On digital blocks of polynomial values and extractions in the Rudin–Shapiro sequence

Thomas Stoll (Université de Lorraine / CNRS, Institut Elie Cartan de Lorraine)

February 24, 2016

Abstract

Let $P(x) \in \mathbb{Z}[x]$ be an integer-valued polynomial taking only positive values and let d be a fixed positive integer. The aim of this short note is to show, by elementary means, that for any sufficiently large integer $N \geq N_0(P,d)$ there exists n such that $P(n)$ contains *exactly* N occurrences of the block $(q-1, q-1, \ldots, q-1)$ of size d in its digital expansion in base q. The method of proof allows to give a lower estimate on the number of "0" resp. "1" symbols in polynomial extractions in the Rudin–Shapiro sequence.

1 Introduction

Any introductory course on automatic sequences starts in one way or another with the example of the Thue–Morse sequence (sequence A010060 in the OEIS [9]; cf. [1, Chapter 1.6]), i.e.,

$$(t_n)_{n\geq 0} = 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, \ldots$$

The maybe second best known example of an automatic sequence is the Rudin–Shapiro sequence (sometimes also known as the Golay–Rudin–Shapiro sequence; see [10, 11]). Similarly to the Thue–Morse sequence, the Rudin–Shapiro sequence can be defined in various equivalent ways [1, Example 3.3.1]. The most common one (for combinatorialists on words) is via the morphism

$$a \to ab, \quad b \to ac, \quad c \to db, \quad d \to dc$$

and the mapping

$$a \to 0, \quad b \to 0, \quad c \to 1, \quad d \to 1,$$

see [6, p.252]. For the aim of this note, we will make use of the numbertheoretic definition of the sequence: Denote by R_n the number of (possibly overlapping) occurrences of the block "11" in the base two expansion of n. For example, $R_{59} = 3$ since $59 = (111011)_2$ written in base two. Let $r_n = R_n \bmod 2$, so that $r_{59} = 1$. Then the sequence

$$(r_n)_{n\geq 0} = 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, \ldots$$

is the Rudin–Shapiro sequence (A020987 in the OEIS; cf. [1, Chapter 3.3]). The overall distribution of the two symbols in the sequence $(r_n)_{n\geq 0}$ is well understood. Brillhart and Morton [2] calculated explicit (sharp) constants c_1, c_2 such that

$$\frac{N}{2} - c_1\sqrt{N} < \sum_{n<N} r_n < \frac{N}{2} - c_2\sqrt{N}, \qquad N \geq 1. \tag{1}$$

On the one hand, this result shows that the symbols 0 and 1 are well-distributed within r_n, and, on the other hand, that there is still a weak preponderance of the 0's over 1's. For the Thue–Morse sequence, one easily verifies that

$$\frac{N}{2} - \frac{1}{2} \leq \sum_{n<N} t_n \leq \frac{N}{2} + \frac{1}{2}, \qquad N \geq 1. \tag{2}$$

Various results on weak preponderance are known under the term "Newman's phenomenon" (see, for example, [4]). Newman [8] originally showed in 1969 that there are more 0's than 1's if the restriction of the summation is to multiples of three in (2).

The rarefication of automatic sequences has its early roots in work of Gelfond [5] from 1967/68. He considered the distribution of the sum-of-digits function evaluated on arithmetic progressions. In particular, his work implies that the symbols 0 and 1 in the Thue–Morse sequence are equidistributed when the restriction is to arithmetic progressions. More difficult rarefications, such as primes and squares, have been considered in recent years, and put in the context of Sarnak's "Möbius randomness principle" and related "prime number theorems". We refer to the work of Mauduit and Rivat [7] on prime numbers in Rudin–Shapiro sequences and to the references given therein. The underlying problem shows that the growth rate of the subsequence is crucial. In that sense, primes and squares have still a "quite large" relative density in the integers whereas subsequences of larger growth (polynomials of large degree, for example) remain still out-of-reach of the current methods. There is no particular reason to believe that the behaviour concerning the distribution along such subsequences should be different than the overall behaviour (concerning the principal leading term $N/2$ in (1) and (2)), but it remains, for example, still a difficult open problem to determine (asymptotically) the number of 1's in the extraction of cubes in the Thue–Morse sequence.

