Automatic sequences and generalised polynomials

Jakub Byszewski, Jakub Konieczny

AUTOMATIC SEQUENCES AND GENERALISED POLYNOMIALS

JAKUB BYSZEWSKI AND JAKUB KONIECZNY

Abstract. We conjecture that bounded generalised polynomial functions cannot be generated by finite automata, except for the trivial case when they are ultimately periodic.

Using methods from ergodic theory, we are able to partially resolve this conjecture, proving that any hypothetical counterexample is periodic away from a very sparse and structured set. In particular, we show that for a polynomial $p(n)$ with at least one irrational coefficient (except for the constant one) and integer $m \geq 2$, the sequence $\lfloor p(n) \rfloor \bmod m$ is never automatic.

We also prove that the conjecture is equivalent to the claim that the set of powers of an integer $k \geq 2$ is not given by a generalised polynomial.

Introduction

Automatic sequences are sequences whose n-th term is produced by a finite-state machine from the base-k digits of n. (A precise definition is given below.) By definition, automatic sequences can take only finitely many values. Allouche and Shallit [AS92, AS03b] have generalised the notion of automatic sequences to a wider class of regular sequences and demonstrated its ubiquity and links with multiple branches of mathematics and computer science. In a broader context, automatic sequences are closely tied to regular languages (a sequence is automatic precisely when its level sets are regular), which are at the lowest level in Chomsky's hierarchy of formal languages [Cho59]. The problem of demonstrating that a certain sequence is or is not automatic or regular has been widely studied, particularly for sequences of arithmetic origin (see, e.g., [AS92, AS03b, Bel07, SY11, MR15, SP11, Mos08, Row10]).

The aim of this article is to continue the aforementioned study for sequences that arise from generalised polynomials, i.e., expressions involving algebraic operations and the floor function. More generally, we are interested in determining the simplest generalised polynomials in terms of Chomsky's hierarchy, apart from the rather trivial examples of ultimately periodic sequences; in practice, only the question concerning regular sequences seems amenable to analysis. Our methods rely on a number of dynamical and ergodic tools. A crucial ingredient in our work is one of the main results from the companion paper [BK18b] concerning the combinatorial structure of the set of times at which an orbit on a nilmanifold hits a semialgebraic subset. This is possible because by the work of Bergelson and Leibman [BL07] generalised polynomials are closely related to dynamics on nilmanifolds.

2010 *Mathematics Subject Classification.* Primary: 11B85, 37A45. Secondary: 37B05, 37B10, 11J71, 11B37, 05C20.

Key words and phrases. Generalised polynomials, automatic sequences, IP sets, nilmanifolds, linear recurrence sequences, regular sequences.

Another motivation for this line of research stems from additive combinatorics. It is natural to study existence and prevalence of linear patterns in sets of integers defined by finite automata. In his seminal work on a Fourier analytic proof of Szemerédi's theorem, Gowers introduced uniformity norms, which control the count of arithmetic progressions and other patterns [Gow01]. The second-named author showed that the Thue–Morse sequence is Gowers uniform of all orders [Kon16], which in particular implies that the number of arithmetic progressions of a given length inside the integers up to N with even sum of binary digits is, up to a small error, the same as for a random set of density $1/2$. The same result holds for the Rudin–Shapiro sequence as well as for all non-periodic k-multiplicative sequences [FK18]. In full generality, one expects that any automatic sequence should have a decomposition as the sum of a simple structured part and a Gowers uniform error term (this can be construed as a variant of the arithmetic regularity lemma [GT10], albeit with a much stronger conclusion, and applicable to a much smaller class of sequences). In absence of such a result, we aim to answer a simpler question. It follows from the celebrated inverse theorem for Gowers norms [GTZ12] that all obstructions to Gowers uniformity come from nilsequences, or – equivalently (cf. [BL07]) – from bounded generalised polynomials. Hence, one is led to investigate the relation between generalised polynomials and automatic sequences.

In [AS03b, Theorem 6.2] it is proved that the sequence $(f(n))_{n\geq 0}$ given by $f(n) = \lfloor \alpha n+\beta \rfloor$ for real numbers α, β is regular if and only if α is rational. The method used there does not immediately generalise to higher degree polynomials in n, but the proof implicitly uses rotation on a circle by an angle of $2\pi\alpha$. Replacing the rotation on a circle by a skew product transformation on a torus (as in Furstenberg's proof of Weyl's equidistribution theorem [Fur61]), we easily obtain the following result. (For more on regular sequences, see Section 1.) A generalisation of this result is obtained in [BK18a].

Theorem A. *Let $p \in \mathbb{R}[x]$ be a polynomial. Then the sequence $f(n) = \lfloor p(n) \rfloor, n \geq 0$, is regular if and only if all the coefficients of p except possibly for the constant term are rational.*

In fact, we show the stronger property that for any integer $m \geq 2$ the sequence $\lfloor f(n) \rfloor \bmod m$ is not automatic unless all the coefficients of p except for the constant term are rational, in which case the sequence is periodic. It is natural to inquire whether a similar result can be proven for more complicated expressions involving the floor function such as, e.g., $f(n) = \lfloor \alpha \lfloor \beta n^2 + \gamma \rfloor^2 + \delta n + \varepsilon \rfloor$. Such sequences are called generalised polynomial and have been intensely studied (see, e.g., [Hål93, Hål94, HK95, BL07, Lei12, GTZ12, GT12]).

Another closely related motivating example comes from the classical Fibonacci word[1] $w_{\mathrm{Fib}} \in \{0,1\}^{\mathbb{N}_0}$, whose systematic study was initiated by Berstel [Ber81, Ber85] (for historical notes, see [AS03a, Sec. 7.12]). There are several ways to define it, each shedding light from a different direction.

(*i*) *Morphic word.* Define the sequence of words $w_0 := 0$, $w_1 := 01$, and $w_{i+2} := w_{i+1}w_i$ for $i \geq 0$. Then w_{Fib} is the (coordinate-wise) limit of w_i as $i \to \infty$.

(*ii*) *Sturmian word.* Explicitly, $w_{\mathrm{Fib}}(n) = \lfloor (2-\varphi)(n+2) \rfloor - \lfloor (2-\varphi)(n+1) \rfloor$.

[1]We will freely identify words in $\Omega^{\mathbb{N}_0}$ with functions $\mathbb{N}_0 \to \Omega$.

(*iii*) *Fib-automatic sequence.* If a positive integer n is written in the form $n = \sum_{i=2}^{d} v_i F_i$, where $v_i \in \{0,1\}$ and there is no i with $v_i = v_{i+1} = 1$, then $w_{\mathrm{Fib}}(n) = v_2$.

The equivalence of (i) and (ii) is well-known, see, e.g., [Lot02, Chpt. 2]. The representation $v_d v_{d-1} \cdots v_2$ of n as a sum of Fibonacci numbers in (iii) is known as the Zeckendorf representation; it exists for each n and is unique. The notion of automaticity using Zeckendorf representation (or, for that matter, a representation from a much wider class) in place of the usual base-k representation of the input n was introduced and studied by Shallit in [Sha88] (see also [Rig00]), where among other things the equivalence of (i) and (iii) is shown. We return to this subject in Section 6.

Hence, w_{Fib} gives a non-trivial example of a sequence which is given by a generalised polynomial and satisfies a variant of automaticity related to the Zeckendorf representation. It is natural to ask if similar examples exist for the usual notion of k-automaticity. Motivated by Theorem A, we believe the answer is essentially negative, except for trivial examples. We say that a sequence f is *ultimately periodic* if it coincides with a periodic sequence except on a finite set. The following conjecture was the initial motivation for the line of research pursued in this paper.

Conjecture A. *Suppose that a sequence f is simultaneously automatic and generalised polynomial. Then f is ultimately periodic.*

In this paper, we prove several slightly weaker variants of Conjecture A. First of all, we prove that the conjecture holds except on a set of density zero. In fact, in order to obtain such a result, we only need a specific property of automatic sequences. For the purpose of stating the next theorem, let us say that a sequence $f\colon \mathbb{N} \to X$ is *weakly periodic* if for any restriction f' of f to an arithmetic sequence given by $f'(n) = f(an+b)$, $a \in \mathbb{N}$, $b \in \mathbb{N}_0$, there exist $q \in \mathbb{N}$, $r, r' \in \mathbb{N}_0$ with $r \neq r'$, such that $f'(qn+r) = f'(qn+r')$. Of course, any periodic sequence is weakly periodic, but not conversely. All automatic sequences are weakly periodic (this follows from the fact that automatic sequences have finite kernels, see Lemma 2.1). Another non-trivial example of a weakly periodic sequence is the characteristic function of the square-free numbers.

