Applicable Analysis and Discrete Mathematics
available online at http://pefmath.etf.rs

Appl. Anal. Discrete Math. **13** (2019), 346–360.
https://doi.org/10.2298/AADM180720014D

ON THE SIZE OF A RESTRICTED SUMSET WITH APPLICATION TO THE BINARY EXPANSION OF $\sqrt{d}$

Artūras Dubickas

For any $A \subseteq \mathbb{N}$, let $U(A,N)$ be the number of its elements not exceeding N. Suppose that $A+A$ has $V(A,N)$ elements not exceeding N, where the elements in the sumset $A+A$ are counted with multiplicities. We first prove a sharp inequality between the size of $U(A,N)$ and that of $V(A,N)$ which, for the upper limits $\omega(A) = \limsup_{N\to\infty} U(A,N)N^{-1/2}$ and $\sigma(A) = \limsup_{N\to\infty} V(A,N)N^{-1}$, implies $\omega(A)^2 \geq 4\sigma(A)/\pi$. Then, as an application, we show that, for any square-free integer $d > 1$ and any $\varepsilon > 0$, there are infinitely many positive integers N such that at least $(\sqrt{8/\pi}-\varepsilon)\sqrt{N}$ digits among the first N digits of the binary expansion of $\sqrt{d}$ are equal to 1.

1. INTRODUCTION

In this paper, for $A = \{a_1 < a_2 < a_3 < \dots\} \subseteq \mathbb{N}$, where $\mathbb{N}$ is the set of positive integers, we will compare the quantities

$$U(A,N) := \#\{i : a_i \leq N\} = \#\{A \cap [1,N]\}$$

and

$$V(A,N) := \#\{(i,j) : a_i + a_j \leq N\},$$

where $\#\mathcal{S}$ stands for the cardinality of a finite set $\mathcal{S}$. The first quantity is simply the number of elements of the set $A \cap [1,N]$, whereas the second counts the number of elements in $(A+A) \cap [1,N]$, where the elements of the sumset $A+A$ are counted with multiplicities.

2010 Mathematics Subject Classification. 11K16, 11K31, 11B13.
Keywords and Phrases. Sumset, binary expansion, quadratic irrationality, Sidon sequence.

Let us fix α in the interval $(0,2]$ and consider the upper limits

$$\omega_\alpha(A) := \limsup_{N\to\infty} \frac{U(A,N)}{N^{\alpha/2}} \quad \text{and} \quad \sigma_\alpha(A) := \limsup_{N\to\infty} \frac{V(A,N)}{N^{\alpha}}.$$

Note that, for any $A \subseteq \mathbb{N}$, we have $\omega_\alpha(A) = 0$ when $\alpha > 2$, so the restriction $\alpha \le 2$ is natural. Furthermore, for $\alpha = 2$, by the trivial inequality $\#U(A,N) \le N$, one has $\omega_2(A) \le 1$.

With this notation, we will prove the following inequality between $\omega_\alpha(A)$ and $\sigma_\alpha(A)$.

Theorem 1. *For each $\alpha \in (0,2]$ and any $A \subseteq \mathbb{N}$, we have*

$$\omega_\alpha(A)^2 \ge \frac{4\Gamma(\alpha)}{\alpha\Gamma(\alpha/2)^2}\sigma_\alpha(A). \tag{1}$$

Furthermore, for each $\alpha \in (0,2]$, the constant

$$K(\alpha) := \frac{4\Gamma(\alpha)}{\alpha\Gamma(\alpha/2)^2} \tag{2}$$

in (1) is best possible in the sense that, for every $\omega \in (0,\infty)$ when $0 < \alpha < 2$ and for every $\omega \in (0,1]$ when $\alpha = 2$, there exists $A \subseteq \mathbb{N}$ such that $\omega_\alpha(A) = \omega$ and $\sigma_\alpha(A) = \omega^2/K(\alpha)$.

One can also compare the quantities $\omega_\alpha(A)$ and $\omega_{2\alpha}(A+A)$, where the elements in $(A+A)\cap[1,N]$ are counted in the usual way (without multiplicities), so that

$$\omega_{2\alpha}(A+A) = \limsup_{N\to\infty} \frac{\#\{(A+A)\cap[1,N]\}}{N^\alpha}.$$

Of course, $\omega_{2\alpha}(A+A) = 0$ for $\alpha > 1$, so in the next statement α is restricted to the interval $(0,1]$.

Corollary 2. *For each $\alpha \in (0,1]$ and any $A \subseteq \mathbb{N}$, we have*

$$\omega_\alpha(A)^2 \ge \frac{8\Gamma(\alpha)}{\alpha\Gamma(\alpha/2)^2}\omega_{2\alpha}(A+A). \tag{3}$$

Furthermore, for each $\alpha \in (0,1/2)$, the constant in (3) is best possible.

