Aperiodicity Measure for Infinite Sequences

Yuri Pritykin[1,2,⋆] and Julya Ulyashkina[1,⋆⋆]

[1] Department of Mechanics and Mathematics,
Moscow State University, Russia
[2] Computer Science Department,
Princeton University, USA

Abstract. We introduce the notion of aperiodicity measure for infinite symbolic sequences. Informally speaking, the aperiodicity measure of a sequence is the maximum number (between 0 and 1) such that this sequence differs from each of its non-identical shifts in at least fraction of symbols being this number. We give lower and upper bounds on the aperiodicity measure of a sequence over a fixed alphabet. We compute the aperiodicity measure for the Thue–Morse sequence and its natural generalization the Prouhet sequences, and also prove the aperiodicity measure of the Sturmian sequences to be 0. Finally, we construct an automatic sequence with the aperiodicity measure arbitrarily close to 1.

1 Introduction

Combinatorics on words is a deeply studied field in theoretical computer science and discrete mathematics. In this paper we focus on infinite words, or sequences, over a finite alphabet. Periodic sequences have the simplest structure, and it is natural to try to measure how far a sequence may be from any periodic sequence.

In this paper we introduce the notion of aperiodicity measure AM for infinite symbolic sequences. Our definition is based on the discrete version of Besicovitch distance that was used by Morse and Hedlund [14] when defining sequences that they called almost periodic[3]. The same approach was also used in [7] when defining α-aperiodic two-dimensional sequences. As it is essentially noticed in [7], if $\mathrm{AM}(x) > \alpha$ for a sequence x, then x has Besicovitch distance at least $\alpha/2$ with every eventually periodic sequence. In [14] it is also proved that $\mathrm{AM}(\mathbf{t}) \geqslant 1/4$ where $\mathbf{t}$ is the Thue–Morse sequence.

Informally speaking, the aperiodicity measure of a sequence is the maximum number (between 0 and 1) such that this sequence differs from each of its non-identical shifts in at least fraction of symbols being this number. Our interest to this notion was mostly inspired by the following conjecture from the personal communication with B. Durand, A. Romashchenko, and A. Shen, that we positively prove as Theorem 6.

⋆ pritykin@cs.princeton.edu
⋆⋆ julya.ulyashkina@gmail.com

[3] This term "almost periodic" from [14] should not be mixed with the recent usage of the term "almost periodic sequence" which stands for sequences also known as uniformly recurrent or minimal, e.g., see [16] for a survey.

Conjecture. For every $\alpha < 1$ there exists an automatic sequence x such that $\mathrm{AM}(x) \geqslant \alpha$.

The solution of this conjecture allows to simplify the construction of a strongly aperiodic tiling from [7].

Besides this conjecture, we believe that the notion of aperiodicity measure is interesting and natural itself, and the main goal of the paper is to support this statement.

Other similar notions and results have appeared in the literature: to name a few, see [14] on Besicovitch-almost-periodic sequences, [8] on tilings of the Thue–Morse sequence, [10] on approximate squares in sequences. The closest to ours seems to be the notion of correlation measure introduced in [12] and then studied in a series of papers currently concluding with [5], see also [13] and many others. Their correlation measure of order 2 is essentially the same as our aperiodicity measure for binary sequences, though in general motivation, frameworks, and approaches are somewhat different. After the submission of our paper, we became aware of the recent paper [9] continuing the investigations of the aforementioned correlation measure. In particular, a result from [9] improves the result of our Theorem 2 below.

The paper is organized as follows. In Section 2 we give necessary preliminaries. In Section 3 we define the aperiodicity measure AM of an infinite sequence and then study some basic properties of this notion. We prove that there exist sequences with AM arbitrarily close to 1 (Theorem 1), though there exists an upper bound strictly less than 1 for AM of sequences over a fixed finite alphabet (Theorem 2). Then we calculate AM for the Thue–Morse sequence (Theorem 3) and for the Sturmian sequences (Theorem 4). In Section 4 we construct an automatic sequence with AM arbitrarily close to 1 (Theorem 6), though first we prove that the Prouhet sequences, natural generalization of the Thue–Morse sequence, do not suffice for this purpose (Theorem 5). Due to space constraints, some proofs are only sketched. Section 5 concludes the paper with a number of open problems.

2 Preliminaries

We use all common definitions and notions of combinatorics on words, which can be found, i.e., in [4] or [11]. We recall some of them here for establishing our notations.

The number of elements in a finite set X is denoted $\#X$. Let $\mathbb{N}$ be the set of natural numbers $\{0, 1, 2, \dots\}$. We use $[i, j]$ for denoting the segment of natural numbers $\{i, i+1, i+2, \dots, j\}$, while the segment $[0, j]$ is simply denoted $[j]$. Let A be a finite alphabet. We consider finite *words* as mappings $u\colon [n-1] \to A$ and denote the length of u by $|u|$, that is, $u = u(0)\,u(1)\,u(2)\,\dots\,u(|u|-2)\,u(|u|-1)$. An empty word is denoted Λ. We also deal with *sequences* over this alphabet, i.e., mappings $x\colon \mathbb{N} \to A$, and denote the set of these sequences by $A^{\mathbb{N}}$. A word of the form $x[0, i]$ for some i is called a *prefix* of x, and respectively a sequence of the form $x(i)x(i+1)x(i+2)\dots$ for some i is called a *suffix* of x. A *left*

shift L maps a sequence to the same sequence with the first symbol cut, that is, $Lx = L(x(0)x(1)x(2)\dots) = x(1)x(2)\dots$.

