappa(r-1)} + aq^{\varkappa r}$$

and

$$(b_0 + B_0 q^{\varkappa(r-1)} + aq^{\varkappa r} \leqslant k + N - 1 < (b_0 + B_0 + 1) q^{\varkappa(r-1)} + aq^{\varkappa r}$$

with $0 \leqslant b_0 < b_0 + B_0 < q^{\varkappa}$. Then the sequence $(t(h+k)x)_{n=0}^{N-1}$ consists of the following parts

$$(t(n+k)x)_{n=0}^{(b_0+1)q^{\varkappa(r-1)} + aq^{\varkappa r} - k - 1},$$

$$(4.4) \qquad \left(t(n+k)x\right)_{n=bq^{\varkappa(r-1)}+aq^{\varkappa r}-k}^{(b+1)q^{\varkappa(r-1)}+aq^{\varkappa r}-k-1} \quad \text{for } b = b_0+1, \ldots, b_0+B_0-1$$

and

$$\left(t(n+k)x\right)_{n=(b_0+B_0)q^{\varkappa(r-1)}+aq^{\varkappa r}-k}^{N-1}.$$

Each of these subsequences is of the form $\xi = (\alpha + t(n)x)_{n=n_0}^{n_1}$ where $0 \leqslant \alpha < 1$ and $0 < n_0 \leqslant n_1 \leqslant q^{\varkappa(r-1)}$. Denoting by $D_{n_0,n_1}(\xi)$ the discrepancy of the sequence ξ we obtain by well known (and simple) properties of the discrepancy

$$(n_1 - n_0 + 1) D_{n_0,n_1}(\xi) \leqslant n_1 D_{n_1}(\tau) + (n_0 - n_1) D_{n_0}(\tau).$$

Hence by [8, Theorem 2.6, p. 115] we have

$$ND_N(\tau^{(k)}) \leqslant 2q^{\varkappa} q^{\varkappa(r-1)} \max_{0 \leqslant j \leqslant q^{\varkappa(r-1)}} D_j(\tau);$$

thus

$$(4.5) \qquad \tilde{D}_N(\tau) \leqslant 2q^{\varkappa} \max_{0 \leqslant j \leqslant N} D_j(\tau).$$

In the case $aq^{\varkappa r} \leqslant k < (a+1)q^{\varkappa r} \leqslant k+N-1$ we divide the sequence $(t(n+k)x)_{n=0}^{N-1}$ into two parts

$$(4.6) \qquad \left(t(n+k)x\right)_{n=0}^{(a+1)q^{\varkappa r}-k-1} \quad \text{and} \quad \left(t(n+(a+1)q^{\varkappa r})\right)_{n=0}^{k+N-(a+1)q^{\varkappa r}-1}.$$

Both sequences satisfy the condition of the above case and therefore

$$(4.7) \qquad \tilde{D}_N(\tau) \leqslant 4q^{\varkappa} \max_{0 \leqslant j \leqslant N} D_j(\tau).$$

Thus the proof of Theorem 4 is complete.

5. Appendix. In the following we establish an "explicit" formula for

$$S(n) = \sum_{k=0}^{n-1} G(k) - \frac{n}{2}\left[\frac{\log n}{\log k}\right].$$

First we note that

$$(5.1) \qquad G(2^k + n) = \begin{cases} G(n) + 2 & \text{for } 0 \leqslant n < 2^{k-1}, \\ G(n) & \text{for } 2^{k-1} \leqslant n < 2^k. \end{cases}$$

Applying (5.1) we obtain for

$$A(n) = \sum_{k=1}^{n-1} G(k)$$

the identity

$$\begin{aligned} A(2^s) &= \sum_{1 \leqslant r < 2^{s-1}} G(r) + \sum_{2^{s-1} \leqslant r < 2^s} G(r) \\ &= A(2^{s-1}) + \sum_{2^{s-1} \leqslant r < 2^{s-2}+2^{s-1}} G(r) + \sum_{2^{s-1}+2^{s-2} \leqslant r < 2^s} G(r) \\ &= A(2^{s-1}) + \sum_{0 \leqslant r < 2^{s-1}} G(r) + 2 \sum_{0 \leqslant r < 2^{s-1}} 1 \\ &= 2A(2^{s-1}) + 2^{s-1}. \end{aligned}$$