Note that $K(1) = 4/\pi$, since the values of the gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ at 1 and $1/2$ are 1 and $\sqrt{\pi}$, respectively. Thus, Theorem 1 with $\alpha = 1$ implies the inequality

$$\omega_1(A)^2 \ge \frac{4}{\pi}\sigma_1(A). \tag{4}$$

As an application of (4), we will derive the following:

Theorem 3. *Let $d > 1$ be a square-free integer, and let $D(\sqrt{d}, N)$ be the number of digits equal to 1 among the first N digits in the binary expansion of $\sqrt{d}$. Then, there is $N_0(d) \in \mathbb{N}$ such that*

$$D(\sqrt{d}, N) > \sqrt{2N} - 2 \tag{5}$$

for every $N \geq N_0(d)$. Furthermore,

$$\limsup_{N\to\infty} \frac{D(\sqrt{d}, N)}{\sqrt{N}} \geq \sqrt{\frac{8}{\pi}}. \tag{6}$$

The distribution of digits of a given irrational number (in its decimal or binary expansion), e.g., $\sqrt{2}$, π, e, etc., is a completely open problem. One expects that those numbers are normal (see, e.g., [**2**], [**4**], [**10**]), but the best known results are very far from this. In [**3**], it was shown that the binary expansion of an algebraic number β of degree $\deg\beta \geq 2$ has at least $c(\beta)N^{1/\deg\beta}$ units among the first N digits. Various generalizations of this result have been given in [**1**], [**7**], [**8**], [**9**], [**11**] (see also [**5**, Theorem 8.5]). In particular, in [**1**] and [**5**] it was shown that the constant $c(\beta)$ is effectively computable. From the proof one can see that the constant $N_0(d)$ of Theorem 3 is effectively computable.

Recently, Vandehey [**13**], with the notation of Theorem 3, showed that

$$D(\sqrt{2}, N) \geq (\sqrt{2} - \varepsilon)\sqrt{N}$$

for each $N \geq N_1(\varepsilon)$ and

$$D(\sqrt{2}, N) \geq \Big(\frac{2}{\sqrt{2\sqrt{2}-1}} - \varepsilon\Big)\sqrt{N}$$

for infinitely many $N \in \mathbb{N}$. The result with $\sqrt{8/\pi}$ is also announced in [**13**]. Theorem 3 shows that this is indeed true, and not only for $\sqrt{2}$, but for arbitrary $\sqrt{d}$, where $d > 1$ is square-free, as well. In this direction, a result of Rivoal implies that $D(\sqrt{d}, N) > (1 + o(1))\sqrt{N}$, where $o(1) \to 1$ as $N \to \infty$ (see Corollary 2 in [**11**]). Note that $\sqrt{2} = 1.414213\ldots$, $2/\sqrt{2\sqrt{2}-1} = 1.479078\ldots$ and $\sqrt{8/\pi} = 1.595769\ldots$. In fact, the proof of Theorem 3 follows that of a particular result for $d = 2$ given in [**13**].

In the next section we will state and solve a minimax type problem (Lemma 4) which is a key ingredient in the proof of the inequalities (1) and (3). Then, in Sections 3, 4 and 5, we will prove Theorem 1, Corollary 2 and Theorem 3, respectively. The proof of Lemma 4 is self-contained. In the proof of Theorem 1 we first reduce the statement to a form when Lemma 4 can be applied and then use some identities for Euler's beta function and gamma function. In the proof of the optimality of the constant in Corollary 2 we use, in addition, a result of Ruzsa [**12**] on Sidon sequences of polynomial type. An extension of such a result would give a wider range for α for which the constant in (3) is best possible.

1. A MINIMAX PROBLEM AND ITS SOLUTION

Lemma 4. *Let m be a positive integer, and let*

$$0 < r_1 < r_2 < \cdots < r_{2m}. \tag{7}$$

Let Ω be a subset of $\mathbb{R}^{2m}$ consisting of the vectors $(z_1, z_2, \dots, z_{2m}) \in \mathbb{R}^{2m}$ satisfying

$$0 \le r_1 z_1 \le r_2 z_2 \le \cdots \le r_{2m} z_{2m} \tag{8}$$

and

$$P(z_1, z_2, \dots, z_{2m}) = 1, \tag{9}$$

where

$$P(z_1, z_2, \dots, z_{2m}) := r_m^2 z_m^2 + 2\sum_{j=1}^{m} r_j z_j (r_{2m+1-j} z_{2m+1-j} - r_{2m-j} z_{2m-j}).$$

Then, for each $(z_1, z_2, \dots, z_{2m}) \in \Omega$, we have

$$\max(z_1, z_2, \dots, z_{2m}) \ge \frac{1}{\sqrt{P(1, 1, \dots, 1)}}. \tag{10}$$

Note that equality in (10) is attained for

$$z_1 = z_2 = \cdots = z_{2m} = \frac{1}{\sqrt{P(1, 1, \dots, 1)}},$$

so that

$$\min_{(z_1, \dots, z_{2m}) \in \Omega} \quad \max_{1 \le j \le 2m} z_j = \frac{1}{\sqrt{P(1, 1, \dots, 1)}}. \tag{11}$$

Proof. We first prove the statement for $m = 1$. By (9), from

$$r_1^2 z_1^2 + 2r_1 z_1 (r_2 z_2 - r_1 z_1) = 1,$$

it follows that $z_1 \ne 0$ and

$$2r_2 z_2 = \frac{r_1^2 z_1^2 + 1}{r_1 z_1}.$$

Since $P(1, 1) = r_1(2r_2 - r_1)$, the inequality (10) is satisfied in the case when

$$z_1^2 \ge \frac{1}{r_1(2r_2 - r_1)}.$$

Assume that the opposite inequality holds, namely, $r_1^2 z_1^2 < 1/(2q-1)$, where $q := r_2/r_1 > 1$. Then, taking into account $1/(2q-1) < 1$ and the fact that the function $x + x^{-1}$ is decreasing in the interval $(0,1)$, we deduce that

$$4r_2^2 z_2^2 = r_1^2 z_1^2 + \frac{1}{r_1^2 z_1^2} + 2 > 2q - 1 + \frac{1}{2q-1} + 2 = \frac{4q^2}{2q-1}.$$

Hence,

$$z_2^2 > \frac{4q^2}{4r_2^2(2q-1)} = \frac{4(r_2/r_1)^2}{4r_2^2(2r_2/r_1 - 1)} = \frac{1}{r_1(2r_2 - r_1)} = \frac{1}{P(1,1)},$$

which is stronger than (10).