Theorem B. *Suppose that a sequence $f\colon \mathbb{N}_0 \to \mathbb{R}$ is weakly periodic and generalised polynomial. Then there exists a periodic function $b\colon \mathbb{N}_0 \to \mathbb{R}$ and a set $Z \subset \mathbb{N}_0$ of upper Banach density zero such that $f(n) = b(n)$ for $n \in \mathbb{N}_0 \setminus Z$.*

(For the definition of Banach density, see Section 1.)

Theorem B is already sufficient to rule out automaticity of many natural examples of generalised polynomials. In particular, sequences such as $\lfloor \sqrt{2} n \lfloor \sqrt{3} n \rfloor \rfloor \bmod 10$ or $\left\lfloor \sqrt{2} n \lfloor \sqrt{3} n \rfloor^2 + \sqrt{5} n + \sqrt{7} \right\rfloor \bmod 10$ are not automatic. For details and more examples, see Corollary 2.7.

To obtain stronger bounds on the size of the "exceptional" set Z, we restrict ourselves to automatic sequences and exploit some finer properties of generalised polynomials studied in the companion paper [BK18b]. We use results concerning growth properties of automatic sequences to derive the following dichotomy: If $a\colon \mathbb{N}_0 \to \{0,1\}$ is an automatic sequence, then the set of integers where a takes the value 1 is either combinatorially rich (it contains what we call an IPS set) or extremely sparse (in particular, the number of its elements up to N grows as $\log^r(N)$

for some integer r); see Theorem 3.13. This result is especially interesting for sparse automatic sequences, i.e., automatic sequences which take non-zero values on a set of integers of density 0. Conversely, in [BK18b] we show that sparse generalised polynomials must be free of similar combinatorial structures. As a consequence, we prove the following result.

Theorem C. *Suppose that a sequence $f\colon \mathbb{N}_0 \to \mathbb{R}$ is automatic and generalised polynomial. Then there exists a periodic function $b\colon \mathbb{N}_0 \to \mathbb{R}$, a set $Z \subset \mathbb{N}_0$, and a constant r such that $f(n) = b(n)$ for $n \in \mathbb{N}_0 \setminus Z$ and*

$$\sup_M |Z \cap [M, M+N)| = O\left(\log^r(N)\right)$$

as $N \to \infty$ for a certain constant r (dependent on f).

In fact, we obtain a much more precise structural description of the exceptional set Z (see Theorem 3.7 for details). Similar techniques allow us to show non-automaticity of some sparse generalised polynomials. For instance, the sequence given by

$$n \mapsto \begin{cases} 1, & \text{if } \left\| \sqrt{2} n \left\lfloor \sqrt{3} n \right\rfloor \right\| < n^{-c}; \\ 0, & \text{otherwise} \end{cases}$$

is not automatic provided that c is small enough. (Here, $\|x\|$ denotes the distance of x from $\mathbb{Z}$.) For details, see Example 4.7.

While Theorem C does not resolve Conjecture A, our proof thereof greatly restricts the number of possible counterexamples. In fact, in order to prove Conjecture A, it would suffice to prove that the characteristic sequence of powers of an integer $k \geq 2$ given by

$$g_k(n) = \begin{cases} 1, & \text{if } n = k^t \text{ for some } t \geq 0; \\ 0, & \text{otherwise} \end{cases}$$

is not a generalised polynomial.

Theorem D. *Let $k \geq 2$ be an integer. Then exactly one of the following statements holds:*

- *(i) All sequences that are simultaneously k-automatic and generalised polynomial are ultimately periodic.*
- *(ii) The characteristic sequence g_k of the powers of k is generalised polynomial.*

Unfortunately, we are currently unable to decide which of the two possibilities in Theorem D holds. Although we expect that g_k should not be a generalised polynomial, in [BK18b] we obtain several examples of algebraic numbers $\lambda > 1$ such that the characteristic function of the set $E_\lambda := \left\{ \langle\!\langle \lambda^i \rangle\!\rangle \;\middle|\; i \in \mathbb{N}_0 \right\}$ is generalised polynomial, where $\langle\!\langle x \rangle\!\rangle$ denotes the closest integer to x. All our examples are Pisot units (a Pisot number is an algebraic integer $\lambda > 1$ all of whose conjugates have modulus < 1; a Pisot unit is a Pisot number whose minimal polynomial has constant term ± 1). Conversely, there is no $\lambda > 1$ for which we can prove that the characteristic function of E_λ is not given by a generalised polynomial. This prompts us to propose the following question.

Question A. *Suppose that $\lambda > 1$ is such that the characteristic function of the set $E_\lambda := \left\{ \langle\!\langle \lambda^i \rangle\!\rangle \;\middle|\; i \in \mathbb{N}_0 \right\}$ is given by a generalised polynomial. Is it then necessarily the case that λ is a Pisot unit?*

For a more detailed discussion of this question, see [BK18b, Section 6]. If λ is a Pisot number, then $\langle\langle \lambda^i \rangle\rangle$ obeys a linear recurrence. We show that for such λ, the characteristic function of E_λ cannot be a counterexample to Conjecture A (see Proposition 4.9) except possibly if λ is an integer.

By Theorem D, determining the validity of Conjecture A is equivalent to answering Question A in the special case when λ is an integer.

Contents. In Section 1, we discuss some basic notions and results concerning automatic sequences and dynamical systems. We intended this section to be accessible to readers familiar with only one (or neither) of these topics. In Section 2, we prove Theorem A and Theorem B using methods from topological dynamics. In Section 3, we use known results on growth and structure of automatic sequences to prove that they are either very sparse and structured (in which case we call them arid) or are combinatorially rich. Together with a result about dynamics on nilmanifolds, this allows us to obtain Theorem C. Section 4 contains four seperate topics concerning examples and non-examples of automatic sets and uniform density of symbols in automatic sequences. Section 5 is devoted to the proof of Theorem D. Finally, Section 6 discusses some open problems and future research topics.

Acknowledgements. The authors thank Ben Green for much useful advice during the work on this project, Vitaly Bergelson and Inger Håland-Knutson for valuable comments on the distribution of generalised polynomials, and Jean-Paul Allouche and Narad Rampersad for information about related results on automatic sequences. We are also grateful to the anonymous referee for his or her many comments, and in particular for suggesting that it should be possible to decompose an arid set into a union of pairwise disjoint basic arid sets.

Thanks also go to Sean Eberhard, Dominik Kwietniak, Freddie Manners, Rudi Mrazović, Przemek Mazur, Sofia Lindqvist, and Aled Walker for many informal discussions.

This research was supported by the National Science Centre, Poland (NCN) under grant no. DEC-2012/07/E/ST1/00185.

Finally, we would like to express our gratitude to the organisers of the conference *New developments around* $\times 2$ $\times 3$ *conjecture and other classical problems in Ergodic Theory* in Cieplice, Poland in May 2016 where we began our project.

1. Background

Notations and generalities. We denote the sets of positive integers and of non-negative integers by $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \{0, 1, \ldots\}$. We denote by $[N]$ the set $[N] = \{0, 1, \ldots, N-1\}$. We use the Iverson convention: whenever φ is any sentence, we denote by $[\![\varphi]\!]$ its logical value (1 if φ is true and 0 otherwise). We denote the number of elements in a finite set A by $|A|$.

For a real number r, we denote its integer part by $\lfloor r \rfloor$, its fractional part by $\{r\} = r - \lfloor r \rfloor$, the nearest integer to r by $\langle\langle r \rangle\rangle = \lfloor r + 1/2 \rfloor$, and the distance from r to the nearest integer by $\|r\| = |r - \langle\langle r \rangle\rangle|$.

We use some standard asymptotic notation. Let f and g be two functions defined for sufficiently large integers. We say that $f = O(g)$ or $f \ll g$ if there exists $c > 0$ such that $|f(n)| \leq c\,|g(n)|$ for sufficiently large n. We say that $f = o(g)$ if for every $c > 0$ we have $|f(n)| \leq c|g(n)|$ for sufficiently large n.

For a subset $E \subset \mathbb{N}_0$, we say that E has *natural density* $d(A)$ if

$$\lim_{N\to\infty} \frac{|E \cap [N]|}{N} = d(A).$$

We say that $E \subset \mathbb{N}_0$ has *upper Banach density* $d^*(A)$ if

$$\limsup_{N\to\infty} \max_{M} \frac{|E \cap [M, M+N)|}{N} = d^*(A).$$

We now formally define generalised polynomials.