A sequence x is *periodic* if for some $T > 0$ we have $x(i) = x(i+T)$ for each $i \in \mathbb{N}$. This T is called a period of x. A sequence is *eventually periodic* if some of its suffixes is periodic.

Let A, B be finite alphabets. A mapping $\phi\colon A^* \to B^*$ is called a *morphism* if $\phi(uv) = \phi(u)\phi(v)$ for all $u, v \in A^*$. Obviously, a morphism is determined by its values on single-letter words. A morphism is *k-uniform* if $|\phi(a)| = k$ for each $a \in A$. A 1-uniform morphism is called a *coding*. For $x \in A^{\mathbb{N}}$ denote $\phi(x) = \phi(x(0))\phi(x(1))\phi(x(2))\dots$ Further we consider only morphisms of the form $A^* \to A^*$. Let $\phi(t) = tu$ for some $t \in A$, $u \in A^*$. Then for all natural $m < n$ the word $\phi^n(t)$ begins with the word $\phi^m(t)$, so $\phi^\infty(t) = \lim_{n\to\infty} \phi^n(t) = tu\phi(u)\phi^2(u)\phi^3(u)\dots$ is correctly defined. If $\phi^n(u) \neq \Lambda$ for all n, then $\phi^\infty(t)$ is infinite. In this case ϕ is said to be *prolongable* on t. Sequences of the form $h(\phi^\infty(t))$ for a coding $h\colon A \to B$ are called *morphic*, of the form $\phi^\infty(t)$ are called *pure morphic*.

Unless stated otherwise, usually in this paper we assume $A = \{0, \dots, k-1\}$ for some $k \in \mathbb{N}$ and assume usual operations $+$ and $\cdot$ in A modulo k. We also assume $t = 0 \in A$, that is, we usually iterate a morphism on symbol 0 to obtain a sequence.

The class of morphic sequences of the form $h(\phi^\infty(t))$ with ϕ being k-uniform coincides with the class of so-called *k-automatic* sequences. Sequences that are k-automatic for some k, are called simply *automatic* (this class was introduced in [6] under the name of uniform tag sequences and was widely studied afterwards, see [4]).

The famous *Thue–Morse sequence* $\mathbf{t} = 01101001100101101001011\dots$ is the automatic sequence generated by the morphism $0 \to 01$, $1 \to 10$. This sequence can also be defined using conditions $\mathbf{t}(0) = 0$ and $\mathbf{t}(2n) = \mathbf{t}(n)$ and $\mathbf{t}(2n+1) = 1 - \mathbf{t}(n)$ for every n. The Thue–Morse sequence has several other names (due to other researchers who discovered it independently), a lot of interesting properties and appears in a lot of contexts; for a survey on the Thue–Morse sequence see [2] and also [4, 11].

Another famous class of sequences is that of *Sturmian sequences*. They were introduced in [15] and has been widely studied since that time, e.g., see [11]. Sturmian sequences have many different equivalent definitions, and we will use the following, via (lower) mechanical sequences. For real numbers α and ρ with $0 < \alpha < 1$, α irrational, and $0 \leqslant \rho < 1$, let $c_{\alpha,\rho}(n) = \lfloor \alpha(n+1) + \rho \rfloor - \lfloor \alpha n + \rho \rfloor$ be the Sturmian sequence with parameters α and ρ. Here $\lfloor y \rfloor$ for a real number y is the maximal integer not greater than y. Let $\mathrm{fr}(y) = y - \lfloor y \rfloor$ for a real y to be its fractional part. Note that $c_{\alpha,\rho}(n) = 0$ if $0 \leqslant \mathrm{fr}(\alpha n + \rho) < 1 - \alpha$ and $c_{\alpha,\rho}(n) = 1$ if $1 - \alpha \leqslant \mathrm{fr}(\alpha n + \rho) < 1$. For example, the well-known *Fibonacci sequence* $\mathbf{f} = 01001010010100101001010\dots$ that can be obtained as $\phi^\infty(0)$ for $\phi(0) = 01$, $\phi(1) = 0$, is the Sturmian sequence $\mathbf{f} = c_{1/\gamma^2, 1/\gamma^2}$ where $\gamma = \frac{1+\sqrt{5}}{2}$ is the golden ratio.

3 Aperiodicity Measure

The Besicovitch distance between sequences x and y is defined as $d(x,y) = \liminf_{n\to\infty} \frac{1}{n}\#\{i : i \in [n-1], x(i) \neq y(i)\}$. Then define the *aperiodicity measure* of a sequence x to be $\mathrm{AM}(x) = \inf\{d(x, L^n x) : n \geqslant 1\}$. In other words, $\mathrm{AM}(x)$ is the maximum number between 0 and 1 such that x differs from every non-trivial shift of x in at least $\mathrm{AM}(x)$ fraction of symbols. Let $s_n = \limsup_{m\to\infty} \frac{1}{m}\#\{i : i \in [m-1], x(i) = x(i+n)\}$. Then $\mathrm{AM}(x) = 1 - \sup\{s_n : n \geqslant 1\}$.