Now, let $m \geq 2$ and assume (10) holds for m replaced by $m-1$. We claim that the minimum of the expression $\max(z_1, z_2, \dots, z_{2m})$ when $(z_1, z_2, \dots, z_{2m}) \in \Omega$ (which is the left hand side of (11)) is attained at some (not necessarily unique) point

$$(z_1^*, z_2^*, \dots, z_{2m}^*) \in \Omega.$$

Indeed, take any particular point $(w_1, \dots, w_{2m}) \in \Omega$, and set $M_0 := \max_{1 \leq i \leq 2m} w_i$. The intersection of the set Ω defined in (8) and (9) with the cube $[0, M_0]^{2m}$ is compact, and so the minimum of the left hand side of (11) is attained. Note that if the coordinates of this point are all equal, that is, $z_1^* = z_2^* = \dots = z_{2m}^*$, then, by (9) and the definition of P, they must all be equal to $1/\sqrt{P(1,1,\dots,1)}$. This implies (10).

Suppose $z_1^*, z_2^*, \dots, z_{2m}^*$ are not all equal and the minimum (of the maximum $\max(z_1, z_2, \dots, z_{2m})$ when $(z_1, z_2, \dots, z_{2m}) \in \Omega$) is equal to

$$M = \max(z_1^*, z_2^*, \dots, z_{2m}^*) < \frac{1}{\sqrt{P(1,1,\dots,1)}}. \tag{12}$$

Then, $z_i^* < M$ for some $i \in \{1, 2, \dots, 2m\}$. In all what follows we will show that then, by a small perturbation of the coordinates (that is, by slightly changing z_i^* so that the new vector is in Ω and satisfies (9)) we can decrease the value of M and so get a contradiction with the minimality of $\max(z_1, z_2, \dots, z_{2m})$ in Ω.

Suppose first that $z_1^* = 0$. Let us remove z_1 and z_{2m} and consider the numbers $z_2, \dots, z_{2m-1}$ only. Note that, with two variables z_1, z_{2m} removed and (8) replaced by $0 \leq r_2 z_2 \leq \dots \leq r_{2m-1} z_{2m-1}$, we have the smaller sum

$$S_{2m-2}^* := P(0, \underbrace{1, \dots, 1}_{2m-2}, 0) < S_{2m} := P(\underbrace{1, 1, \dots, 1}_{2m}),$$

because $S_{2m} - S_{2m-2}^* = 2r_1(r_{2m} - r_{2m-1}) > 0$. Applying (10) to $2m-2$ numbers $z_2, \dots, z_{2m-1}$ and Ω replaced by $\Omega' \subset \mathbb{R}^{2m-2}$, by induction hypotheses on m (namely, the assumption that (10) holds for $m-1$), we have

$$\max(z_2, \dots, z_{2m-1}) \geq \frac{1}{\sqrt{S_{2m-2}^*}} > \frac{1}{\sqrt{P(1,1,\dots,1)}},$$

which contradicts (12). Hence, $z_1^* > 0$. Consequently, by the definition of Ω (see (8)), we must have $z_2^*, \dots, z_{2m}^* > 0$ as well.

Assume first that there is some $j \in \{m+1, \dots, 2m\}$ satisfying $z_j^* < M$. Take the largest such j. By the definition, P is a linear polynomial in z_j^*. So, by (9), there are $U, V \in \mathbb{R}[z_1^*, \dots, z_{j-1}^*, z_{j+1}^*, \dots, z_{2m}^*]$ such that

$$1 = Uz_j^* + V. \tag{13}$$

Consider two cases $U = 0$ and $U \neq 0$. If $U = 0$ we can simply replace z_j^* by M. The condition (8), that is, $0 \le r_1 z_1^* \le \dots \le r_{2m} z_{2m}^*$ will be satisfied, since $z_{j+1}^* = \dots = z_{2m}^* = M$. The condition (9) (or, more precisely, $P(z_1^*, \dots, z_{2m}^*) = 1$) will be satisfied too. Thus, the new point $(z_1^*, \dots, z_{j-1}^*, M, \dots, M)$ belongs to Ω and $z_j^* = M$. In case $U \neq 0$ we can write $z_j^* = (1-V)/U$ by (13). Replace each z_i^*, $i \neq j$, by $z_i' = z_i^*/(1+\varepsilon)$, where $\varepsilon > 0$ is so small that the new z_j' obtained as the value of $(1-V)/U$ satisfies

$$\frac{r_{j-1} z_{j-1}^*}{1+\varepsilon} \le r_j z_j' \le \frac{r_{j+1} M}{1+\varepsilon}.$$

This is clearly possible if we take ε small enough. In this way we get a new vector $(z_1', \dots, z_{2m}') \in \Omega$ close to $(z_1^*, \dots, z_{2m}^*) \in \Omega$, with $\max(z_1', \dots, z_{2m}') < M$, which contradicts to the minimality of M. This proves that there is no such j and therefore $z_{m+1}^* = \dots = z_{2m}^* = M$.