Definition 1.1 (Generalised polynomial)**.** The family GP of *generalised polynomials* is the smallest set of functions $\mathbb{Z} \to \mathbb{R}$ containing the polynomial maps and closed under addition, multiplication, and the operation of taking the integer part. Whenever it is more convenient, we regard generalised polynomials as functions on $\mathbb{N}_0$.

A set $E \subset \mathbb{Z}$ (or $E \subset \mathbb{N}_0$) is called *generalised polynomial* if its characteristic function given by $f(n) = [\![n \in E]\!]$ is a generalised polynomial. (Note that this definition depends on whether we are regarding the generalised polynomial as a function on $\mathbb{Z}$ or on $\mathbb{N}_0$ and a generalised polynomial set $E \subset \mathbb{N}_0$ might a priori not be generalised polynomial when considered as a subset of $\mathbb{Z}$. It will always be clear from the context which meaning we have in mind.)

An example of a generalised polynomial is therefore a function f given by the formula $f(n) = \sqrt{3}\lfloor\sqrt{2}n^2 + 1/7\rfloor^2 + n\lfloor n^3 + \pi\rfloor$.

Automatic sequences. Whenever A is a (finite) set, we denote the free monoid with basis A by A^*. It consists of finite words in A, including the empty word ϵ, with the operation of concatenation. We denote the concatenation of two words $v, w \in A^*$ by vw and we denote the length of a word $w \in A^*$ by $|w|$. In particular, $|\epsilon| = 0$. We say that a word $v \in A^*$ is a factor of a word $w \in A^*$ if there exist words $u, u' \in A^*$ such that $w = uvu'$. We denote by $w^{\mathrm{R}} \in A^*$ the reversal of the word $w \in A^*$ (the word in which the elements of A are written in the opposite order).

Let $k \geq 2$ be an integer and denote by $\Sigma_k = \{0, 1, \ldots, k-1\}$ the set of digits in base k. For $w \in \Sigma_k^*$, we denote by $[w]_k$ the integer whose expansion in base k is w, i.e., if $w = v_l v_{l-1} \cdots v_1 v_0$, $v_i \in \Sigma_k$, then $[w]_k = \sum_{i=0}^{l} v_i k^i$. Conversely, for an integer $n \geq 0$, we write $(n)_k \in \Sigma_k^*$ for the base-k representation of n (without an initial zero). In particular, $(0)_k = \epsilon$.

The class of automatic sequences consists, informally speaking, of finite-valued sequences $(a_n)_{n\geq 0}$ whose values a_n are obtained via a finite procedure from the digits of base-k expansion of an integer n.

The most famous example of an automatic sequence is arguably the Thue–Morse sequence, first discovered by Prouhet in 1851. Let $s_2(n)$ denote the sum of digits of the base 2 expansion of an integer n. Then the Thue–Morse sequence $(t_n)_{n\geq 0}$ is given by $t_n = 1$ if $s_2(n)$ is odd and $t_n = 0$ if $s_2(n)$ is even.

We will introduce the basic properties of automatic sequences. For more information, we refer the reader to the canonical book of Allouche and Shallit [AS03a]. To formally introduce the notion of automatic sequences, we begin by discussing finite automata.

Definition 1.2. A deterministic finite *k-automaton with output* (which we will just call a k-automaton) $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \Omega, \tau)$ consists of the following data:

(*i*) a finite set of states S;
(*ii*) an initial state $s_0 \in S$;
(*iii*) a transition map $\delta\colon S \times \Sigma_k \to S$;
(*iv*) an output set Ω;
(*v*) an output map $\tau\colon S \to \Omega$.

We extend the map δ to a map $\delta\colon S \times \Sigma_k^* \to S$ (denoted by the same letter) by the recurrence formula

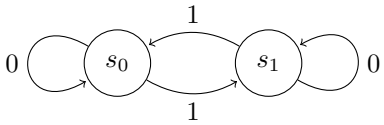
$$\delta(s, \epsilon) = s, \quad \delta(s, wv) = \delta(\delta(s, w), v), \quad s \in S, w \in \Sigma_k^*, v \in \Sigma_k.$$

We call a sequence *k-automatic* if it can be produced by a k-automaton in the following manner: one starts at the initial state of the automaton, follows the digits of the base-k expansion of an integer n, and then uses the output function to print the n-th term of the sequence. This is stated more precisely in the following definition.

Definition 1.3. A sequence $(a_n)_{n\geq 0}$ with values in a finite set Ω is *k-automatic* if there exists a k-automaton $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \Omega, \tau)$ such that $a_n = \tau\left(\delta(s_0, (n)_k)\right)$. We call a set E of nonnegative integers *automatic* if the characteristic sequence $(a_n)_{n\geq 0}$ of E given by $a_n = [\![n \in E]\!]$ is automatic.

For some applications, it will be useful to consider the following variant of the definition. A function $\tilde{a}\colon \Sigma_k^* \to \Omega$ is *automatic* if there exists a k-automaton $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \Omega, \tau)$ such that $\tilde{a}(u) = \tau(\delta(s_0, u))$ for $u \in \Sigma_k^*$.

The values of the Thue–Morse sequence are given by the 2-automaton



with nodes depicting the states of the automaton, edges describing the transition map, $\tau(s_0) = 0$, and $\tau(s_1) = 1$. Thus, the Thue–Morse sequence is 2-automatic.

In the definition above, the automaton reads the digits starting with the most significant one. In fact, we might equally well demand that the digits be read starting with the least significant digit or that the automaton produce the correct answer even if the input contains some leading zeros. Neither of these modifications changes the notion of automatic sequence [AS03a, Theorem 5.2.3] (though of course for most sequences we would need to use a different automaton to produce a given automatic sequence).

There is a number of equivalent definitions of the notion of automatic sequence connecting them to different branches of mathematics (stated for example in terms of algebraic power series over finite fields or letter-to-letter projections of fixed points of uniform morphisms of free monoids). We will need one such definition that has a combinatorial flavour and is expressed in terms of the k-kernel.

Definition 1.4. The *k-kernel* $\mathcal{N}_k((a_n))$ of a sequence $(a_n)_{n\geq 0}$ is the set of its subsequences of the form

$$\mathcal{N}_k((a_n)) = \{(a_{k^l n+r})_{n\geq 0} \mid l \geq 0, 0 \leq r < k^l\}.$$

Automaticity of a sequence is equivalent to finiteness of its kernel, originally due to Eilenberg [Eil74].

Proposition 1.5. [AS03a, Theorem 6.6.2] *Let $(a_n)_{n\geq 0}$ be a sequence. Then the following conditions are equivalent:*

(*i*) *The sequence (a_n) is k-automatic.*
(*ii*) *The k-kernel $\mathcal{N}_k((a_n))$ is finite.*

For the Thue–Morse sequence we have the relations $t_{2n} = t_n$, $t_{2n+1} = 1 - t_n$, and hence one easily sees that the 2-kernel $\mathcal{N}_2((t_n))$ consists of only two sequences $\mathcal{N}_2((t_n)) = \{t_n, 1 - t_n\}$. This gives another argument for the 2-automaticity of the Thue–Morse sequence.

An automatic sequence by definition takes only finitely many values. In 1992 Allouche and Shalit [AS92] generalised the notion of automatic sequences to the wider class of k-regular sequences that are allowed to take values in a possibly infinite set. The definition of regular sequences is stated in terms of the k-kernel. For simplicity, we state the definition over the ring of integers, though it could also be introduced over a general (noetherian) ring.

Definition 1.6. Let $(a_n)_{n\geq 0}$ be a sequence of integers. We say that the sequence (a_n) is *k-regular* if its k-kernel $\mathcal{N}_k((a_n))$ spans a finitely generated abelian subgroup of $\mathbb{Z}^{\mathbb{N}_0}$.

For example, the following sequences are easily seen to be 2-regular: $(t_n)_{n\geq 0}$, $(n^3 + 5)_{n\geq 0}$, $(s_2(n))_{n\geq 0}$. (The corresponding subgroups spanned by the 2-kernel have rank 2, 4, and 2, respectively. In the case of $t = (t_n)_{n\geq 0}$, the subgroup spanned by the 2-kernel is free abelian with basis consisting of t and the constant sequence $(1)_{n\geq 0}$.) In fact, every k-automatic (integer-valued) sequence is obviously k-regular, and the following converse result holds.