In the sequel, let $P(x) \in \mathbb{Z}[x]$ denote an integer-valued polynomial that takes only positive values. A lower bound for the Thue–Morse sequence on general polynomial extractions is due to the author [12]. He proved that

$$\sum_{n<N} t_{P(n)} \gg_P \quad N^{4/(3\deg P+1)}, \qquad N \to \infty, \tag{3}$$

where the implied constant depends on P. This result improved on the lower bound obtaind by Dartyge and Tenenbaum [3] who had

$$\sum_{n<N} t_{P(n)} \gg_P \quad N^{2/(\deg P)!}, \qquad N \to \infty.$$

In the present note we show (with a suitably modified application of the method used in [12]) that for each sufficiently large integer N we can find an integer n such that the number of digital blocks of length d (overlapping or non-overlapping) of the form $(q-1,\dots,q-1)$, i.e., blocks consisting of digits $q-1$ repeated d times, in $P(n)$ is *exactly* N. The method of proof allows to give a lower estimate similar to (3) for polynomial extractions within the Rudin–Shapiro sequence.

2 Notation and Main Result

Let $q \geq 2$ be an integer. For $n \in \mathbb{N}$ we write

$$\sum_{i\geq 0} \varepsilon_i(n) q^i, \qquad \varepsilon_i(n) \in \{0,1,\dots,q-1\}$$

for its digital expansion in base q. For fixed q we denote by $e_d(n)$ the number of occurrences of the block $(q-1,q-1,\dots,q-1)$ of length $d \geq 1$ (possibly overlapping) in the base q representation of n, by $U(n)$ the number of leading digits $(q-1)$ in the expansion of n and by $L(n)$ the number of trailing digits $(q-1)$ in the representation of n. For instance, for $q=10$ and $n = 9184399992399$ we have $e_2(n) = 4$, $U(n) = 1$ and $L(n) = 2$.

Our main result is as follows:

Theorem 1. *There is $N_0(q,P,d) > 1$ such that for all $N \geq N_0(q,P,d)$ there is an n with $e_d(P(n)) = N$.*

We also get a quantitative result if we look at arithmetic progressions.

Theorem 2. *Let $m \geq 2$. There exist $C = C(q,P,d,m) > 0$ and $N_0 = N_0(q,P,d,m) \geq 1$ such that for all $a \in \mathbb{Z}$ and all $N \geq N_0$,*

$$\#\{0 \leq n < N : \quad e_d(P(n)) \equiv a \bmod m\} \geq C N^{4/(3\deg P+1)}.$$

The constant C in Theorem 2 is effectively computable. However, we refrain here from calculating this constant, and refer the interested reader to [12] where such a calculation is presented. A statement about the polynomially rarefied Rudin–Shapiro sequence follows by taking $q = d = m = 2$.

Corollary 1. *We have*

$$\sum_{n<N} r_{P(n)} \gg_P N^{4/(3\deg P+1)}, \qquad N \to \infty.$$

3 Proofs

Our results are based on a crucial lemma about polynomials with a certain sign structure in their l-th power [12]. For the sake of completeness, we restate the proof here.