Assume next that $z_m^* < M$. By the definition of P and (9), there is $W \in \mathbb{R}[z_1^*, \dots, z_{m-1}^*, z_{m+1}^*, \dots, z_{2m}^*]$ such that

$$1 = -r_m^2 {z_m^*}^2 + 2 r_m z_m^* r_{m+1} z_{m+1}^* + W.$$

Setting $W_1 := (1-W)/r_m^2$, we obtain the quadratic equation

$${z_m^*}^2 - 2\frac{r_{m+1}}{r_m} z_{m+1}^* z_m^* + W_1 = 0.$$

In view of $r_m z_m^* \le r_{m+1} z_{m+1}^*$ this leads to

$$z_m^* = \frac{r_{m+1}}{r_m} z_{m+1}^* - \sqrt{W_2},$$

where $W_2 := (r_{m+1} z_{m+1}^*/r_m)^2 - W_1$. Here, $W_2 > 0$, since $z_{m+1}^* = M$ and $r_m < r_{m+1}$ implies $r_m z_m^* < r_{m+1} M$. Now, as above, let us replace each z_i^*, $i \neq m$, by $z_i' = z_i^*/(1+\varepsilon)$, where $\varepsilon > 0$ is so small that the new z_m' obtained as the value of $r_{m+1} z_{m+1}'/r_m - \sqrt{W_2}$ satisfies

$$\frac{r_{m-1} z_{m-1}^*}{1+\varepsilon} \le r_m z_m' \le \frac{r_{m+1} M}{1+\varepsilon}.$$

This is possible for ε small enough. So, as above, we get a new vector $(z_1', \dots, z_{2m}') \in \Omega$ close to $(z_1^*, \dots, z_{2m}^*) \in \Omega$, with $\max(z_1', \dots, z_{2m}') < M$, which contradicts to the minimality of M. This proves that $z_m^* = M$.

Finally, assume that there is some $j \in \{1, \ldots, m-1\}$ satisfying $z_j^* < M$. Take the largest such j. By the definition, P is a linear polynomial in z_j^*. In this way we get (13) and obtain a contradiction by the same argument as in the case $j \in \{m+1, \ldots, 2m\}$. Therefore, z_j^*, $j = 1, \ldots, 2m$, must be all equal to M, contrary to our assumption. □

2. PROOF OF THEOREM 1

Take an increasing sequence N_k, $k = 1, 2, \ldots$, of $\mathbb{N}$ such that

$$V(A, N_k)/N_k^\alpha \to \sigma_\alpha(A) \quad \text{as} \quad k \to \infty.$$

Fix $m \in \mathbb{N}$ and take $N = N_k$ with k large enough. Consider the $2m$ intervals

$$I_j := \big((j-1)N/(2m), jN/(2m)\big],$$

where $j = 1, \ldots, 2m$. Suppose I_j contains e_j elements of A. Then,

$$e_j = U(A, jN/(2m)) - U(A, (j-1)N/(2m)).$$

Setting $s_j = e_1 + \cdots + e_j = U(A, jN/(2m))$, by the definitions of $V(A, N)$ and $U(A, N)$, we have

$$V(A, N) \le s_m^2 + 2(e_{2m}s_1 + e_{2m-1}s_2 + \cdots + e_{m+1}s_m).$$

Put

$$y_j := \frac{U(A, jN/(2m))}{U(A, N)} = \frac{s_j}{U(A, N)}$$

for $j = 1, 2, \ldots, 2m$. Clearly,

$$0 \le y_1 \le y_2 \le \cdots \le y_{2m-1} \le y_{2m} = 1. \tag{14}$$

Using $e_j = s_j - s_{j-1}$, we deduce that

$$V(A, N) \le U(A, N)^2 \Big(y_m^2 + 2\sum_{j=1}^{m} y_j (y_{2m+1-j} - y_{2m-j})\Big). \tag{15}$$

Setting

$$Y_m := y_m^2 + 2\sum_{j=1}^{m} y_j (y_{2m+1-j} - y_{2m-j}) \tag{16}$$

and dividing (15) by $N^\alpha Y_m$, we get

$$\frac{U(A, N)^2}{N^\alpha} \ge \frac{V(A, N)}{Y_m N^\alpha}.$$

This implies the result when $\sigma_\alpha(A) = 0$ or $\sigma_\alpha(A) = \infty$. From now on, we assume that $0 < \sigma_\alpha(A) < \infty$.

Similarly, for each $j = 1, \dots, 2m$, from (15) and (16), it follows that

$$\frac{U(A, jN/(2m))^2}{(jN/(2m))^\alpha} = \frac{y_j^2 U(A,N)^2}{(jN/(2m))^\alpha} \geq \frac{(2m/j)^\alpha y_j^2 V(A,N)}{Y_m N^\alpha}.$$

For each $j = 1, \dots, 2m$, letting $N = N_k \to \infty$, we find that

$$\omega_\alpha(A)^2 \geq \limsup_{N_k \to \infty} \frac{U(A, jN_k/(2m))^2}{(jN_k/(2m))^\alpha} \geq \frac{(2m/j)^\alpha y_j^2}{Y_m} \sigma_\alpha(A).$$

So, putting

$$C_m := \max_{1 \leq j \leq 2m} \frac{(2m/j)^\alpha y_j^2}{Y_m} \tag{17}$$

and using $\sigma_\alpha(A) \neq 0$, we deduce that

$$\frac{\omega_\alpha(A)^2}{\sigma_\alpha(A)} \geq C_m$$

for each fixed $m \in \mathbb{N}$.