Theorem 1.7. [AS03a, Theorem 16.1.5] *Let $(a_n)_{n\geq 0}$ be a sequence of integers. Then the following conditions are equivalent:*

(*i*) *The sequence (a_n) is k-automatic.*
(*ii*) *The sequence (a_n) is k-regular and takes only finitely many values.*

Corollary 1.8. [AS03a, Corollary 16.1.6] *Let $(a_n)_{n\geq 0}$ be a sequence of integers that is k-regular and let $m \geq 1$ be an integer. Then the sequence $(a_n \bmod m)$ is k-automatic.*

A convenient tool for ruling out that a given sequence is automatic is provided by the pumping lemma.

Lemma 1.9. [AS03a, Lemma 4.2.1] *Let $(a_n)_{n\geq 0}$ be a k-automatic sequence. Then there exists a constant N such that for any $w \in \Sigma_k^*$ with $|w| \geq N$ and any integer $0 \leq L \leq |w| - N$ there exist $u_0, u_1, v \in \Sigma_k^*$ such that $v \neq \epsilon$, $w = u_0 v u_1$, $L \leq |u_0| \leq L + N - |v|$, and a_n takes the same value for all $n \in \{[u_0 v^t u_1]_k \mid t \in \mathbb{N}_0\}$.*

The final issue that we need to discuss is the dependence of the notion of k-automaticity on the base k. While the Thue–Morse sequence is 2-regular, and is also easily seen to be 4-regular, it is not 3-regular. This follows from the celebrated result of Cobham [Cob69]. We say that two integers $k, l \geq 2$ are *multiplicatively independent* if they are not both powers of the same integer (equivalently, $\log k / \log l \notin \mathbb{Q}$).

Theorem 1.10. [AS03a, Theorem 11.2.2] *Let $(a_n)_{n\geq 0}$ be a sequence with values in a finite set Ω. Assume that the sequence (a_n) is simultaneously k-automatic and*

l-automatic with respect to two multiplicatively independent integers $k, l \geq 2$. Then (a_n) is eventually periodic.

We will have no use for Cobham's theorem. We will, however, use the following much easier related result.

Theorem 1.11. [AS03a, Theorem 6.6.4] *Let $(a_n)_{n\geq 0}$ be a sequences with values in a finite set Ω. Let $k, l \geq 2$ be two multiplicatively* dependent *integers. Then the sequence (a_n) is k-automatic if and only if it is l-automatic.*

Let A denote a finite alphabet and let L and L' be languages, i.e., subsets of A^*. We denote by $LL' = \{wv \mid w \in L, v \in L'\}$ the concatenation of L and L'. For an integer $i \geq 0$, we denote by $L^i = L \cdots L$ the concatenation of i copies of L with the understanding that $L^0 = \{\epsilon\}$. The *Kleene closure* of L is $L^* = \bigcup_{i\geq 0} L^i$. A language L is *regular* if it can be obtained from the empty set and the letters of the alphabet using the operations of union, concatenation, and the Kleene closure.

Regular languages are intimately connected with automatic sequences via Kleene's theorem [Kle56] (see also [AS03a, Thm. 4.1.5]), which says that a language L over the alphabet Σ_k is regular if and only if the sequence $(a_n)_{n\geq 0}$ given by $a_n = [\![(n)_k \in L]\!]$ is k-automatic.

Dynamical systems. An (invertible, topological) dynamical system is given by a compact metrisable space X and a continuous homeomorphism $T\colon X \to X$. We say that X is minimal if for every point $x \in X$ the orbit $\{T^n x \mid n \in \mathbb{Z}\}$ is dense in X. (Equivalently, the only closed subsets $Y \subset X$ such that $T(Y) = Y$ are $Y = X$ or $Y = \emptyset$.) We say that X is totally minimal if the system (X, T^n) is minimal for all $n \geq 1$.

Let (X, T) be a dynamical system. We say that a Borel measure μ on X is invariant if for every Borel subset $A \subset X$ we have $\mu(T^{-1}(A)) = \mu(A)$. By the Krylov–Bogoliubov theorem (see, e.g., [EW11, Thm. 4.1]), each dynamical system has at least one invariant measure. We say that a dynamical system in *uniquely ergodic* if it has exactly one invariant measure.

If (X, T) is minimal, $x \in X$, and $U \subset X$ is open, then the set $\{n \in \mathbb{Z} \mid T^n x \in U\}$ is syndetic, i.e., has bounded gaps [Fur81, Thm. 1.15].

We will need the following standard consequence of the ergodic theorem [EW11, Thm 4.10], which we also note in [BK18b, Corollary 1.4]. (Below and elsewhere, ∂S denotes the boundary of the set S.)

Corollary 1.12. *Let (X, T) be a uniquely ergodic dynamical system with the invariant measure μ. Then for any $x \in X$ and any $S \subset X$ with $\mu(\partial S) = 0$, the set $E = \{n \in \mathbb{N}_0 \mid T^n x \in S\}$ has upper Banach density $\mu(S)$.*

In fact, in this case the limit superior in the definition of upper Banach density can be replaced by a limit.

The connection between generalised polynomials and dynamics of nilsystems has been intensely studied by Bergelson and Leibman in [BL07] (see also [Lei12]). Nilsystems are a widely studied class of dynamical systems of algebraic origin. Here, we only need several properties which these systems enjoy; in particular, we shall spare the reader the definition of a nilsystem. A good introduction to nilsystems may be found in the initial sections of [BL07].

A nilsystem (X, T) is minimal if and only if it is uniquely ergodic; the unique invariant measure μ_X has then full support. If (X, T) is minimal but not totally

minimal, then X splits into finitely many connected components $X_1, \dots, X_n$, each X_i is preserved by T^n, and each (X_i, T^n) is a totally minimal nilsystem.

As a special case of the aforementioned connection between nilsystems and generalised polynomials [BL07, Thm. A], we have the following result. (For more details, see also [BK18b].)

Theorem 1.13 (Bergelson–Leibman)**.** *Let $g\colon \mathbb{Z} \to \mathbb{R}$ be a generalised polynomial taking finitely many values $\{c_1, \dots, c_r\}$. Then there exists a minimal nilsystem (X, T) as well as a point $z \in X$ and a partition $X = S_1 \cup S_2 \cup \ldots \cup S_r$ such that $\mu_X(\partial S_j) = 0$ and*

$$g(n) = c_j \text{ if and only if } T^n z \in S_j$$

for each $1 \leq j \leq r$.

Remark 1.14. Let $g\colon \mathbb{Z} \to \mathbb{R}$ be a generalised polynomial taking finitely many values. Then there exists $a \in \mathbb{N}$ such that for any $b \in \mathbb{Z}$ the generalised polynomial $g_{a,b}(n) := g(an + b)$ has a representation as in Theorem 1.13 with (X, T) totally minimal.

2. Density 1 results

Polynomial sequences. Our first purpose in this section is to prove Theorem A. Recall that we aim to show that the sequence $n \mapsto \lfloor p(n) \rfloor$ is not regular if $p(x) \in \mathbb{R}[x]$ has at least one irrational coefficient other than the constant term. We will show more, namely that the sequence $n \mapsto \lfloor p(n) \rfloor \bmod m$ is not automatic for any $m \geq 2$. In fact, we will only need to work with the weaker property of weak periodicity, defined in the introduction.

Lemma 2.1. *Every automatic sequence is weakly periodic.*

Proof. Let f be a k-automatic sequence. Since the restriction of a k-automatic sequence to an arithmetic progression is again k-automatic [AS03a, Theorem 6.8.1], it will suffice to find $q \in \mathbb{N}$ and $r, r' \in \mathbb{N}_0$ with $r \neq r'$ such that $f(qn + r) = f(qn + r')$.

The k-kernel $\mathcal{N}_k(f)$ of f, consisting of the functions $f(k^t n + r)$ for $0 \leq r < k^t$, is finite. Pick t sufficiently large that $k^t > |\mathcal{N}_k(f)|$. By the pigeonhole principle, there exist $r \neq r'$ such that $f(k^t n + r) = f(k^t n + r')$. □

The proof of the following proposition is closely analogous to Furstenberg's proof [Fur61] of Weyl's equidistribution theorem [Wey16] (see also [EW11, Section 4.4.3]).

Proposition 2.2. *Let $p(x) \in \mathbb{R}[x]$ be a polynomial, and let $m \geq 2$ be an integer. Then the sequence $(\lfloor p(n) \rfloor \bmod m)_{n \geq 0}$ is weakly periodic if and only if it is periodic. This happens precisely when all non-constant coefficients of $p(x)$ are rational.*

Proof. If all coefficients of $p(x)$ are rational (except possibly for the constant term) then the sequence $(\lfloor p(n) \rfloor \bmod m)$ is easily seen to be periodic, hence weakly periodic.