Hence, by induction,

$$(5.2) \qquad A(2^s) = \frac{s}{2} 2^s.$$

Every odd number has a unique representation of the form

$$(5.3) \qquad n_m = 2^{t_0} + 2^{t_0+t_1} + \ldots + 2^{t_0+\ldots+t_m}$$

with integers $m \geqslant 0$, $t_0 = 0$, $t_j \geqslant 1$ $(j \geqslant 1)$. Furthermore we set

$$(5.4) \qquad n_0 = 1, \quad n_i = 1 + 2^{t_1} + \ldots + 2^{t_1+\ldots+t_i} \quad \text{for } 1 \leqslant i \leqslant m.$$

Then we have

$$\begin{aligned} (5.5) \qquad A(n_m) &= A(2^{t_1+\ldots+t_m}) + \sum_{2^{t_1+\ldots+t_m} \leqslant r < n_m} G(r) \\ &= A(2^{t_1+\ldots+t_m}) + \sum_{0 \leqslant r < n_{m-1}} G(2^{t_1+\ldots+t_m} + r). \end{aligned}$$

In the case $t_m = 1$ (i.e. $n_{m-1} > 2^{t_1+\ldots+t_{m-1}}$) we obtain by (5.1) and (5.5) $(m \geqslant 2)$

$$\begin{aligned} (5.6) \qquad A(n_m) &= A(2^{t_1+\ldots+t_m}) + A(n_{m-1}) + 2 \sum_{0 \leqslant r < 2^{t_1+\ldots+t_{m-1}}} 1 \\ &= A(2^{t_1+\ldots+t_m}) + A(n_{m-1}) + 2(n_{m-1} - n_{m-2}). \end{aligned}$$

In the case $t_m > 1$ (i.e. $n_{m-1} < 2^{t_1+\ldots+t_{m-1}}$) we obtain by (5.1) and (5.5)

$$\begin{aligned} (5.7) \qquad A(n_m) &= A(2^{t_1+\ldots+t_m}) + \sum_{0 \leqslant r < n_{m-1}} (G(r) + 2) \\ &= A(2^{t_1+\ldots+t_m}) + A(n_{m-1}) + 2n_{m-1}. \end{aligned}$$

Combining (5.6) and (5.7) gives

$$(5.8) \qquad A(n_m) = A(2^{t_1+\ldots+t_m}) + A(n_{m-1}) + 2n_{m-1} - 2\delta_m n_{m-2} \qquad (m \geqslant 1),$$

where

$$\delta_m = \begin{cases} 0 & \text{for } t_m > 1, \\ 1 & \text{for } t_m = 1, \end{cases}$$

$$n_{-1} = 0.$$

Summing up and applying (5.2) (note that $A(n_0) = 0$) yields

$$\begin{aligned} A(n_m) &= \sum_{r=1}^{m} (A(2^{t_1+\dots+t_r}) + 2n_{r-1}) - 2\sum_{r=1}^{m} \delta_r n_{r-2} \\ &= \sum_{r=1}^{m} \left(\frac{t_1+\dots+t_r}{2} 2^{t_1+\dots+t_r} + 2n_{r-1}\right) - 2\sum_{r=1}^{m} \delta_r n_{r-2} \\ &= \sum_{r=1}^{m} \left(\frac{t_1+\dots+t_r}{2}(n_r - n_{r-1}) + 2n_{r-1}\right) - 2\sum_{r=1}^{m} \delta_r n_{r-2} \\ &= \sum_{r=1}^{m} \left(2 - \frac{t_r}{2}\right) n_{r-1} + \frac{(t_1+\dots+t_m) n_m}{2} - 2\sum_{r=1}^{m} \delta_r n_{r-2}. \end{aligned}$$

Observing that $t_r + \dots + t_m = \left[\dfrac{\log n_m}{\log 2}\right]$ we have proved the following formula

$$\frac{S(n_m)}{n_m} = \frac{1}{n_m} \sum_{r=1}^{m} \left(\left(2 - \frac{t_r}{2}\right) n_{r-1} - 2\delta_r n_{r-2}\right). \tag{5.10}$$

It should be remarked that the expression (5.10) does not change its value if n_m is replaced by $2^\beta n_m$ (β an arbitrary positive integer). Hence the explicit formula (2.10) is valid also in the general case (not only for odd numbers n_m).