Clearly if a sequence x is eventually periodic with a period p then $s_p = 1$ and therefore $\mathrm{AM}(x) = 0$. The opposite is not necessarily true, though if $\mathrm{AM}(x)$ is relatively small then it is reasonable to think of x as close to a periodic sequence. We prove that $\mathrm{AM}(x) = 0$ for the Sturmian sequences x (Theorem 4) that are often considered as close to periodic. Note that if $\mathrm{AM}(x) > \alpha$ for a sequence x, then x has Besicovitch distance at least $\alpha/2$ with every eventually periodic sequence. Indeed, suppose $d(x,y) < \alpha/2$ for an eventually periodic sequence y with a period p. Then $d(L^p x, x) \leqslant d(L^p x, L^p y) + d(L^p y, y) + d(y, x) = d(x,y) + 0 + d(y,x) < \alpha$, since the Besicovitch distance is symmetric and satisfies the triangle inequality.

In Section 4 we prove that there exist automatic sequences with aperiodicity measure arbitrarily close to 1. However, if we do not require a sequence to be automatic, then the sequence with aperiodicity measure arbitrarily close to 1 can easily be proved to exist.

Theorem 1. *For every $\alpha < 1$ there exists a sequence x such that $\mathrm{AM}(x) \geqslant \alpha$.*

Proof. Let x be a sequence over an alphabet with k symbols. It is not difficult to prove that x can be chosen so that for every n the fraction of i's such that $x(i) = x(i+n)$, exists and is equal to $\frac{1}{k}$. Indeed, for every fixed n the Lebesgue measure (should not be mixed with aperiodicity measure!) of sequences not satisfying the above condition, is 0. Therefore, total Lebesgue measure of "bad" sequences is 0.

Let us choose such x. Then $s_n = \frac{1}{k}$ for every n. Therefore $\mathrm{AM}(x) = 1 - \frac{1}{k}$. □

However, for every fixed alphabet the aperiodicity measure can be bounded from above by some number strictly less than 1.

Theorem 2. *If a sequence x has no more than k symbols, then $\mathrm{AM}(x) \leqslant 1 - \frac{1}{2k}$.*

Proof. Suppose the alphabet of the sequence is $A = \{0, \ldots, k-1\}$.

The first observation we can make is that among any $k+1$ consecutive symbols of x there is one that occurs twice, by pigeonhole principle. After proper calculation similar to what we do below, one gets $\mathrm{AM}(x) \leqslant 1 - \frac{1}{k^2}$.

Generalizing this idea, let $u = x[l, l+N]$ be some segment of x, and for $0 \leqslant j \leqslant k-1$ denote by r_j how many times symbol j occurs in u. We have $\sum_{j=0}^{k-1} r_j = N+1$. For a symbol $j \in A$, there are $\frac{r_j(r_j-1)}{2}$ pairs (p,q) with $p, q \in [l, l+N]$, $p < q$, such that $x(p) = x(q) = j$. Therefore $\#\{(p,q) : p, q \in [l, l+N], p < q, x(p) = x(q)\} = \sum_{j=0}^{k-1} \frac{1}{2} r_j(r_j - 1) = \sum_{j=0}^{k-1} \frac{1}{2} r_j^2 - \frac{N+1}{2} \geqslant \frac{1}{2k}(N+1)^2 - \frac{1}{2}(N+1)$, where we used the Cauchy's inequality $\sum_{j=0}^{k-1} r_j^2 \geqslant \frac{1}{k}\left(\sum_{j=0}^{k-1} r_j\right)^2$.

Now let us approximate $\sum_{n=1}^{N} s_n$:

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{1}{m}\#\{i \in [m-1] : x(i) = x(i+n)\} \\
&\quad = \frac{1}{m}\#\{(i,n) : i \in [m-1], n \in [1,N], x(i) = x(i+n)\} \\
&\quad \geqslant \sum_{t=0}^{\lfloor m/N \rfloor - 1} \frac{1}{m}\#\{(i,n) : i \in [Nt, N(t+1)-1], n \in [1,N], x(i) = x(i+n)\} \\
&\quad \geqslant \sum_{t=0}^{\lfloor m/N \rfloor - 1} \frac{1}{m}\#\{(p,q) : p,q \in [Nt, N(t+1)], p < q, x(p) = x(q)\} \\
&\quad \geqslant \frac{\lfloor m/N \rfloor}{m}\left(\frac{(N+1)^2}{2k} - \frac{N+1}{2}\right) \to \frac{(N+1)^2}{2kN} - \frac{N+1}{2N}
\end{aligned}$$

as $m \to \infty$. Thus $\sum_{n=1}^{N} s_n \geqslant \frac{(N+1)^2}{2kN} - \frac{N+1}{2N}$, and therefore $s_n \geqslant \frac{(N+1)^2}{2kN^2} - \frac{N+1}{2N^2}$ for some n such that $1 \leqslant n \leqslant N$. Tending $N \to \infty$, we can find s_n arbitrarily close to $\frac{1}{2k}$, therefore $\mathrm{AM}(x) \leqslant 1 - \frac{1}{2k}$. □

Note that in [9] the upper bound $1 - \frac{1}{k}$ was obtained for aperiodicity measure of sequences over k-letter alphabet which matches the lower bound from Theorem 1.