Now suppose that at least one non-constant coefficient of $p(x)$ is irrational. Replacing $p(x)$ with $p(hx + r)$ for multiplicatively large h and $r = 0, 1, \dots, h - 1$, we may assume that the leading coefficient of $p(x)$ is irrational. We will prove marginally more than claimed, namely that for any $0 \leq l < m$ the sequence f given by

$$f(n) = [\![\lfloor p(n) \rfloor \equiv l \pmod{m}]\!] \tag{1}$$

fails to be weakly periodic. For a proof by contradiction, suppose this claim is false for some choice of l.

It will be convenient to expand $p(x)/m = \sum_{i=0}^{d} a_i \binom{x}{i}$, where $d = \deg p$, $a_i \in \mathbb{R}$, and $\binom{x}{i} = x(x-1)(x-2)\cdots(x-i+1)/i!$. Note that $a_d \in \mathbb{R} \setminus \mathbb{Q}$ and

$$f(n) = \left[\!\!\left[\frac{p(n)}{m} \bmod 1 \in \left[\frac{l}{m}, \frac{l+1}{m} \right) \right]\!\!\right]. \tag{2}$$

We will represent the sequence p dynamically. Let X be the d-dimensional torus $\mathbb{T}^d$ and define the self-map $T\colon X \to X$ by

$$(x_1, x_2, x_3, \ldots, x_d) \mapsto (x_1 + a_d, x_2 + x_1 + a_{d-1}, \ldots, x_d + x_{d-1} + a_1). \tag{3}$$

Put $a_j = 0$ for $j > d$. A direct computation shows that for $z = (0, 0, \ldots, 0, a_0)$ and $j = 1, \ldots, d$ we have

$$(T^n z)_j = z_j + \sum_{i \geq 1} a_{d-j+i} \binom{n}{i}, \tag{4}$$

and in particular $(T^n z)_d = p(n)/m$. Putting $A = \mathbb{T}^{d-1} \times \left[\frac{l}{m}, \frac{l+1}{m}\right)$, we thus find that

$$f(n) = [\![T^n z \in A]\!]. \tag{5}$$

Since f is weakly periodic, we may find q and $r \neq r'$ such that $f(qn + r) = f(qn + r')$.

The dynamical system (X, T) can be obtained as a sequence of iterated group extension over an irrational rotation, and hence is totally minimal (this follows easily from the results in, e.g., [EW11, Section 4.4.3]). In particular, for any point $y \in \operatorname{cl} A$ we may find a sequence $(n_i)_{i \geq 0}$ such that $T^{qn_i + r} z \to y$ and $T^{qn_i + r} z \in A$. It follows that the points $T^{qn_i + r'} z$ converge to $T^{r' - r} y$ and lie in A. Thus, $T^{r'-r}(\operatorname{cl} A) \subset \operatorname{cl} A$. In light of total minimality of T, this is only possible if $\operatorname{cl} A = X$ or $\operatorname{cl} A = \emptyset$ — but this is absurd. □

Corollary 2.3. *With the notation of Proposition 2.2, the sequence $n \mapsto \lfloor p(n) \rfloor \bmod m$ is automatic if and only if it is periodic, and if and only if all the non-constant coefficients of $p(x)$ are rational.*

Proof. Immediate from Proposition 2.2 and Lemma 2.1. □

Proof of Theorem A. Suppose first that all non-constant coefficients of $p(n)$ are rational, and fix an integer $k \geq 2$. Let $h \in \mathbb{N}$ be such that $hp(n)$ has integer coefficients, except possibly for the constant term. Then $f_1(n) = \lfloor hp(n) \rfloor$ is an integer-valued polynomial, hence is k-regular ($\mathcal{N}_k(f_1)$ is contained in the $(\deg p + 1)$-dimensional $\mathbb{Z}$-module consisting of integer-valued polynomials of degree $\leq \deg p$). Also, $f_2(n) = \lfloor hp(n) \rfloor - hf(n) = \lfloor h \{p(n)\} \rfloor$ is periodic, hence k-automatic, hence k-regular. It follows that $f(n) = \frac{1}{h}(f_1(n) - f_2(n))$ is regular.

Conversely, suppose that $f(n)$ is regular. Then by Theorem 1.7 for any choice of $m \geq 2$ the sequence $f(n) \bmod m$ is automatic. Now, it follows from Corollary 2.3 that all non-constant coefficients of $p(x)$ are rational. □

Generalised polynomials. Having dealt with the case of polynomial maps, we move on to a more general context. Our next goal is to prove Theorem B. We begin by abstracting and generalising some of the key steps from the proof of Theorem A.

Recall that a set of integers is *thick* if it contains arbitrarily long segments of consecutive integers, and *syndetic* if it has bounded gaps; every thick set intersects every syndetic set.

Lemma 2.4. *Let (X,T) be a totally minimal dynamical system. Let $A \subset X$ be a set which is neither empty nor dense and such that $\operatorname{cl} A = \operatorname{cl}\operatorname{int} A$. Let $z \in X$. Suppose that $f\colon \mathbb{N}_0 \to \{0,1\}$ is a sequence such that the set of n with $f(n) = [\![T^n z \in A]\!]$ is thick. Then f is not weakly periodic.*

Proof. Suppose for the sake of contradiction that f is weakly periodic. In particular there exist $q \in \mathbb{N}$, $r, r' \in \mathbb{N}_0$ with $r \neq r'$ such that $f(qn+r) = f(qn+r')$. Put $d = r' - r$.

We will show that $T^d(\operatorname{cl} A) \subset \operatorname{cl} A$. Since T is continuous and $\operatorname{cl}\operatorname{int} A = \operatorname{cl} A$, it will suffice to prove that $T^d(\operatorname{int} A) \subset \operatorname{cl} A$. Once this is accomplished, the contradiction follows immediately, because (X, T^d) is minimal, while $\operatorname{cl} A \neq \emptyset, X$.

Pick any $y \in \operatorname{int} A$ and an open neighbourhood V of $T^d y$; we aim to show that $V \cap A \neq \emptyset$. Put $U = T^{-d}V \cap \operatorname{int} A$, and consider the set S of those n for which $T^{qn+r}z \in U$. Since (X, T^q) is minimal and $U \neq \emptyset$, the set S is syndetic. Let R_0 be the set of those n for which $f(n) = [\![T^n z \in A]\!]$ and put $R = \{n \in \mathbb{N}_0 \mid qn + r \in R_0\}$ and $R' = \{n \in \mathbb{N}_0 \mid qn + r' \in R_0\}$.

Since R_0 is thick, so is $R \cap R'$. Since S is syndetic, $S \cap R \cap R'$ is non-empty. Pick any $n \in S \cap R \cap R'$ and put $x = T^{qn+r}z$. Since $n \in S$, we have $x \in U \subset A$, and so $T^d x \in V$. Since $n \in R$, we have $f(qn+r) = [\![x \in A]\!] = 1$, and hence also $f(qn+r') = 1$. Finally, since $n \in R'$, we have $1 = f(qn+r') = [\![T^d x \in A]\!]$, meaning that $T^d x \in V \cap A$. In particular, $V \cap A \neq \emptyset$, which was our goal. □

Remark 2.5. Some mild topological restrictions on the target set A are, of course, necessary in the above lemma. Note that any open, non-dense and non-empty subset of X will satisfy the stated assumptions.

The assumption that the map T is totally minimal is essential. Indeed, take X to be the Thue–Morse shift, i.e., the closed orbit under the shift map of the Thue–Morse sequence. Let

$$A = \{(a_n)_{n\in\mathbb{N}_0} \in X \mid a_{2k} = a_{2k+1} \text{ for some } k \in \mathbb{N}_0\},$$
$$B = \{(a_n)_{n\in\mathbb{N}_0} \in X \mid a_{2k+1} = a_{2k+2} \text{ for some } k \in \mathbb{N}_0\}.$$

Since the Thue–Morse sequence (t_n) has the property $t_{2n} \neq t_{2n+1}$ for all n and since the Thue–Morse word contains no cubes (i.e., no occurences of factors of the form www with $w \in \Sigma_k^*$, $w \neq \epsilon$), we see that $A \cap B \neq \emptyset$, $X = A \cup B$ and A and B are clopen. Let $z = (t_n) \in X$ be the Thue–Morse sequence. Then the function $f(n) = [\![T^n z \in A]\!]$ is periodic with period 2, and while X is minimal, it is not totally minimal.

The analogue of the representation of a polynomial sequence using a skew rotation on the torus in (5) is provided by the Bergelson–Leibman Theorem 1.13. We are now ready to state and prove the main result of this section, from which Theorem B easily follows.