Suppose

$$\frac{\omega_\alpha(A)^2}{\sigma_\alpha(A)} < K(\alpha),$$

where $K(\alpha)$ is the constant defined in (2). Then, there is a positive ε such that

$$\frac{\omega_\alpha(A)^2}{\sigma_\alpha(A)} < K(\alpha) - \varepsilon,$$

and so

$$C_m < K(\alpha) - \varepsilon. \tag{18}$$

We will show, however, that (18) does not hold for sufficiently large m, and so get a contradiction.

Put

$$z_j := \frac{y_j (2m/j)^{\alpha/2}}{\sqrt{Y_m}}. \tag{19}$$

Inserting $y_j = z_j \sqrt{Y_m} r_j$, where $r_j := (j/(2m))^{\alpha/2}$, into (16) we deduce that

$$1 = r_m^2 z_m^2 + 2 \sum_{j=1}^{m} r_j z_j (r_{2m+1-j} z_{2m+1-j} - r_{2m-j} z_{2m-j}).$$

Furthermore, by (14), we have

$$0 \le r_1 z_1 \le r_2 z_2 \le \cdots \le r_{2m} z_{2m}.$$

Hence, by Lemma 4 applied to $r_j = (j/(2m))^{\alpha/2}$, $j = 1, \dots, 2m$, and (17), (19), we deduce that

$$C_m = \max(z_1^2, z_2^2, \dots, z_{2m}^2) \ge \frac{1}{S_{2m}},$$

where

$$\begin{aligned} S_{2m} := P(\underbrace{1, 1, \dots, 1}_{2m}) &= r_m^2 + 2 \sum_{j=1}^{m} r_j (r_{2m+1-j} - r_{2m-j}) \\ &= 2^{-\alpha} + \frac{2}{(2m)^\alpha} \sum_{j=1}^{m} j^{\alpha/2} \big((2m+1-j)^{\alpha/2} - (2m-j)^{\alpha/2} \big). \end{aligned}$$

Our aim is to show that

$$S_{2m} < \frac{1}{K(\alpha)} + O\big(\frac{1}{m}\big). \tag{20}$$

Then, by (10) and (20), one gets the inequality

$$C_m = \max(z_1^2, z_2^2, \dots, z_{2m}^2) > K(\alpha) + O\big(\frac{1}{m}\big) \tag{21}$$

for each $m \in \mathbb{N}$, which gives the desired contradiction to (18).

To prove (20) let us first observe that, by the mean value theorem and $\alpha/2 - 1 \le 0$, for $j = 1, \dots, m$, one has

$$(2m+1-j)^{\alpha/2} - (2m-j)^{\alpha/2} = \frac{\alpha}{2} (2m - j + \theta)^{\alpha/2 - 1} \le \frac{\alpha}{2} (2m-j)^{\alpha/2-1},$$

where $\theta = \theta_{m,j,\alpha} \in [0, 1]$. Consequently,

$$S_{2m} - 2^{-\alpha} \le \frac{\alpha}{(2m)^\alpha} \sum_{j=1}^{m} j^{\alpha/2} (2m-j)^{\alpha/2-1}. \tag{22}$$

Note that the right hand side of (22),

$$\frac{\alpha}{2m} \sum_{j=1}^{m} \Big(\frac{j}{2m}\Big)^{\alpha/2} \Big(1 - \Big(\frac{j}{2m}\Big)\Big)^{\alpha/2-1},$$

is the right Riemann sum of the increasing function

$$\varphi(x) := \alpha x^{\alpha/2} (1-x)^{\alpha/2-1}$$

in the interval $[0,1/2]$. Hence, from (22), we further deduce that

$$\begin{aligned} S_{2m} &\le 2^{-\alpha} + \int_0^{1/2} \varphi(x)dx + \frac{1}{2m}\big(\varphi(1/2) - \varphi(0)\big) \\ &= 2^{-\alpha} + \alpha \int_0^{1/2} x^{\alpha/2}(1-x)^{\alpha/2-1}dx + \frac{\alpha 2^{-\alpha}}{m}. \end{aligned}$$

In order to prove (20) we need to show that

$$2^{-\alpha} + \alpha \int_0^{1/2} x^{\alpha/2}(1-x)^{\alpha/2-1}dx = \frac{1}{K(\alpha)}, \tag{23}$$

where $K(\alpha)$ is defined in (2).