Now we compute the aperiodicity measure for some well known sequences.

Theorem 3. $\mathrm{AM}(\mathbf{t}) = \frac{1}{3}$.

Proof. Let $s_n^m = \frac{1}{m}\#\{i \in [m-1] : \mathbf{t}(i) = \mathbf{t}(i+n)\}$.

First of all, clearly $s_0^m = 1$ for every m. Then one can obtain the following equations for s_n^m:

$$\begin{aligned}
s_{2n}^{2m} &= s_n^m, \\
s_{2n}^{2m+1} &= \frac{m+1}{2m+1}s_n^{m+1} + \frac{m}{2m+1}s_n^m, \\
s_{2n+1}^{2m} &= 1 - \frac{1}{2}(s_n^m + s_{n+1}^m), \\
s_{2n+1}^{2m+1} &= \frac{m+1}{2m+1}(1 - s_n^{m+1}) + \frac{m}{2m+1}(1 - s_{n+1}^m)
\end{aligned} \tag{1}$$

for every m and n.

The idea is to consider separately even and odd indices of the sequence. Let us prove for instance the fourth equation. Indeed,

$$\begin{aligned} s_{2n+1}^{2m+1} &= \frac{1}{2m+1}\#\{i \in [2m] : \mathbf{t}(i) = \mathbf{t}(i+2n+1)\} \\ &= \frac{1}{2m+1}\#\{i \in [m] : \mathbf{t}(2i) = \mathbf{t}(2i+2n+1)\} + \\ &\quad \frac{1}{2m+1}\#\{i \in [m-1] : \mathbf{t}(2i+1) = \mathbf{t}(2i+1+2n+1)\} \\ &= \frac{1}{2m+1}\#\{i \in [m] : \mathbf{t}(i) \neq \mathbf{t}(i+n)\} + \\ &\quad \frac{1}{2m+1}\#\{i \in [m-1] : \mathbf{t}(i) \neq \mathbf{t}(i+n+1)\} \\ &= \frac{m+1}{2m+1}(1 - s_n^{m+1}) + \frac{m}{2m+1}(1 - s_{n+1}^m), \end{aligned}$$

where we used equations $\mathbf{t}(2i) = \mathbf{t}(i)$ and $\mathbf{t}(2i+1) = 1 - \mathbf{t}(i)$ for the Thue–Morse sequence. Other equations from (1) are proved in a similar way.

From (1) we derive

$$s_1^{2m} = \frac{1}{2} - \frac{1}{2}s_1^m, \qquad s_1^{2m+1} = \frac{m}{2m+1}(1 - s_1^m).$$

Note that $s_1^1 = 0$. Our goal is to prove that $\lim_{m\to\infty} s_1^m = \frac{1}{3}$. Let $s_1^m = \frac{1}{3} + a_m$. Then we have $a_1 = -\frac{1}{3}$, $a_{2m} = -\frac{1}{2}a_m$, and $a_{2m+1} = -\frac{m}{2m+1}a_m - \frac{1}{6m+3}$. Let $b_m = 3ma_m$. Then $b_1 = -1$, $b_{2m} = 6ma_{2m} = -3ma_m = -b_m$ and $b_{2m+1} = 3(2m+1)a_{2m+1} = -3ma_m - 1 = -b_m - 1$, from what it can easily be seen that $|b_m| = O(\log m)$, and therefore $\lim_{m\to\infty} a_m = 0$ and there exists $\lim_{m\to\infty} s_1^m = s_1 = \frac{1}{3}$.

Now from (1) one can prove that $s_n = \lim_{m\to\infty} s_n^m = \lim_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : \mathbf{t}(i) = \mathbf{t}(i+n)\}$ exists for every $n \geqslant 2$, and moreover one gets

$$s_{2n} = s_n, \qquad s_{2n+1} = 1 - \frac{1}{2}(s_n + s_{n+1})$$

for every n, and $s_0 = 1$, $s_1 = \frac{1}{3}$.

Now it is easy to see by induction that $\frac{1}{3} \leqslant s_n \leqslant \frac{2}{3}$ for every $n \geqslant 1$. And since $s_3 = \frac{2}{3}$, then $\mathrm{AM}(\mathbf{t}) = 1 - \frac{2}{3} = \frac{1}{3}$. □

Note that most part of the proof of Theorem 3 was spent on proving the existence of limits $s_n = \lim_{m\to\infty} s_n^m$. If one is ready to assume that these limits exist, then the proof becomes much simpler and shorter. Though we do not know how to prove the existence of these limits simpler, we will be omitting such proofs later, since they are all similar to each other and rather technical.

Theorem 4. *If x is Sturmian, then* $\mathrm{AM}(x) = 0$.

Proof. Let $x = c_{\alpha,\rho}$ be a Sturmian sequence. Recall that by definition $0 < \alpha < 1$, α irrational, and $0 \leqslant \rho < 1$.

Our goal is to show that s_n can be arbitrarily close to 1. Then from the definition of aperiodicity measure it follows that $\mathrm{AM}(x) = 0$.