Lemma 1. *For $m_0, m_1, m_2, m_3 \in \mathbb{R}^+$ and $\ell \geq 1$ denote*

$$t(x) = m_3x^3 + m_2x^2 - m_1x + m_0, \qquad T_\ell(x) = t(x)^\ell = \sum_{i=0}^{3\ell} c_i x^i, \tag{4}$$

with $c_i = c_i(m_3, m_2, m_1, m_0, \ell)$. If

$$1 \leq m_0, m_2, m_3 < q, \qquad 0 < m_1 < \ell^{-1}(6q)^{-\ell},$$

then $c_i > 0$ for $i = 0, 2, 3, \ldots, 3\ell$ and $c_i < 0$ for $i = 1$. Moreover, for all i,

$$|c_i| \leq (4q)^\ell. \tag{5}$$

Proof. The bound (5) follows from easy considerations. For the first statement, observe that $c_0 = m_0^\ell > 0$ and $c_1 = -\ell m_1 m_0^{\ell-1}$ which is negative. Assume now that $2 \leq i \leq 3\ell$ and consider the coefficient of x^i in

$$T_\ell(x) = (m_3x^3 + m_2x^2 + m_0)^\ell + r(x), \tag{6}$$

where

$$r(x) = \sum_{j=1}^{\ell} \binom{\ell}{j} (-m_1x)^j \left(m_3x^3 + m_2x^2 + m_0\right)^{\ell-j} = \sum_{j=1}^{3\ell-2} d_j x^j.$$

First, consider the first summand in (6). Since $m_0, m_2, m_3 \geq 1$ the coefficient of x^i in the expansion of $(m_3x^3 + m_2x^2 + m_0)^\ell$ is ≥ 1. Note also that all the powers $x^2, x^3, \ldots, x^{3\ell}$ appear in the expansion of this term due to the fact that every $i \geq 2$ allows at least one representation as $i = 3i_1 + 2i_2$ with non-negative integers i_1, i_2. We prove that for sufficiently small $m_1 > 0$ the coefficient of x^i in the first summand in (6) is dominant. Suppose that $m_1 < 1$ so that $m_1 > m_1^j$ for $2 \leq j \leq \ell$. Then

$$|d_j| < \ell 2^\ell m_1 (3q)^\ell = \ell\,(6q)^\ell\, m_1, \qquad 1 \leq j \leq 3\ell - 2.$$

Therefore, if $m_1 < \ell^{-1}(6q)^{-\ell}$ then all of $x^2, \ldots, x^{3\ell}$ in the polynomial $T_\ell(x)$ have positive coefficients. □

In [12] we considered the sum-of-digits function in base q which is a strictly q-additive function. Counting blocks, as we do here, is certainly not a q-additive process but we are not far off as seen in the following proposition.

Proposition 1. *Let $1 \leq q^{u-1} \leq b < q^u \leq q^k$ and $a, k \geq 1$.*

(i) *If $b < q^{k-1}$ then*

$$e_d(aq^k + b) = e_d(a) + e_d(b).$$

(ii) *If $k-u \geq d$ then*

$$e_d(aq^k-b) = k-u-d+1+e_d(a-1)+e_d(q^u-b) \\ +\min(d-1, L(a-1))+\min(d-1, U(q^u-b)).$$

Proof. The inequality $b < q^{k-1}$ guarantees that $\varepsilon_{k-1}(aq^k+b)=0$, so that there are no blocks $(q-1,\dots,q-1)$ that span over the a and b parts, and (i) follows. For (ii), we first write

$$\begin{aligned} e_d(aq^k-b) &= e_d((a-1)q^k+q^k-q^u+q^u-b) \\ &= e_d\left((a-1)q^k+\sum_{i=0}^{k-u-1}(q-1)q^{i+u}+q^u-b\right) \\ &= e_d(a-1)+e_d\left(\sum_{i=0}^{k-u-1}(q-1)q^i\right)+e_d\left(q^u-b\right)+\sum_{i=1}^{\min(d-1,L(a-1))}1+\sum_{i=1}^{\min(d-1,U(q^u-b))}1. \end{aligned}$$

