ON POLYNOMIAL EXTRACTIONS OF THE RUDIN–SHAPIRO SEQUENCE

THOMAS STOLL

Abstract. Let $P(x) \in \mathbb{Z}[x]$ be an integer-valued polynomial taking only positive values and let d be any 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 $(q-1, q-1, \dots, q-1)$ in its digital expansion in base q. The method of proof is constructive. It allows to give a lower estimate on the number of "0" resp. "1" symbols in polynomial extractions of 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 [5]), i.e.,

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

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 [6, 7]). Similarly to the Thue–Morse sequence, the Rudin–Shapiro sequence can be defined in various equivalent ways. The most common one (for combinatorialists on words) is via the substitution $a \mapsto ab$, $b \mapsto ac$, $c \mapsto db$, $d \mapsto dc$ and the mapping $a \mapsto 0$, $b \mapsto 0$, $c \mapsto 1$, $d \mapsto 1$. 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

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

is the Rudin–Shapiro sequence (see also [1]; A020987 in the OEIS). The overall distribution of the two symbols in the sequence (r_n) is well understood. Brillhart and Morton [2] calculated explicit (sharp) constants c_1, c_2 such that

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

This means that there is a weak preponderance of the symbol 0 over symbol 1 in the Rudin–Shapiro sequence. For the Thue–Morse sequence, one easily verifies that

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

2010 *Mathematics Subject Classification.* Primary 11A63; Secondary 11B85.

Key words and phrases. Rudin–Shapiro sequence; automatic sequences; polynomials.

La Région Lorraine "Operation realisée avec le concours financier du Conseil Régional de Lorraine."

The rarefication of automatic sequences has its early roots in work of Gelfond [3] 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 [4] and 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, 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, i.e., as $N \to \infty$,

$$\{n < N : \quad t_{n^3} = 1\} \sim \frac{N}{2} \quad ?$$

In the sequel, let $P[x] \in \mathbb{Z}[x]$ denote an integer-valued polynomial that takes only positive values. The best known lower bound for the Thue–Morse sequence is due to the author [8]. He proved that

$$\{n < N : \quad t_{P(n)} = 1\} \gg N^{4/(3\deg P+1)}. \tag{2}$$

In the present note we show (with an application of the same method) that the symbols 0 and 1 appear infinitely many often in the extraction along indices $P(n)$ within the Rudin–Shapiro sequence and give a lower estimate similar to (2). On our way, we prove 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 $(q-1,\ldots,q-1)$, i.e., blocks consisting of digits $q-1$ repeated d times, in $P(n)$ is *exactly* N.

2. Notation and Main Result

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

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

for its digital expansion in base q. For fixed q we denote by $e_d(n)$ the number of occurrences of the block $(q-1,q-1,\ldots,q-1)$ 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 $L(n) = 2$.

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 actually get an in some respect stronger 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 $N \geq N_0$,*

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

A statement about the 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

The result is based on a crucial lemma about polynomials with a certain sign structure in their l-th power [8]. For the sake of completeness, we also give the proof here.

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

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

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

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

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

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

Proof. The bound (4) follows from easy considerations. For the first statement, observe that $c_0 = m_0^l > 0$ and $c_1 = -lm_1m_0^{l-1}$ which is negative. Assume now that $2 \leq i \leq 3l$ and consider the coefficient of x^i in

$$T_l(x) = (m_3x^3 + m_2x^2 + m_0)^l + r(x), \tag{5}$$

where

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

First, consider the first summand in (5). Since $m_0, m_2, m_3 \geq 1$ the coefficient of x^i in the expansion of $(m_3x^3 + m_2x^2 + m_0)^l$ is ≥ 1. Note also that all the powers $x^2, x^3, \ldots, x^{3l}$ 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 (5) is dominant. Suppose that $m_1 < 1$ so that $m_1 > m_1^j$ for $2 \leq j \leq l$. Then

$$|d_j| < l2^l m_1 (3q)^l = l\,(6q)^l\, m_1, \qquad 1 \leq j \leq 3l-2.$$

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

Counting blocks, as we do, is certainly not a q-additive process in the strict sense (compared to the case of the sum-of-digits function and the Thue–Morse sequence), but we are not far off as seen in the following proposition.

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

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

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

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

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

Proof. The condition in (i) guarantees that there are no blocks $(q-1, \ldots, q-1)$ that span over the a and b parts. The statement (ii) follows from $e_d(aq^k - b) = e_d((a-1)q^k + q^k - q^u + q^u - b)$ and by considering the various possibilities for the block. □

We start with the easier case of monomials,

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

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

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

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

$$q^v \ge 2hq(6q)^h \tag{7}$$

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

$$k > hv + 2h + 1. \tag{8}$$

With these inequalities at hand, the interval for m_1 in (6) is non-empty and

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

where c_i is the coefficient of x^i in $T_h(x)$. We now use Proposition 1 (i) to get

$$e_d(t(q^k)^h) = e_d\left(\sum_{i=2}^{3h} c_i q^{ik} - |c_1| q^k + c_0\right) = \sum_{i=3}^{3h} e_d(c_i) + e_d(c_2 q^k - |c_1|) + e_d(c_0).$$

Let u be such that $q^{u-1} \le |c_1| < q^u$. Since $|c_1| = hm_1 m_0^{h-1}$ we see that u only depends on m_0, m_1. Suppose that, in addition to (8) we also have

$$k \ge d + u. \tag{9}$$

Then by Proposition 1 (ii) we get

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

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 (6), the quantity M does not depend on k and is constant whenever k satisfies (8) and (9), say, $k \geq k_0$. A simple calculation shows that we may take

$$k_0 = hv + 2h + d + 1. \tag{10}$$

This already proves Theorem 1 for the case of monomials x^h.

Now, since

$$e_d(t(q^k)^h), \qquad \text{for } k = k_0,\ k_0+1,\ \ldots,\ k_0+m-1, \tag{11}$$

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 (6) we find at least

$$(q^v - q^{v-1})^3 (q^v/(hq(6q)^h) - 1) \gg_{q,h} \ q^{4v} \tag{12}$$

integers n that by (8), (9) and (11) 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

$$q^{3(2h+d+m)} \cdot q^{v(3h+1)} \leq N < q^{3(2h+d+m)} \cdot q^{(v+1)(3h+1)}. \tag{13}$$

By (12) and (13), we finally find

$$\gg_{q,h,d,m} \ N^{4/(3h+1)}$$

integers n with $0 \leq n < N$ and $e_d(n^h) \equiv a \bmod m$, thus we also 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 $f = \delta(P)$ depending only on δ 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.,

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

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

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

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

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

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

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

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

1. Université de Lorraine, Institut Elie Cartan de Lorraine, UMR 7502, Vandoeuvre-lès-Nancy, F-54506, France; 2. CNRS, Institut Elie Cartan de Lorraine, UMR 7502, Vandoeuvre-lès-Nancy, F-54506, France

E-mail address: `thomas.stoll@univ-lorraine.fr`