Let $\varepsilon > 0$. Since α is irrational, we can find n such that $\mathrm{fr}(n\alpha) < \varepsilon$. We have $\frac{1}{m}\#\{i \in [m-1] : x(i) = x(i+n)\} = 1 - \frac{1}{m}\#\{i \in [m-1] : x(i) \neq x(i+n)\}$. Recall that $x(j) = 0$ if $0 \leqslant \mathrm{fr}(\alpha j + \rho) < 1 - \alpha$ and $x(j) = 1$ if $1 - \alpha \leqslant \mathrm{fr}(\alpha j + \rho) < 1$ for every j. Note also that $\mathrm{fr}(\alpha(i+n) + \rho) = \mathrm{fr}(\mathrm{fr}(\alpha i + \rho) + \mathrm{fr}(\alpha n))$. Therefore

$$\begin{aligned}
&\{i \in [m-1] : x(i) \neq x(i+n)\} \\
&\subseteq \{i \in [m-1] : 1 - \alpha - \varepsilon \leqslant \mathrm{fr}(\alpha i + \rho) < 1 - \alpha\} \\
&\qquad \cup \{i \in [m-1] : 1 - \varepsilon \leqslant \mathrm{fr}(\alpha i + \rho) < 1\}.
\end{aligned}$$

Therefore, $\#\{i \in [m-1] : x(i) \neq x(i+n)\} \leqslant \#\{i \in [m-1] : 1 - \alpha - \varepsilon \leqslant \mathrm{fr}(\alpha i + \rho) < 1 - \alpha\} + \#\{i \in [m-1] : 1 - \varepsilon \leqslant \mathrm{fr}(\alpha i + \rho) < 1\}$ which is asymptotically $< 2\varepsilon m$ as $m \to \infty$. Indeed, it is well known that for every irrational β, every real γ, every real a, b such that $0 \leqslant a < b \leqslant 1$, we have $\lim_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : a \leqslant \mathrm{fr}(\beta i + \gamma) \leqslant b\} = b - a$, that is, the sequence $(\mathrm{fr}(\beta i + \gamma))_{i=0}^{\infty}$ is uniformly distributed in $[0, 1]$ (the Kronecker–Weyl Theorem).

Thus $s_n > 1 - 2\varepsilon$. Since ε can be chosen arbitrarily small, it follows that $\mathrm{AM}(x) = 0$. □

4 Automatic Sequences with High Aperiodicity Measure

The following generalization of the Thue–Morse sequence was called Prouhet sequences in [1] (see [17]) and has been widely studied (e.g., see [3, 19] etc.).

Let $\phi\colon \{0, \ldots, k-1\}^* \to \{0, \ldots, k-1\}^*$ be as follows:

$$\begin{aligned}
\phi(0) &= 0\,1\,2\,3 \ldots (k-2)\,(k-1) \\
\phi(1) &= 1\,2\,3 \ldots (k-2)\,(k-1)\,0 \\
\phi(2) &= 2\,3 \ldots (k-2)\,(k-1)\,0\,1 \\
&\ldots \\
\phi(k-1) &= (k-1)\,0\,1\,2\,3 \ldots (k-2),
\end{aligned}$$

in other words, $(\phi(i))(j) = i + j$ (where $+$ is modulo k) for $0 \leqslant i, j \leqslant k-1$. Let $\mathbf{t}_k = \phi^\infty(0)$. Initially it was conjectured that $\mathbf{t}_k$ may have high aperiodicity measure. However, it turns out to be not the case.

Theorem 5. $\mathrm{AM}(\mathbf{t}_k) \leqslant \frac{2}{k+1} - \frac{2}{k^{k-1}(k+1)}$.

Proof (sketch). Remind that $s_n = \limsup_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : x(i) = x(i+n)\}$. Let us generalize this and define $s_n(d) = \limsup_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : x(i+n) - x(i) = d\}$. That is, $s_n(0) = s_n$.

It is clear that $s_0(0) = 1$ and $s_0(d) = 0$ for $1 \leqslant d \leqslant k-1$.

In the same manner as what we did in the proof of Theorem 3, one can prove the existence of limits $s_n(d) = \lim_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : x(i+n) - x(i) = d\}$ and obtain the following equations:

$$s_{kn+p}(d) = \frac{k-p}{k} s_n(d-p) + \frac{p}{k} s_{n+1}(d-p) \tag{2}$$

for every n and every d, p such that $0 \leqslant d, p \leqslant k-1$.

In particular, one can derive the following equations $s_1(d) = \frac{k-1}{k} s_0(d-1) + \frac{1}{k} s_1(d-1)$ for $0 \leqslant d \leqslant k-1$ and prove that

$$s_1(0) = \frac{k-1}{k^k-1} \text{ and } s_1(d) = k^{k-d}\frac{k-1}{k^k-1}$$

for $1 \leqslant d \leqslant k-1$.