Moreover, we have

$$e_d\left(\sum_{i=0}^{k-u-1}(q-1)q^i\right)=\sum_{i=0}^{(k-u-1)-d+1}1=k-u-d+1$$

and (ii) follows. □

Proofs of Theorems 1 and 2. We start the proofs of Theorems 1 and 2 with the easier case of monomials,

$$P(x)=x^h, \qquad h\geq 1,$$

and generalize in a second step to general polynomials $P(x)\in\mathbb{Z}[x]$. We regard d and h as fixed quantities. Lemma 1 shows that for all integers m_0,m_1,m_2,m_3 with

$$q^{v-1}\leq m_0,m_2,m_3<q^v, \qquad 1\leq m_1<q^v/(hq(6q)^h), \tag{7}$$

the polynomial $T_h(x)=(t(x))^h=P(t(x))$ has all positive integer coefficients with the only exception of the coefficient of x^1 which is negative. Let v be an integer such that

$$q^v\geq 2hq(6q)^h \tag{8}$$

and let $k\in\mathbb{Z}$ be such that

$$k>hv+2h+1. \tag{9}$$

With these inequalities at hand, the interval for m_1 in (7) is non-empty and

$$q^{k-1}>q^{hv}\cdot q^{2h}\geq(4q^v)^h\geq|c_i|, \qquad \text{for all } i=0,1,\dots,3h,$$

where c_i is the coefficient of x^i in $T_h(x)$ which only depends on m_0,m_1,m_2 and m_3. We now use twice Proposition 1 (i) to get

$$e_d(t(q^k)^h)=e_d\left(\sum_{i=2}^{3h}c_iq^{ik}-|c_1|q^k+c_0\right)=\sum_{i=3}^{3h}e_d(c_i)+e_d(c_2q^k-|c_1|)+e_d(c_0).$$

Let u be such that

$$q^{u-1}\leq|c_1|<q^u. \tag{10}$$

Since

$$|c_1|=hm_1m_0^{h-1} \tag{11}$$

we see that u only depends on m_0, m_1. Suppose that, in addition to (9) we also have

$$k \geq d + u. \tag{12}$$

Then by Proposition 1 (ii) we get

$$e_d(t(q^k)^h) = \sum_{i=3}^{3h} e_d(c_i) + e_d(c_0) + k - u - d + 1 + e_d(c_2 - 1) + e_d(q^u - |c_1|) \\ + \min(d-1, L(c_2-1)) + \min(d-1, U(q^u - |c_1|))$$

which means that

$$e_d(t(q^k)^h) = k + M$$

with $M = M(m_0, m_1, m_2, m_3)$. Once we fix m_0, m_1, m_2 and m_3 (with fixed d and h) in the ranges (7), the quantity M does not depend on k and is constant whenever k satisfies (9) and (12), say, $k \geq k_0$. Since $hv + 2h + 1 < hv + 2h + d + 1$, and by (10), (11) and (7),

$$q^{u-1} < h\,\frac{q^v}{hq(6q)^h} \cdot q^{v(h-1)} < q^{vh-1},$$

we see that we can take

$$k_0 = hv + 2h + d + 1. \tag{13}$$

Summing up, given h, q and d we choose v and integers m_0, m_1, m_2, m_3 according to (8) and (7), set k_0 as given in (13), and find that for all $N = M + k \geq M + k_0 =: N_0$ we have $e_d(n^h) = N$ with $n = t(q^k)$. This already proves Theorem 1 for the case of monomials $P(x) = x^h$.

Now, since

$$e_d(t(q^k)^h), \qquad \text{for } k = k_0,\ k_0 + 1,\ \ldots,\ k_0 + m - 1, \tag{14}$$

runs through a complete set of residues mod m, we hit a fixed arithmetic progression mod m for some k with $k_0 \leq k \leq k_0 + m - 1$. Therefore, by (7) we find at least

$$(q^v - q^{v-1})^3 (q^v/(hq(6q)^h) - 1) \gg_{q,h} q^{4v} \tag{15}$$

integers n that by (9), (12) and (14) are all smaller than

$$q^v \cdot q^{3(hv+2h+d+m)} = q^{3(2h+d+m)} \cdot q^{v(3h+1)}$$

and satisfy $e_d(n^h) \equiv a \bmod m$ for fixed a and m. Note that by our construction all these integers are distinct. We denote

$$N_0 = N_0(q, h, d, m) = q^{3(2h+d+m)} \cdot q^{v_0(3h+1)},$$

where

$$v_0 = \left\lceil \log_q \left(2hq(6q)^h\right) \right\rceil = O_{q,h}(1).$$

Then for all $N \geq N_0$ we find $v \geq v_0$ with

$$q^{3(2h+d+m)} \cdot q^{v(3h+1)} \leq N < q^{3(2h+d+m)} \cdot q^{(v+1)(3h+1)}. \tag{16}$$