Let us first evaluate the integral

$$J(a) := \int_0^{1/2} x^a(1-x)^{a-1}dx$$

for $a > 0$. Defining

$$J_1(a) := \int_0^{1/2} x^{a-1}(1-x)^a dx,$$

we clearly obtain

$$\begin{aligned} J(a) + J_1(a) &= \int_0^{1/2} x^{a-1}(1-x)^{a-1}dx = \frac{1}{2}\int_0^1 x^{a-1}(1-x)^{a-1}dx \\ &= \frac{B(a,a)}{2}, \end{aligned}$$

where $B(a,b)$ is the Euler beta function $\int_0^1 x^{a-1}(1-x)^{b-1}dx$. On the other hand,

$$J_1(a) - J(a) = \int_0^{1/2} x^{a-1}(1-x)^{a-1}(1-2x)dx = \frac{(x-x^2)^a}{a}\Big|_0^{1/2} = \frac{1}{a2^{2a}}.$$

Consequently,

$$2J(a) = \frac{B(a,a)}{2} - \frac{1}{a2^{2a}} = \frac{\Gamma(a)^2}{2\Gamma(2a)} - \frac{1}{a2^{2a}}.$$

Inserting $a = \alpha/2$, we thus obtain

$$J(\alpha/2) = \frac{\Gamma(\alpha/2)^2}{4\Gamma(\alpha)} - \frac{1}{\alpha 2^{\alpha}}.$$

Hence, by (2),

$$2^{-\alpha} + \alpha J(\alpha/2) = \frac{\alpha\Gamma(\alpha/2)^2}{4\Gamma(\alpha)} = \frac{1}{K(\alpha)}.$$

This proves (23), and so completes the proof (20). The proof of the inequality (1) is now completed.

To show that the constant $K(\alpha)$ in (1) is best possible we consider the set

$$A := \{\lceil (n/\omega)^{2/\alpha} \rceil \; : \; n = n_0+1, n_0+2, n_0+3, \dots\},$$

where $n_0 = n_0(\omega,\alpha) \in \mathbb{N} \cup \{0\}$ is so large that the sequence A defined above is increasing. Here, $\omega \in (0,\infty)$ for $0<\alpha<2$ and $0<\omega\le 1$ for $\alpha=2$.

It is evident that $U(A,N) \sim \omega N^{\alpha/2}$ as $N \to \infty$, and hence $\omega_\alpha(A) = \omega$. Furthermore, one can easily see that, as $N \to \infty$, the number of pairs $(i,j) \in \mathbb{N}^2$ satisfying $i,j \ge n_0+1$ and

$$\lceil (i/\omega)^{2/\alpha} \rceil + \lceil (j/\omega)^{2/\alpha} \rceil \le N$$

(that is, $V(A,N)$) is approximately

$$\begin{aligned}\sum_{i=1}^{\lfloor \omega N^{\alpha/2} \rfloor} \bigl(\omega^{2/\alpha}N - i^{2/\alpha}\bigr)^{\alpha/2} &= \omega N^{\alpha/2} \sum_{i=1}^{\lfloor \omega N^{\alpha/2} \rfloor} \bigl(1 - N^{-1}(i/\omega)^{2/\alpha}\bigr)^{\alpha/2} \\ &= \omega^2 N^\alpha \int_0^1 (1-x^{2/\alpha})^{\alpha/2} dx + O(N^{\alpha/2}).\end{aligned}$$

(In particular, the case $\alpha=\omega=1$ corresponds to the famous Gauss circle problem. A better error term $O(N^{0.315})$ follows from a result of Huxley [**6**].) Consequently, setting

$$I(\alpha) := \int_0^1 (1-x^{2/\alpha})^{\alpha/2} dx,$$

we find that $V(A,N) \sim \omega^2 N^\alpha I(\alpha)$ as $N \to \infty$, and therefore $\sigma_\alpha(A) = \omega^2 I(\alpha)$.

To see that this completes the proof of the optimality of the constant $K(\alpha)$ in (1) we need to evaluate the integral $I(\alpha)$. Indeed, by the well known identities for the Euler beta function, gamma function and (2), one has

$$\begin{aligned} I(\alpha) &= \int_0^1 (1-x^{2/\alpha})^{\alpha/2} dx = \frac{\alpha}{2}\int_0^1 (1-x)^{\alpha/2} x^{\alpha/2-1} dx \\ &= \frac{\alpha}{2} B(\alpha/2, \alpha/2+1) = \frac{\alpha\Gamma(\alpha/2)\Gamma(\alpha/2+1)}{2\Gamma(\alpha+1)} \\ &= \frac{\alpha\Gamma(\alpha/2)(\alpha/2)\Gamma(\alpha/2)}{2\alpha\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha/2)^2}{4\Gamma(\alpha)} = \frac{1}{K(\alpha)}. \end{aligned}$$

Hence, $\sigma_\alpha(A) = \omega^2 I(\alpha) = \omega^2/K(\alpha)$.

3. PROOF OF COROLLARY 2

Note that each element in $(A+A)\cap[1,N]$ occurs at least twice, except perhaps for at most $U(A,N/2)$ elements of the form $2a_i$, where $a_i \in A$. Consequently,

$$2U(A+A,N) \le V(A,N) + U(A,N/2) \le V(A,N) + U(A,N).$$

Without loss of generality, suppose that $\omega_\alpha(A) < \infty$ (otherwise the claim is trivial). Then, $U(A,N)N^{-\alpha} \to 0$ as $N \to \infty$. Hence, dividing by N^α and letting $N \to \infty$, by Theorem 1, we obtain

$$\begin{aligned} 2\omega_{2\alpha}(A+A) &= \limsup_{N\to\infty} \frac{2U(A+A,N)}{N^\alpha} \\ &\le \limsup_{N\to\infty} \frac{V(A,N)}{N^\alpha} + \limsup_{N\to\infty} \frac{U(A,N)}{N^\alpha} \\ &= \sigma_\alpha(A) + 0 \le \omega_\alpha(A)^2 \frac{\alpha\Gamma(\alpha/2)^2}{4\Gamma(\alpha)}, \end{aligned}$$

whence the result.