One can prove for $0 \leqslant i \leqslant k$ by induction on i that

$$s_{k^i-1}(k-i) = \frac{1}{k^i}\left(1 + \frac{k-1}{k+1}\frac{k^{2i}-1}{k^k-1}\right)$$

using $s_0(0) = 1$ and $s_{k^{i+1}-1}(k-(i+1)) = \frac{1}{k} s_{k^i-1}(k-i) + \frac{k-1}{k} s_{k^i}(k-i) = \frac{1}{k} s_{k^i-1}(k-i) + \frac{k-1}{k} k^i \frac{k-1}{k^k-1}$ (follows from equations (2)). In particular,

$$s_{k^k-1}(0) = \frac{1}{k^k}\left(1 + \frac{k-1}{k+1}\frac{k^{2k}-1}{k^k-1}\right) = 1 - \frac{2}{k+1} + \frac{2}{k^{k-1}(k+1)},$$

from what it follows that $\mathrm{AM}(\mathbf{t}_k) \leqslant \frac{2}{k+1} - \frac{2}{k^{k-1}(k+1)}$. □

We believe that $\mathrm{AM}(\mathbf{t}_k) = \frac{2}{k+1} - \frac{2}{k^{k-1}(k+1)}$ though did not manage to show this. To prove this, one has to find the maximum of the above sequence $s_n(0)$, and we believe that this maximum is indeed reached in $n = k^k - 1$. This statement is supported by computer tests we performed.

An additional interest to study sequences $\mathbf{t}_k$ is in the following alternative definition for these sequences. Let $f_k(i)$ be the sum of digits of i written in base k. Then $\mathbf{t}_k(i) \equiv f_k(i) \pmod k$. This representation is well known for the Thue–Morse sequence. One may ask the following question: what is the number n such that the fraction of numbers i for which $f_k(i) \equiv f_k(i+n) \pmod k$ is maximum possible? We conjecture that this n is $k^k - 1$, that is, the number consisting of $k-1$ digits $k-1$ in base k, and this maximum possible fraction is $1 - \frac{2}{k+1} + \frac{2}{k^{k-1}(k+1)}$, that is, approximately $1 - \frac{2}{k+1}$ for large k.

Other interesting regularities we noticed while performing some computer tests, are the following. Let $s_n^{(k)}(d)$ in this paragraph be the value of $s_n(d)$ for $\mathbf{t}_k$. Let $\operatorname{argmax}_n f(n)$ for $f\colon \mathbb{N} \to \mathbb{R}$ be the smallest value of the argument n on which $f(n)$ reaches its maximum. It seems that $\operatorname{argmax}_n s_n^{(k)}(0) = k^k - 1$ (see above), $\operatorname{argmax}_n s_n^{(k)}(-1) = 1$, $\operatorname{argmax}_n s_n^{(k)}(-2) = k^{k-1} + 1$. It seems also for instance that $\operatorname{argmax}_n s_n^{(4)}(1) = 3332333_4$ (here lower index k means base k representation), and $\operatorname{argmax}_n s_n^{(5)}(1) = 444434444_5$. It also seems that

$\operatorname{argmax}_n s_n^{(5)}(2) = 100010001_5$. Sequences $\mathbf{t}_k$ and the aforementioned regularities should definitely be studied more properly, especially keeping in mind the alternative definition from the previous paragraph.

Now we construct automatic sequences with the aperiodicity measure arbitrarily close to 1.

Theorem 6. *For every $\alpha < 1$ there exists an automatic sequence x such that $\mathrm{AM}(x) \geqslant \alpha$.*

Proof (sketch). Let $k \geqslant 3$ and let $\phi\colon \{0,\dots,k-1\}^* \to \{0,\dots,k-1\}^*$ be such that $(\phi(i))(j) = i+1+2+\dots+(j-1)+j$ (where $+$ is always modulo k) for $0 \leqslant i,j \leqslant k-1$. Let $x_k = \phi^\infty(0)$. For instance, if $k=5$, then ϕ is as follows:

$$\begin{aligned}\phi(0) &= 01310\\ \phi(1) &= 12421\\ \phi(2) &= 23032\\ \phi(3) &= 34143\\ \phi(4) &= 40204,\end{aligned}$$

and $x_5 = 013101242134143124210131012421\dots$

Claim. If $k \geqslant 3$ is prime, then $\mathrm{AM}(x_k) = 1 - \frac{2}{k}$.

Let us define $s_n^m(d) = \frac{1}{m}\#\{i \in [m-1] : x_k(i+n) - x_k(i) = d\}$.

It follows from the definition that for every m we have $s_0^m(0) = 1$ and $s_0^m(d) = 0$ for $1 \leqslant d \leqslant k-1$. Analogously to the proof of Theorem 3, for every n, every $m \geqslant 1$, and every d, p, t such that $0 \leqslant d,p,t \leqslant k-1$, one can obtain the following equations (compare with equations (1)):

$$\begin{aligned} &s_{kn+p}^{km+t}(d)\\ &= \frac{1}{km+t}\sum_{j=0}^{k-p-1} m_j s_n^{m_j}\left(d - \frac{p(p+1)}{2} - jp\right)\\ &+ \frac{1}{km+t}\sum_{j=k-p}^{k-1} m_j s_{n+1}^{m_j}\left(d - \frac{p(p+1)}{2} - jp\right), \end{aligned} \tag{3}$$

where $m_j = m+1$ for $j < t$ and $m_j = m$ for $j \geqslant t$. The idea again is to consider separately sets of indices $\{ik+j : i \in \mathbb{N}\}$ for different j such that $0 \leqslant j \leqslant k-1$.