By (15) and (16), we finally find at least $c' N^{4/(3h+1)}$ integers n with $0 \leq n < N$ and $e_d(n^h) \equiv a \bmod m$, where the constant c' depends at most on q, h, d and m. We therefore get the statement of Theorem 2 for the case of monomials $P(x) = x^h$ with $h \geq 1$.

Finally, let $P(x) = a_h x^h + \cdots + a_0 \in \mathbb{Z}[x]$. Without loss of generality we may assume that all a_i are positive, since otherwise there exists $\delta = \delta(P)$ depending only on P such that $P(x + \delta)$ has all positive

coefficients. By Lemma 1 we see that the polynomial $P(t(x))$ has all positive coefficients with the exception of a negative coefficient to the power x^1. Choosing k sufficiently large, e.g.,

$$k > hv + 2h + d + \log_q \left(\max_{0 \le i \le h} a_i \right),$$

we can again split the digital structure of $P(t(q^k))$ exactly as above and can apply the same reasoning to obtain the general statements of Theorems 1 and 2.

□

4 Concluding remarks

We conclude this note with two open questions that naturally arise in this context:

(i) Is it possible to extend Theorem 1 to *arbitrary* blocks of size d?

(ii) Can we sharpen the result in Corollary 1 for $P(n) = n^2$? Numerical simulations suggest that there exists $c_3 > 0$ and $0 < \alpha < 1$ such that for $N \ge N_0$,

$$\sum_{n<N} r_{n^2} < \frac{N}{2} - c_3 N^{\alpha}.$$

This is supported by the inequalities (1).

References

[1] J.-P. Allouche, J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge University Press, Cambrdige, 2003.

[2] J. Brillhart, P. Morton, A case study in mathematical research: the Golay–Rudin–Shapiro sequence, *Amer. Math. Monthly* **103** (1996), 854–869.

[3] C. Dartyge, G. Tenenbaum, Congruences de sommes de chiffres de valeurs polynomiales, *Bull. London Math. Soc.* **38** (2006), no. 1, 61–69.

[4] M. Drmota, T. Stoll, Newman's phenomenon for generalized Thue–Morse sequences, *Discrete Math.* **308** (7) (2008), 1191–1208.

[5] A. O. Gelfond, Sur les nombres qui ont des propriétés additives et multiplicatives données, *Acta Arith.* **13** (1967/1968), 259–265.

[6] M. Lothaire, Applied combinatorics on words, Encyclopedia of Mathematics and its Applications **105**, *Cambridge Univ. Press*, Cambridge, 2005.

[7] C. Mauduit, J. Rivat, Prime numbers along Rudin–Shapiro sequences, *J. Eur. Math. Soc.*, to appear.

[8] D. J. Newman, On the number of binary digits in a multiple of three, *Proc. Amer. Math. Soc.* **21** (1969), 719–721.

[9] The Online Encyclopedia of Integer Sequences (OEIS), N.J.A. Sloane, `https://oeis.org/`.

[10] W. Rudin, Some theorems on Fourier coefficients, *Proc. Amer. Math. Soc.* **10** (1959), 855–859.

[11] H. S. Shapiro, Extremal Problems for Polynomials and Power Series, MS thesis, M.I.T., 1951.

[12] T. Stoll, The sum of digits of polynomial values in arithmetic progressions, *Functiones et Approximatio* **47** (2) (2012), 233–239.