Since $n_{r-1}/n_r \leqslant 1/2$ (for every $r \geqslant 1$), an application of Foster's lower bound [6, Theorem 1] yields

$$\frac{S(n_m)}{n_m} = \frac{1}{2n_m} \sum_{r=1}^{m} (2 - t_r) n_{r-1} + \frac{1}{n_m} \sum_{r=1}^{m} (n_{r-1} - 2\delta_r n_{r-2}) \geqslant -\frac{1}{13}. \tag{5.11}$$

An elementary observation shows that $S(n_m)/n_m$ increases if all $t_r > 2$ are replaced by $t_r = 2$. Hence, for determining an upper bound for $S(n_m)/n_m$ it suffices to consider the case $t_r = 1$ or 2 $(r = 1, \dots, m)$. We prove $S(n_m)/n_m \leqslant 7/10$ by induction; the cases $m = 0, 1$ are trivial. Assuming this bound for $m-1$ we obtain in the case $t_m = 2$ by (5.10)

$$\begin{aligned} \frac{S(n_m)}{n_m} &\leqslant \frac{1}{n_m}\left(S(n_{m-1}) + \left(2 - \frac{t_m}{2}\right) n_{m-1} - 2\delta_m n_{m-2}\right) \\ &\leqslant \frac{7}{10}\frac{n_{m-1}}{n_m} + \frac{n_{m-1}}{n_m} \leqslant \frac{7}{10}, \end{aligned} \tag{5.12}$$

since $n_{m-1}/n_m \leqslant 1/(1+2^{t_m - 1})$.

In the case $t_m = 1$ we have

$$\frac{S(n_m)}{n_m} = \frac{n_{m-1}}{n_m}\left(\frac{S(n_{m-1})}{n_{m-1}} - \frac{1}{2}\right) + 2\frac{n_{m-1} - n_{m-2}}{n_m}.$$

The expression $(n_{m-1} - n_{m-2})/n_m$ takes its maximal value if $t_1 = \dots = t_l = 1$ and $t_{l+1} = \dots = t_{m-1}$ (for some l with $1 \leqslant l \leqslant m-1$). Hence after a simple calculation

$$\frac{n_{m-1} - n_{m-2}}{n_m} = \frac{2^{2m-l-2}}{(2^{l+1}-1) + \frac{1}{3}2^{l+2}(4^{m-l-1}-1) + 2^{2m-l-1}} \leqslant \frac{3}{10}.$$

From this estimate and the induction hypothesis we derive

$$\frac{S(n_m)}{n_m} \leqslant \frac{1}{2}\left(\frac{7}{10} - \frac{1}{2}\right) + \frac{3}{5} = \frac{7}{10}, \tag{5.13}$$

which completes the induction argument. Combining (5.11) and (5.13) we have for all positive integers n (note the remark after (5.10))

$$-\frac{1}{13} \leqslant \frac{S(n)}{n} \leqslant \frac{7}{10}. \tag{5.14}$$

These bounds seem to be quite far away from the optimal bounds. It may be possible to derive the optimal bounds from (5.10) using inductive arguments similar to Foster's [6]. However, some numerical calculations become rather extensive in this case.

References

[1] J. P. Allouche and M. Mendès-France, *On an extremal property of the Rudin–Shapiro sequence*, Mathematika 32 (1985), 33–38.

[2] J. Coquet, *Sur certaines suites uniformément équireparties modulo* 1, Acta Arith. 36 (1980), 157–162.

[3] H. Delange, *Sur la fonction sommatoire de la fonction 'somme des chiffres'*, Enseignement Math. 2.21 (1975), 31–47.

[4] M. P. Drazin and J. S. Griffiths, *On the decimal representation of integers*, Proc. Cam. Phil. Soc., (4), 48 (1952), 555–565.