In particular, one can derive from (3) that $s_1^{km+t}(0) = \frac{m}{km+t}s_1^m(0)$ and $s_1^{km+t}(d) = \frac{1}{km+t}\left(m_{d-1} + ms_1^m(d)\right)$ for $1 \leqslant d \leqslant k-1$. Our goal is to prove that there exist $\lim_{m\to\infty} s_1^m(0) = 0$ and $\lim_{m\to\infty} s_1^m(d) = \frac{1}{k-1}$ for $1 \leqslant d \leqslant k-1$.

The former equation is clear, since $s_1^t(0) = 0$ for $1 \leqslant t \leqslant k-1$ can be checked easily. For the latter, fix some d such that $1 \leqslant d \leqslant k-1$ and let b_m be such that $s_1^m(d) = \frac{1}{k-1} + \frac{b_m}{m}$. Then one gets $b_{km+t} = m_{d-1} - m - \frac{t}{k-1} + b_m$, from what it is easy to see that $|b_m| = O(\log m)$. Therefore there exists $\lim_{m\to\infty} s_1^m(d) = s_1(d) = \frac{1}{k-1}$.

Using (3), now one can prove by induction on n the existence of limits $s_n(d) = \lim_{m\to\infty} \frac{1}{m}\#\{i \in [m-1] : x(i+n) - x(i) = d\}$ and to obtain the following equations:

$$s_{kn+p}(d) = \frac{1}{k}\left(\sum_{j=0}^{k-p-1} s_n\left(d - \frac{p(p+1)}{2} - jp\right) + \sum_{j=k-p}^{k-1} s_{n+1}\left(d - \frac{p(p+1)}{2} - jp\right)\right) \tag{4}$$

for every n and every d, p such that $0 \leqslant d, p \leqslant k-1$.

For instance, for $k = 5$ we get

$$s_{5m}(d) = s_m(d)$$

$$s_{5m+1}(d) = \frac{1}{5}(s_m(d-1) + s_m(d-2) + s_m(d-3) + s_m(d-4) + s_{m+1}(d-5))$$

$$s_{5m+2}(d) = \frac{1}{5}(s_m(d-3) + s_m(d-5) + s_m(d-7) + s_{m+1}(d-9) + s_{m+1}(d-11))$$

...

Clearly, $s_0(0) = 1$ and $s_0(d) = 0$ for $1 \leqslant d \leqslant k-1$. We already proved that $s_1(0) = 0$ and $s_1(d) = \frac{1}{k-1}$ for $1 \leqslant d \leqslant k-1$.

Using (4), it is easy to see that $s_{k-1}(0) = \frac{1}{k}s_0(0) + \frac{1}{k}\sum_{j=1}^{k-1} s_1(j) = \frac{2}{k}$.

Now it is easy to prove using equations (4) that $s_n(d) \leqslant \frac{2}{k}$ for every d and $n \geqslant 1$. Indeed, note that k is prime (this is the first time we use it), and therefore $\{d - \frac{p(p+1)}{2} - p, d - \frac{p(p+1)}{2} - 2p, \ldots, d - \frac{p(p+1)}{2} - kp\} = \{0, \ldots, k-1\}$. Thus for $n \geqslant 0$ and $1 \leqslant p \leqslant k-1$ we have

$$\begin{aligned} &s_{kn+p}(d) \\ &= \frac{1}{k}\left(\sum_{j=0}^{k-p-1} s_n\left(d - \frac{p(p+1)}{2} - jp\right) + \sum_{j=k-p}^{k-1} s_{n+1}\left(d - \frac{p(p+1)}{2} - jp\right)\right) \\ &\leqslant \frac{1}{k}\sum_{j=0}^{k-1} s_n(j) + \frac{1}{k}\sum_{j=0}^{k-1} s_{n+1}(j) = \frac{1}{k} + \frac{1}{k} = \frac{2}{k}, \end{aligned}$$

and we also need $s_{kn}(d) = s_n(d)$ for $n \geqslant 1$.

Therefore $\mathrm{AM}(x_k) = 1 - \sup\{s_n(0) : n \geqslant 1\} = 1 - s_{k-1}(0) = 1 - \frac{2}{k}$. □

5 Conclusion and Open Problems

In this paper we introduced the notion of aperiodicity measure for infinite symbolic sequences. It seems that this notion was not studied before, though looks very natural at least from a combinatorial point of view. However, the results of our paper are far from sufficient before we could say that the notion of aperiodicity measure is properly studied. Here we formulate some open questions, in addition to those listed throughout the paper, that we think may be interesting for future research.