On the other hand, by a result of Ruzsa [**12**], for any positive $a > 4$, there are real numbers $0 < b < a$ and $0 \le \xi \le 1$ such that the set

$$A := \{\lfloor n^a + \xi n^b \rfloor \,:\, n > n_0\} = \{a_1 < a_2 < a_3 < \dots\}$$

is a *Sidon set* for a suitable constant n_0. This means that equality $a_i + a_j = a_t + a_l$ holds if and only if $(t,l) = (i,j)$ or $(t,l) = (j,i)$. Since $\alpha < 1/2$, selecting $a = 2/\alpha > 4$, we clearly have $U(A,N) \sim N^{\alpha/2}$ as $N \to \infty$. Thus, $\omega_\alpha(A) = 1$.

On the other hand, the number of pairs $(i,j) \in \mathbb{N}^2$ satisfying $n_0 + 1 \le i \le j$ and

$$\lfloor i^{2/\alpha} + \xi i^b \rfloor + \lfloor j^{2/\alpha} + \xi j^b \rfloor \le N$$

is $N^\alpha I(\alpha)/2$ plus the error term $o(N^\alpha)$. (Compared to the previous section we only have half of the integral $I(\alpha) = \int_0^1 (1 - x^{2/\alpha})^{\alpha/2} dx$ in view of the restriction $i \le j$.) Hence, using $I(\alpha) = 1/K(\alpha)$, where $K(\alpha)$ is defined in (2), we deduce that $U(A+A,N) \sim N^\alpha/(2K(\alpha))$ as $N \to \infty$. It follows that $\omega_{2\alpha}(A+A) = 1/(2K(\alpha))$. Therefore, in view of $\omega_\alpha(A) = 1$, the constant $2K(\alpha)$ in (3) cannot be improved for $\alpha \in (0, 1/2)$.

4. PROOF OF THEOREM 3

Write

$$\sqrt{d} = \sum_{i=0}^{\infty} d_i 2^{q-i}, \tag{24}$$

where $d_0 = 1$, $d_i \in \{0,1\}$ for $i \in \mathbb{N}$ and $q := \lfloor \log d/(2\log 2) \rfloor$. Squaring this expansion, we deduce that

$$d = \sum_{i=0}^{\infty} d_i 2^{q-i} \cdot \sum_{j=0}^{\infty} d_j 2^{q-j} = 2^{2q} \sum_{m=0}^{\infty} \frac{r(m)}{2^m},$$

where

$$r(m) := \sum_{i+j=m} d_i d_j. \tag{25}$$

Below, we shall also use the quantity

$$K := \lceil \log_2(4N+8) \rceil. \tag{26}$$

Let us investigate the sums

$$T(R) := 2^{2q} \sum_{m=0}^{\infty} \frac{r(m+R)}{2^m}$$

for $R = 0, 1, 2, \dots$. Evidently, $T(0) = d$ and

$$T(R) = 2(T(R-1) - 2^{2q} r(R-1))$$

for $R = 1, 2, \dots$. So, $T(R) \in \mathbb{N}$ for each $R \geq 0$.

In particular, $T(1)$ is divisible by 2. Next, for each $j = 0, 1, \dots, 2q$, going step by step we derive that $T(j+1)$ is divisible by 2^{j+1}. In particular, $2^{2q+1} | T(2q+1)$. Furthermore, for each $R \geq 2q+2$, since $2^{2q+1} | T(R-1)$, the number $T(R)$ is divisible by 2^{2q+2}, unless $r(R-1)$ is odd. The latter happens only when R is odd and $d_{(R-1)/2} = 1$. So, we may assume that with at most $\sqrt{2N}$ exceptions (otherwise, the inequality $D(\sqrt{d}, N) \geq \sqrt{2N}$ follows immediately) for each $R = 2q+2, \dots, N-K$, the number $T(R)$ is divisible by 2^{2q+2}. Therefore, using (26), we get

$$\sum_{R=2q+2}^{N-K} T(R) \geq 2^{2q+2}(N - K - 2q - 1 - \sqrt{2N}) > 2^{2q+2}(N - \sqrt{3N}) \tag{27}$$

for each $N \geq N_1(d)$.

On the other hand, we have

$$2^{-2q} \sum_{R=2q+2}^{N-K} T(R) = \sum_{R=2q+2}^{N-K} \sum_{m=0}^{\infty} \frac{r(m+R)}{2^m} = \sum_{j=2q+2}^{\infty} \kappa_j r(j),$$

where

$$\kappa_j := 1 + \frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^{j-2q-2}} < 2$$

for $j = 2q+2, \dots, N-K$ and

$$\kappa_j := \frac{1}{2^{j-N+K}} + \dots + \frac{1}{2^{j-2q-2}} < \frac{1}{2^{j-N+K-1}}$$

for $j = N-K+1, N-K+2, \dots$. Consequently,

$$\begin{aligned} 2^{-2q} \sum_{R=2q+2}^{N-K} T(R) &< \sum_{j=2q+2}^{N-K} 2r(j) + \sum_{j=N-K+1}^{\infty} \frac{r(j)}{2^{j-N+K-1}} \\ &< \sum_{j=2q+2}^{N-1} 2r(j) + \sum_{j=N}^{\infty} \frac{r(j)}{2^{j-N+K-1}}. \end{aligned}$$

Using the trivial bound $r(j) \le j+1$ for $j \ge N$ and (26), we deduce that

$$\sum_{j=N}^{\infty} \frac{r(j)}{2^{j-N+K-1}} \le \sum_{j=N}^{\infty} \frac{j+1}{2^{j-N+K-1}} = \frac{1}{2^K} \sum_{j=0}^{\infty} \frac{N+j+1}{2^{j-1}} = \frac{4N+8}{2^K} \le 1.$$

Consequently,

$$\sum_{R=2q+2}^{N-K} T(R) < 2^{2q+1} \sum_{R=2q+2}^{N-1} r(R) + 2^{2q}.$$

Combining this inequality with (27) we obtain

$$\sum_{R=0}^{N-1} r(R) > \sum_{R=2q+2}^{N-1} r(R) > 2(N-\sqrt{3N}) - 1/2 \tag{28}$$

for $N \ge N_1(d)$.