[5] P. Flajolet and L. Ramshaw, *Gray code and odd–even merge*, SIAM J. Comput. 9 (1980), 142–158.

[6] D. M. E. Foster, *Estimates for a remainder term associated with the sum of digits function*, to appear in Glasgow Math. J.

[7] E. Hlawka, *Theorie der Gleichverteilung*, Bibl. Inst. Mannheim–Wien–Zürich, 1979.

[8] L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*, Wiley and Sons, New York 1974.

[9] G. Larcher, *Exponential sums of digit-depending sequences and uniform distribution*, to appear.

[10] M. Mendès-France, *Nombres normaux. Applications aux fonctions pseudo-aléatoires*. J. Analyse Math. 20 (1967), 1–56.

[11] R. Sedgewick, *Data movement in odd-even merging*, SIAM J. Comput. 7 (1978), 239–272.
[12] K. B. Stolarsky, *Power and exponential sums related to binomial digit parity*, SIAM J. Appl. Math. 32 (1977), 717–730.
[13] R. F. Tichy and G. Turnwald, *On the discrepancy of some special sequences*, J. Number Theory 26 (1987), 68–78.
[14] –, – *Gleichmässige Diskrepanzabschätzung für Ziffernsummen*, Anz. Österr. Akad. Wiss. (1986), 17–21.

MATHEMATISCHES INSTITUT
DER UNIVERSITÄT SALZBURG
Hellbrunner Strasse 34
A-5020 Salzburg, Austria

ABTEILUNG FÜR TECHNISCHE MATHEMATIK
TU WIEN
Wiedner Hauptstrasse 8–10
A-1040 Wien, Austria

Received on 14.8.1987 (1744)

ACTA ARITHMETICA
LII (1989)

On the number of values taken by a polynomial over a finite field

by

J. F. Voloch (Rio de Janeiro)

Let $\boldsymbol{F}_q$ be the finite field with q elements and $f(x) \in \boldsymbol{F}_q[x]$ a polynomial of degree n. Let $r(f) = \#f(\boldsymbol{F}_q)$, considering f as a function $f\colon \boldsymbol{F}_q \to \boldsymbol{F}_q$. A classical problem, raised by Chowla [3] (see [4] for other references), is to estimate $r(f)$ an terms of n and q. One has the trivial bounds $q/n \leqslant r(f) \leqslant q$. The lower bound is essentially best possible and a characterization of the cases with equality when q is prime was obtained in [2].

On the other hand, if f is a "general" polynomial (in a sense that can be made precise, see below) Uchiyama [6] proved that $r(f) \geqslant q/2 + O(q^{1/2})$ and Birch and Swinnerton-Dyer [1] found the precise result

$$r(f) = q\left(\sum_{i=1}^{n} \frac{(-1)^{i-1}}{i!}\right) + O(q^{1/2}).$$

They proved this when the Galois group of $f(x) = y$ over $\bar{\boldsymbol{F}}_q(y)$ is the full symmetric group. Of course these results are interesting only when q is large compared to n. The purpose of this paper is to give lower bounds for $r(f)$, valid for f "general", which improves on the above bounds in several cases.

Uchiyama's condition is that the polynomial

$$f^*(u, v) = (f(u) - f(v))/(u - v)$$

is absolutely irreducible. When this is the case he could apply Weil's estimate ([7]) on the number of points of $f^*(u, v) = 0$ over $\boldsymbol{F}_q$ to get his result.

To relate the number of solutions of $f^*(u, v) = 0$ in $\boldsymbol{F}_q^2$ with $r(f)$, Uchiyama [6] proved the following:

LEMMA 1. *Let N be the number of solutions of $f^*(u, v) = 0$ in $\boldsymbol{F}_q^2$ and n_0 the number of solutions of $f'(x) = 0$ in $\boldsymbol{F}_q$. Then*

$$r(f) \geqslant q^2/(N + q - n_0).$$

Proof. First notice that $f^*(u, v) = 0$ and $u \neq v$ if and only if $f(u) = f(v)$ and that $f^*(u, u) = f'(u)$. Let $\{a_1, \ldots, a_r\} = f(\boldsymbol{F}_q)$, $r = r(f)$ and n_i