Theorem 2.6. *Let $g\colon \mathbb{Z} \to \mathbb{R}$ be a generalised polynomial taking finitely many values, and let $f\colon \mathbb{N}_0 \to \mathbb{R}$ be a weakly periodic sequence which agrees with g on a thick set $R \subset \mathbb{N}_0$. Then there exists a set $Z \subset R$ with $d^*(Z) = 0$ such that the common restriction of f and g to $R \setminus Z$ is periodic.*

Proof. Let the minimal nilsystem (X, T), $z \in X$, and a partition $X = \bigcup_{j=1}^{r} S_j$ be as in Theorem 1.13, so that in particular

$$g(n) = \sum_{j=1}^{r} [\![T^n z \in S_j]\!]\, c_j. \tag{6}$$

If X is not totally minimal, then (as in Remark 1.14) we may find $a \in \mathbb{N}$ such that for any $b \in \mathbb{Z}$, $g'_{a,b}(n) = g(an + b)$ has a representation as in (6) on a totally minimal nilsystem. Clearly, $f'_{a,b}(n) = f(an + b)$ is weakly periodic and agrees with $g'_{a,b}(n)$ on the thick set $R'_{a,b} = \{n \mid an + b \in R\}$. Thus, it will suffice to prove the theorem under the additional assumption that (X, T) is totally minimal.

We may write

$$g(n) = \sum_{j=1}^{r} [\![T^n z \in \operatorname{int} S_j]\!]\, c_j + h(n), \tag{7}$$

where $h(n) = 0$ unless $T^n z \in \bigcup_{j=1}^{r} \partial S_j$. In particular (by Corollary 1.12), the set $Z \subset \mathbb{N}_0$ of n with $h(n) \neq 0$ has upper Banach density 0. Note that $R \setminus Z$ is then thick.

For $j \in \{1, \dots, r\}$, put $g'_j(n) = [\![T^n z \in \operatorname{int} S_j]\!]$ and $f'_j(n) = [\![f(n) = c_j]\!]$. Then $g'_j(n) = f'_j(n)$ for $n \in R \setminus Z$. By Lemma 2.4, this is only possible if for each j, the set $\operatorname{int} S_j$ is either empty or dense. Since $\mu_X(X \setminus \bigcup_{j=1}^{r} \operatorname{int} S_j) = 0$, there is i such that $\operatorname{int} S_i$ is dense, and $\operatorname{int} S_j = \emptyset$ for $j \neq i$. Denoting by $Z' \supset Z$ the set of $n \in R$ with $T^n z \in X \setminus \operatorname{int} S_i$ we have $d^*(Z') = 0$ and $f(n) = g(n) = c_i$ for $n \in R \setminus Z'$, as needed. □

Proof of Theorem B. This is a direct application of Theorem 2.6 with $f = g$ and $R = \mathbb{N}_0$ □

It is not a trivial matter to determine whether a given generalised polynomial is periodic away from a set of density 0, although it can be accomplished by the techniques in [BL07, Lei12]. In order to give explicit examples, we restrict ourselves to generalised polynomials of a specific form, which is somewhat more general than the one considered in Proposition 2.2.

Corollary 2.7. *Suppose that $q\colon \mathbb{Z} \to \mathbb{R}$ is a generalised polynomial with the property that $\lambda q(an) \bmod 1$ is equidistributed in $[0, 1)$ for any $\lambda \in \mathbb{Q} \setminus \{0\}$ and $a \in \mathbb{N}$, and let $m \geq 2$. Then the sequence $f(n) = \lfloor q(n) \rfloor \bmod m$ is not automatic.*

Proof. Suppose $f(n)$ were automatic. By Theorem B, there exist $a \in \mathbb{N}$ and $Z \subset \mathbb{N}_0$ with $d^*(Z) = 0$ such that $f(an)$ is constant for $n \in \mathbb{N}_0 \setminus Z$. Hence, there is some $0 \leq l < m$ such that $\frac{1}{m} q(an) \in \left[\frac{l}{m}, \frac{l+1}{m}\right)$ for $n \in \mathbb{N}_0 \setminus Z$, contradicting the equidistribution assumption. □

The uniform distribution of generalised polynomials has been extensively studied by Håland-Knutson [Hål93, Hål94, HK95], and later a very general theory was developed by Bergelson and Leibman [BL07, Lei12]. In view of the the results in

[Hål93], it is fair to say that a "generic" generalised polynomial $q(n)$ is equidistributed modulo 1. Hence, the assumptions on $q(n)$ in Corollary 2.7 are not overly restrictive.

To make the last remark precise, let us define the (multi)set of coefficients of a generalised polynomial q as follows. If $q(n) = \sum_j \alpha_j n^j$ is a polynomial, then the coefficients of $q(n)$ are the non-zero terms among the α_j. If $q(n) = r_1(n) + r_2(n)$ or $q(n) = r_1(n) \cdot r_2(n)$, then the coefficients of $q(n)$ are the union of the coefficients of $r_1(n)$ and $r_2(n)$. Finally, if $q(n) = p(n) \lfloor r(n) \rfloor^d$, then the coefficients of $q(n)$ are the union of the coefficients of $r(n)$ and the coefficients of $p(n)$. The set of coefficients will depend on the choice of a representation of the generalised polynomial at hand; we fix one such choice. We cite a slightly simplified version of the main theorem of [Hål93].

Theorem 2.8. *Suppose that $q(n)$ is a generalised polynomials, and all of the products of subsets of the coefficients of $q(n)$ are $\mathbb{Q}$-linearly independent. Then $q(n)$ is equidistributed modulo* 1*.*

As an example of an application, we conclude that $\lfloor \sqrt{2} n \lfloor \sqrt{3} n \rfloor \rfloor \bmod 10$ is not an automatic sequence.

3. Combinatorial structure of automatic sets

In this section, we begin the investigation of sparse sequences. Here, we call a sequence $f\colon \mathbb{N}_0 \to \{0,1\} \subset \mathbb{R}$ *sparse* if it is the characteristic function of a set of density 0 (if such a sequence comes from a generalised polynomial or is automatic, it also has upper Banach density 0, cf. [BK18b] and Lemma 4.8 below). Note that for such sparse sequences, Theorem B conveys no useful information. Conversely, to prove Conjecture A, it would suffice (in light of Theorem B) to verify it for sparse sequences; this observation will be made precise in the proof of Theorem C below.

Arid sets. To formulate our main result, it is convenient to introduce the following piece of terminology, inspired by Kedlaya [Ked06]. Such sets appear in the papers of Szilard–Yu–Zhang–Shallit [SYZS92], Gawrychowski–Krieger–Rampersad–Shallit [GKRS10], Derksen [Der07] and Adamczewski–Bell [AB08] (among many others) under different names (regular languages of polynomial growth/sparse/polyslender/bounded) or without any name. A closely related class of sets known as p-normal sets plays a significant rôle in the study of zero sets of linear recurrences in positive characteristic; see also [DM15, AB12]. Other related classes of sets include Saguaro sets of [AB08] and F-sets of [MS02]. Since we will use the notation simultaneously for languages and for the associated sets of integers, and since some of the existing terminology might be confusing in our context, we have decided to use a different term.

Definition 3.1 (Arid sets)**.** Let $k \geq 2$, $r \geq 0$ be integers. A *basic k-arid set* (of rank $\leq r$) is a set of the form

$$A = \left\{ v_0 w_1^{l_1} v_1 w_2^{l_2} \cdots w_r^{l_r} v_r \;\middle|\; l_1, \dots, l_r \in \mathbb{N}_0 \right\}, \tag{8}$$

where $v_0, \dots, v_r \in \Sigma_k^*$ and $w_1, \dots, w_r \in \Sigma_k^*$. A set $A \subset \Sigma_k^*$ is *k-arid* (of rank $\leq r$) if it is a finite union of *basic arid sets* (of rank $\leq r$). If k is clear from the context, we speak simply of (basic) arid sets.

We similarly define these notions for set of integers: A set $E \subset \mathbb{N}_0$ is k-arid (of rank $\leq r$) if it has the form $\{[u]_k \mid u \in A\}$ where $A \subset \Sigma_k^*$ is arid (of rank $\leq r$). A sequence $f\colon \mathbb{N}_0 \to \{0,1\}$ is arid if the set $\{n \in \mathbb{N}_0 \mid f(n) = 1\}$ is arid.

Using the Kleene star notation, the k-arid set A in (8) can be alternatively written as

$$A = v_0 w_1^* v_1 w_2^* \cdots w_r^* v_r.$$

In the following, we will not use this notation, and rather use the former notation which seems more appropriate for our context. It is possible to find a somewhat natural decomposition of an arid set into a union of basic arid sets; in particular, one can demand that the basic arid sets in the decomposition are pairwise disjoint. We elaborate on this in the next subsection.