Consider the set A consisting of positive integers

$$\{a_1 < a_2 < a_3 < \dots\} = \{i+1 \; : \; i \ge 0, d_i = 1\},$$

where d_i are defined in (24). Note that the first digit that is equal to 1 is d_0, so that $a_1 = 1$. By the definition of $r(m)$ in (25), the sum $\sum_{R=0}^{N-1} r(R)$ is the number of integer pairs (i,j) satisfying $0 \le i, j \le N-1$, $i+j \le N-1$ and $d_i = d_j = 1$, or, equivalently, the number of positive integer pairs (i,j) with $a_i + a_j \le N+1$, that is, $V(A, N+1)$. The number of digits among $d_0, \dots, d_{N-1}$ that are equal to 1 is

$$D(\sqrt{d}, N) = U(A, N). \tag{29}$$

By (28), we thus obtain

$$V(A, N+1) > 2(N-\sqrt{3N}) - 1/2. \tag{30}$$

Now, the trivial inequality $U(A,N)^2 \ge V(A, N+1)$ combined with (29) and (30) yields

$$D(\sqrt{d}, N) \ge \sqrt{2(N-\sqrt{3N}) - 1/2} > \sqrt{2N} - 2$$

for $N \ge N_0(d)$, which is (5).

Furthermore, (30) implies that $\sigma_1(A) \ge 2$. Thus, from (4) and(29), it follows that

$$\limsup_{N\to\infty} \frac{D(\sqrt{d}, N)}{\sqrt{N}} = \limsup_{N\to\infty} \frac{U(A,N)}{\sqrt{N}} = \omega_1(A) \ge \sqrt{\frac{4}{\pi}\sigma_1(A)} \ge \sqrt{\frac{8}{\pi}}.$$

This completes the proof of (6).

Acknowledgments. I thank both referees for very careful reading and pointing out some misprints and inaccuracies left in the first version. This research was funded by the European Social Fund according to the activity 'Improvement of researchers' qualification by implementing world-class R&D projects' of Measure No. 09.3.3-LMT-K-712-01-0037.

REFERENCES

1. B. Adamczewski and C. Faverjon: *Chiffres non nuls dans le développement en base entière des nombres algébriques irrationnels.* C. R. Math. Acad. Sci. Paris, **350** (2012), 1–4.
2. D. H. Bailey, J. M. Borwein, C. S. Calude, M. J. Dinneen, M. Dumitrescu and A. Yee: *An empirical approach to the normality of* π*.* Exp. Math., **21** (2012), 375–384.
3. D. H. Bailey, J. M. Borwein, R. E. Crandall and C. Pomerance: *On the binary expansions of algebraic numbers.* J. Théor. Nombres Bordeaux, **16** (2004), 487–518.
4. É. Borel: *Sur les chiffres décimaux de* $\sqrt{2}$ *et divers problèmes de probabilités en chaîne.* C. R. Acad. Sci. Paris, **230** (1950), 591–593.
5. Y. Bugeaud: *Distribution modulo one and Diophantine approximation.* Cambridge Tracts in Mathematics, 193, CUP, Cambridge, 2012.
6. M. N. Huxley: *Exponential sums and lattice points. III.* Proc. London Math. Soc. (3), **87** (2003), 591–609.
7. H. Kaneko: *On the binary digits of algebraic numbers.* J. Aust. Math. Soc., **89** (2010), 233–244.
8. H. Kaneko: *On the beta-expansions of* 1 *and algebraic numbers for a Salem number beta.* Ergodic Theory Dyn. Syst., **35** (2015), 1243–1262.
9. H. Kaneko: *On the number of nonzero digits in the beta-expansions of algebraic numbers.* Rend. Semin. Mat. Univ. Padova, **136** (2016), 205–223.
10. J. S. B. Nielsen and J. G. Simonsen: *An experimental investigation of the normality of irrational algebraic numbers.* Math. Comp., **82** (2013), 1837–1858.
11. T. Rivoal: *On the bits counting function of real numbers.* J. Aust. Math. Soc., **85** (2008), 95–111.
12. I. Z. Ruzsa: *An almost polynomial Sidon sequence.* Studia Sci. Math. Hungar., **38** (2001), 367–375.
13. J. Vandehey: *On the binary digits of* $\sqrt{2}$*.* Integers, **18** (2018), paper #A30.

Artūras Dubickas
Institute of Mathematics,
Faculty of Mathematics and Informatics,
Vilnius University, Naugarduko 24,
LT-03225 Vilnius, Lithuania
E-mail: *arturas.dubickas@mif.vu.lt*

(Received 20.07.2018)
(Revised 03.05.2019)