1. As we already discussed, $\mathrm{AM}(x)$ of a sequence x over the alphabet with k symbols ranges from 0 to $1-\frac{1}{k}$. What values in this range may the aperiodicity measure have?
2. For each $k \geqslant 2$, what is the maximum possible $\mathrm{AM}(x)$ for an automatic sequence x over the alphabet of k symbols? For a morphic sequence? What values can the aperiodicity measure of a morphic sequence have?
3. For each $k \geqslant 2$, what is the maximum possible $\mathrm{AM}(x)$ for a k-automatic x? What values can the aperiodicity measure of a k-automatic sequence have?
4. For which sequences x can one take lim instead of lim sup in the definition of $s_n(x) = \limsup \frac{1}{m}\#\{i \in [m-1] : x(i) = x(i+n)\}$, that is, when does this limit exist? In particular, does it exist for every automatic sequence x? Is it true that if this limit exists for $n = 1$, then it exists for all n?
5. Study the behavior of the sequence s_n, and more generally, of the sequence $s_n(i)$ for different i, more properly. In particular, describe its set of accumulation points.
6. Calculate the aperiodicity measure for some other sequences and classes of sequences, for instance, for the Toeplitz sequences, some morphic sequences, some generalizations of the Sturmian sequences, etc.
7. We characterized the aperiodicity measure of some sequences and suggested that this work should be continued. However, one can also ask the inverse question: to characterize the set of sequences with some fixed aperiodicity measure α. This is specifically interesting for the extremal values $\alpha = 0$ and $\alpha = 1 - \frac{1}{k}$ for k-letter sequences.
8. There is a generic way (an algorithm) to calculate the aperiodicity measure for morphic sequences. For example, this can be seen in the following way. Let the Cartesian product of sequences x and y be the sequence $x \times y$ such that $(x \times y)(n) = (x(n), y(n))$. Then for a morphic x, the sequence $x \times L^p x$ is morhic, since it can be obtained from x by a finite transduction (e.g., see [4]). Then it only remains to calculate the frequency of letters in $x \times L^p x$ (e.g., see [18]). However, this method is clearly non-practical. Is there a sufficiently simpler method to calculate the aperiodicity measure for morphic sequences? Or at least for automatic sequences? In particular, can one generalize the method used in the proofs of Theorems 3, 5, 6?

Acknowledgements

The authors are grateful to A. Semenov and N. Vereshchagin for their permanent support, and also to S. Avgustinovich, B. Durand, A. Frid, A. Romashchenko, A. Shen for fruitful discussions, as well as to anonymous referees for many useful comments.

References

1. A. Adler and S.-Y. R. Li. Magic cubes and Prouhet sequences. *American Mathematical Monthly*, 84:618–627, 1977.

2. J.-P. Allouche and J. Shallit. The ubiquitous Prouhet–Thue–Morse sequence. In *Sequences and their applications, Proceedings of SETA'98*, pages 1–16. Springer-Verlag, 1999.
3. J.-P. Allouche and J. Shallit. Sums of digits, overlaps, and palindromes. 4:1–10, 2000.
4. J.-P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, 2003.
5. J. Cassaigne, C. Mauduit, and A. Sárközy. On finite pseudorandom binary sequences. VII. The measures of pseudorandomness. *Acta Arithmetica*, 103(2):97–118, 2002.
6. A. Cobham. Uniform tag sequences. *Mathematical Systems Theory*, 6:164–192, 1972.
7. B. Durand, A. Romashchenko, and A. Shen. Fixed point and aperiodic tilings. In *Proceedings of DLT 2008*, volume 5257 of *Lecture Notes in Computer Science*, pages 276–288. Springer-Verlag, 2008. See also `http://arxiv.org/abs/0802.2432`.
8. S. Ferenczi. Tiling and local rank properties of the morse sequence. *Theoretical Computer Science*, 129:369–383, 1994.
9. E. Grant, J. Shallit, and T. Stoll. Bounds for the discrete correlation of infinite sequences on k symbols and generalized Rudin–Shapiro sequences. Preprint arXiv:0812.3186, 2008.
10. D. Krieger, P. Ochem, N. Rampersad, and J. Shallit. Avoiding approximate squares. In *Proceedings of DLT 2007*, volume 4588 of *Lecture Notes in Computer Science*, pages 278–289. Springer-Verlag, 2007.
11. M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.
12. C. Mauduit and A. Sárközy. On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol. *Acta Arithmetica*, 82(4):365–377, 1997.
13. C. Mauduit and A. Sárközy. On the measures of pseudorandomness of binary sequences. *Discrete Mathematics*, 271:195–207, 2003.
14. M. Morse and G. A. Hedlund. Symbolic dynamics. *American Journal of Mathematics*, 60(4):815–866, 1938.
15. M. Morse and G. A. Hedlund. Symbolic dynamics ii: Sturmian trajectories. *American Journal of Mathematics*, 62:1–42, 1940.
16. An. A. Muchnik, Yu. L. Pritykin, and A. L. Semenov. Sequences close to periodic. Preprint arXiv:0903.5316, 2009 (in Russian, to appear in English).
17. M. E. Prouhet. Mémoire sur quelques relations entre les puissances des nombres. *Comptes Rendus des Séances de l'Académie des Sciences*, 33:225, 1851. Available at `http://gallica.bnf.fr/ark:/12148/bpt6k29901.image.f227.langFR`.
18. K. Saari. On the frequency of letters in morphic sequences. In *Proceedings of CSR 2006*, volume 3967 of *Lecture Notes in Computer Science*, pages 334–345, 2006.
19. P. Séébold. On some generalizations of the Thue–Morse morphism. *Theoretical Computer Science*, 292:283–298, 2003.