Lemma 3.2. *Any k-arid sequence is k-automatic.*

Proof. It is clear that any k-arid set is given by a regular expression and hence it is k-automatic by Kleene's theorem. Alternatively, in this simple case one can construct the required automata by hand. □

Cobham [Cob72] proved that there is a gap in the growth rate of automatic sets.

Proposition 3.3. *Let $E \subset \mathbb{N}_0$ be a non-empty automatic set. Then exactly one of the following two conditions holds:*

(*i*) *There exists an integer $r \geq 0$ and a real number $c > 0$ such that*

$$\lim_{N\to\infty} \frac{|E \cap [N]|}{\log^r(N)} = c.$$

(*ii*) *There exists $\alpha > 0$ such that*

$$\liminf_{N\to\infty} \frac{|E \cap [N]|}{N^\alpha} = \infty.$$

Proof. This follows from [Cob72, Theorem 11 & 12] □

According to the theorem above, automatic sets have either poly-logarithmic or polynomial rate of growth. Szilard–Yu–Zhang–Shallit [SYZS92] showed that the class of automatic sets of poly-logarithmic growth coincides with the class of arid sets. To state a more precise version of this result, we recall that a state s in a k-automaton $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \{0,1\}, \tau)$ with output $\{0,1\}$ is called *accessible* if there exists $v \in \Sigma_k^*$ such that $\delta(s_0, v) = s$ and is called *coaccessible* is there exists $v \in \Sigma_k^*$ such that $\tau(\delta(s, v)) = 1$.

Proposition 3.4. *Let $E \subset \mathbb{N}_0$ be a k-automatic set and let $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \{0,1\}, \tau)$ be a k-automaton with output $\{0,1\}$ that produces E, in the sense that an integer n is in E if and only $\tau(\delta(s_0, n)) = 1$. Then the following conditions are equivalent:*

(i) The set E is arid.
(ii) There exists an integer r such that $|E \cap [N]| = O(\log^r(N))$.
(iii) There does not exist an accessible and coaccessible state $s \in S$ and $v_1, v_2 \in \Sigma_k^*$ such that $v_1 v_2 \neq v_2 v_1$ and $\delta(s, v_1) = \delta(s, v_2) = s$.

Moreover, if E is arid of rank r, then the limit $\lim_{N\to\infty} |E \cap [N]| / \log^r(N)$ exists and is finite.

Proof. This is essentially proved in [SYZS92]; our formulation is influenced by [BHS18, Lemmas 2.1–2.3] (for more details and related results see references therein). □

Remark. Some similar results are also implicit in [AB08, Lemma 6.7] and [Der07, Proposition 7.9]; see also [Ked06].

Remark 3.5. Let $a \geq 1$ be an integer. Then the notions of k-arid sets and k^a-arid sets coincide. This follows either from a direct argument or from Proposition 3.4. We will use this observation several times.

We will in fact need a slight improvement on the information on the rate of growth of arid sets from Proposition 3.4.

Lemma 3.6. *Let $E \subset \mathbb{N}_0$ be arid of rank (exactly) r. Then*

$$\max_{M \in \mathbb{N}_0} |E \cap [M, M+N)| = O(\log^r(N)).$$

Proof. It suffices to deal with basic arid sets given by

$$E = \Big\{ [v_0 w_1^{l_1} v_1 w_2^{l_2} \cdots w_r^{l_r} v_r]_k \;\Big|\; l_1, \ldots, l_r \in \mathbb{N}_0 \Big\}. \tag{9}$$

We begin with some standard reductions. Replacing w_i with suitably chosen powers, altering v_i accordingly, and passing to basic arid subsets, we may assume that all w_i have the same length a. Replacing k with k^a and using Remark 3.5 enables us to assume that $|w_i| = 1$ for each i. If r is minimal, we further know that if $w_i = w_{i+1}$ for some i, then v_i is not a power of w_i. Finally, we may assume that $N = k^L$ is a large power of k, and that $M = k^L M'$ is divisible by N.

Since an element of $E \cap [M, M+N)$ is uniquely determined by its final L digits, the bound $|E \cap [N]| \ll L^r$ follows immediately from counting the r-tuples $(l_1, \ldots, l_r)$ with $\sum_{i=1}^r l_i + \sum_{i=0}^r |v_i| \leq L$. □

We are now ready to state the main theorem of this section in a more convenient language.

Theorem 3.7. *Suppose that a sparse set $E \subset \mathbb{N}_0$ is simultaneously k-automatic and generalised polynomial. Then E is k-arid.*

For the proof of this result, we need to use the notion of IPS sets introduced in [BK18b].

Standard k-arid sets. The aim of this subsection is to show that any arid set can be written as a union of basic k-arid sets which are "well-behaved". One of the pleasant properties of such a decomposition is that basic arid sets in question are then pairwise disjoint. The existence of such a decomposition has been suggested to us by the anonymous referee, and we hope that it might be of some use. Nevertheless, we will not use the results of this subsection, and hence the reader might omit it without any loss of understanding of the forthcoming results.

Let $r \geq 0$ and $s \geq 1$ be integers. A *standard k-arid word of rank r and step s* is a word of the form

$$a(\mathbf{v}, \mathbf{w}, \mathbf{l}) = v_0 w_1^{l_1} v_1 \cdots w_r^{l_r} v_r,$$

where $\mathbf{v} = (v_i)_{i=0}^r$, $\mathbf{w} = (w_i)_{i=1}^r$, $\mathbf{l} = (l_i)_{i=1}^r$, and

(i) $|w_i| = s$ for each $1 \leq i \leq r$;
(ii) $|l_i| \geq 3$ for each $1 \leq i \leq r$;

(iii) v_{i-1} does not have a non-empty common suffix with w_i for each $1 \leq i \leq r$;
(iv) w_i is not a prefix of $v_i w_{i+1}$ (where $w_{r+1} := \epsilon$) for each $1 \leq i \leq r$;
(v) $|v_i| < 2s$ for each $0 \leq i \leq r$.

Let $\mathbf{m} = (m_i)_{i=1}^r$ be a sequence of integers satisfying the condition

(ii') $|m_i| \geq 3$ for each $1 \leq i \leq r$.

We denote by $A(\mathbf{v}, \mathbf{w}, \mathbf{m})$ the set of all standard k-arid words $a(\mathbf{v}, \mathbf{w}, \mathbf{l})$ of rank r and step s such that $l_i \geq m_i$ for each $1 \leq i \leq r$. We call $A(\mathbf{v}, \mathbf{w}, \mathbf{m})$ the *standard k-arid set of rank r, step s, and exponent $\mathbf{m}$*.

Lemma 3.8. *For a standard k-arid word of step s, the rank and the defining parameters are uniquely determined. More precisely, if a word*

$$u = a(\mathbf{v}, \mathbf{w}, \mathbf{l}) = a(\mathbf{v}', \mathbf{w}', \mathbf{l}')$$

has two representations as a standard k-arid word of rank r and r' (resp.) and step s, then $r = r'$ and $(\mathbf{v}, \mathbf{w}, \mathbf{l}) = (\mathbf{v}', \mathbf{w}', \mathbf{l}')$.

Proof. We proceed by induction on rank r. If $r = 0$, then by (v) $|u| < 2s$, whence by (i) and (ii) $r' = 0$ and the claim follows. Suppose now that $r \geq 1$ and (by symmetry) $r' \geq 1$, and assume without loss of generality that $|v_0'| \geq |v_0|$. Then $v_0' = v_0 x'$ for some x' with $|x'| < 2s$ (by (v)), and hence by (ii) we have

$$w_1 w_1 w_1 \cdots = x' w_1' w_1' w_1' \cdots .$$

Since by (i) $|w_1| = |w_1'| = s$, we see from this that either x' has a common suffix with w_1' or $x' = \epsilon$, the former being impossible by (iii). Hence $x' = \epsilon$, $v_0' = v_0$, and $w_1' = w_1$. We may assume without loss of generality that $l_1' \geq l_1$. If l_1' was strictly larger than l_1, then we would have

$$v_1 w_2^{l_2} \cdots w_r^{l_r} v_r = (w_1')^{l_1' - l_1} v_1' (w_2')^{l_2'} \cdots (w_r')^{l_r'} v_r',$$

from which we see that $w_1 = w_1'$ is a prefix of $v_1 w_2$, which is impossible by (iii). Hence, $l_1' = l_1$. It remains to remove the prefix $v_0 w_1^{l_1} = v_0' (w_1')^{l_1'}$ from u and apply the inductive assumption. □

Lemma 3.9. *Suppose that $B \subset \Sigma_k^*$ is k-arid of rank r. Then B is a union of standard k-arid sets of rank $\leq r$ and equal step s. Moreover, there exists $s_0 = s_0(B)$ such that s can be chosen to be any multiple of s_0.*

Proof. We proceed by induction on r. The claim is obvious for $r = 0$. Assume that $r \geq 1$ and that the claim holds for arid sets of rank $\leq r - 1$. Writing B as a finite union, we may assume without loss of generality that B is a basic k-arid set

$$B = \{v_0 w_1^{l_1} v_1 \cdots w_r^{l_r} v_r \mid l_i \geq 0\}.$$

Applying the induction hypothesis to the set

$$B' = \{v_0 w_1^{l_1} v_1 \cdots w_{r-1}^{l_{r-1}} v_{r-1} \mid l_i \geq 0\},$$

we may write B' as a finite union of standard k-arid sets of rank $\leq r - 1$ and equal step s (which may be chosen to be an arbitrary multiple of some integer). We choose s large enough (to be determined later) and divisible by $|w_r|$. Writing B as a finite union of sets corresponding to the terms in the decomposition of B', we may assume that B is of the form

$$B = \left\{ u w_r^{l_r} v_r \;\middle|\; u \in A(\mathbf{v}, \mathbf{w}, \mathbf{m}), l_r \geq 0 \right\}$$

for some standard k-arid set $A(\mathbf{v}, \mathbf{w}, \mathbf{m})$ of rank $r-1$ and step s.

The elements $\mathbf{v} = (v_i)_{i=0}^{r-1}$, $\mathbf{w} = (w_i)_{i=1}^{r-1}$ and $\mathbf{m} = (m_i)_{i=1}^{r-1}$ satisfy the conditions in the definition of a standard k-arid set, but this is not necessarily the case for w_r and v_r. We will now modify them to obtain this. Properties (i) and (ii') will be easy: we just replace w_r by an appropriate power, and split off the terms corresponding to small exponents. We will not do it yet, however, and proceed instead with the remaining properties.

We begin with (iii). Considering the common suffix of v_{r-1} and a large power of w_r, we may write

$$v_{r-1} = v'_{r-1} y w_r^p, \qquad w_r = xy$$

for some $p \geq 0$, $x \neq \epsilon$, and so that v'_{r-1} and x have no common suffix. Replacing v_{r-1} with v'_{r-1}, w_r with $w'_r = yx$, and v_r with $v'_r = yv_r$, we write

$$B = \{u(w'_r)^{l_r} v'_r \mid u \in A(\mathbf{v}', \mathbf{w}, \mathbf{m}), l_r \geq p\}, \tag{10}$$

where $\mathbf{v}' = (v_0, v_1, \ldots, v_{r-2}, v'_{r-1})$ (note that $A(\mathbf{v}', \mathbf{w}, \mathbf{m})$ is still a standard k-arid set, and that moreover $|v'_r| < |v_r| + |w_r|$).

Our next aim is property (iv). (This is vacuous if $r = 1$.) We know by the induction assumption that w_{r-1} is not a prefix of v_{r-1}, and hence a fortiori of v'_{r-1}. If w_{r-1} is not a prefix of $v'_{r-1}(w'_r)^{q'}$ for any $q' \geq 1$, we put $\mathbf{v}'' = \mathbf{v}'$, $\mathbf{w}'' = \mathbf{w}$, and $\mathbf{m}'' = \mathbf{m}$. Otherwise, write

$$w_{r-1} = v'_{r-1}(w'_r)^q x', \qquad w'_r = x'y'$$

for some $q \geq 0$ and $y' \neq \epsilon$. We first split off the elements in (10) corresponding to the values of l_r which are smaller than $q+1$. This is an arid set of rank $\leq r-1$, and hence satisfies the claim by induction. Putting $m''_{r-1} = m_{r-1} + 1$, $v''_{r-1} = \epsilon$, $w''_r = y'x'$, $v''_r = y'v'_r$, we can thus write the set B (up to the previously considered set of rank $\leq r-1$) as

$$B = \left\{u(w''_r)^{l_r} v''_r \;\middle|\; u \in A(\mathbf{v}'', \mathbf{w}'', \mathbf{m}''), l_r \geq \max(p-q-1, 0)\right\},$$

where $\mathbf{w}''$, $\mathbf{v}''$ and $\mathbf{m}''$ have the obvious meaning, the set $A(\mathbf{v}'', \mathbf{w}'', \mathbf{m}'')$ is a standard k-arid set, and $|v''_r| < |v_r| + 2|w_r|$. By construction, either w''_{r-1} is not a prefix of $v''_{r-1}(w''_r)^{q'}$ for any $q' \geq 0$ or w''_{r-1} coincides with a power of w''_r. In the latter case $v''_{r-1} = \epsilon$ and we may rewrite the set as an arid set of rank $\leq r-1$ for which the claim holds by induction. Hence assume the former condition holds.

To obtain the remaining properties, we write B as a union of $s/|w_r|$ sets depending on $l_r \bmod s/|w_r|$. Recall that $|w''_r| = |w_r|$ and replace B with a union of sets B_j by replacing w''_r with $w'''_r = (w''_r)^{s/|w_r|}$ and v''_r with $v'''_r = (w''_r)^j v''_r$ for $0 \leq j < s/|w_r|$. The sets B_j take the form

$$B_j = \left\{u(w'''_r)^{l_r} v'''_r \;\middle|\; u \in A(\mathbf{v}'', \mathbf{w}'', \mathbf{m}''), l_r \geq m'''_r\right\},$$

with $m'''_r = \lfloor \max(p-q-1, 0)|w_r|/s \rfloor$. Let $\mathbf{v}''', \mathbf{w}''', \mathbf{m}'''$ denote the vectors $\mathbf{v}'', \mathbf{w}'', \mathbf{m}''$ with v'''_r, w'''_r and m'''_r appended at the end. The parameters $\mathbf{v}''', \mathbf{w}''', \mathbf{m}'''$ clearly satisfy conditions (i) and (iii). Condition (iv) holds due to our assumption that w''_{r-1} is not a prefix of $v''_{r-1}(w''_r)^{q'}$ for any $q' \geq 0$. For condition (v), note that $|v'''_r| < |v_r| + 2|w_r| + (s - s/|w_r|) < 2s$ for s large enough. Finally, we may assume that $m''_r \geq 3$ by splitting off if necessary arid sets of rank $\leq r-1$ corresponding to the values of $l_r \leq 2$. This ends the proof of the inductive claim, and shows that s can be chosen to be an arbitrary multiple of some integer. $\square$

Corollary 3.10. *Any k-arid set can be written as a finite union of pairwise disjoint basic k-arid sets.*

Proof. Follows immediately from Lemmas 3.8 and 3.9. □

IPS sets and automatic sequences. The following notion generalises the classical notion of an IP set that is of importance in combinatorial number theory and ergodic theory (for origin of the term IP, which stands either for infinite-dimensional parallelepiped or idempotent, see, e.g., [BL16]). This notion is discussed in more detail in [BK18b] (in particular, an equivalent definition of IPS sets in terms of ultrafilters is given there).

Definition 3.11 (IP and IPS sets)**.** For a sequence $(n_i)_{i\in\mathbb{N}} \subset \mathbb{N}$, the corresponding set of *finite sums* is

$$\mathrm{FS}(n_i) = \{n_\alpha \mid \alpha \subset \mathbb{N},\ 0 < |\alpha| < \infty\}, \tag{11}$$

where $n_\alpha = \sum_{i\in\alpha} n_i$. Any set containing a set of the form $\mathrm{FS}(n_i; N_t)$ for some (n_i), (N_t) is called an IPS *set*.

For a sequence $(n_i)_{i\in\mathbb{N}} \subset \mathbb{N}$ and shifts $(N_t)_{t\geq 1} \subset \mathbb{N}_0$, the corresponding set of *shifted finite sums* is

$$\mathrm{FS}(n_i; N_t) = \{n_\alpha + N_t \mid t \in \mathbb{N},\ \alpha \subset \{1, 2, \dots, t\}, \alpha \neq \emptyset\}, \tag{12}$$

where again $n_\alpha = \sum_{i\in\alpha} n_i$. Any set containing a set of the form $\mathrm{FS}(n_i; N_t)$ for some (n_i), (N_t) is called an IPS *set*.

Example 3.12. Fix $k \geq 2$. Let $v_1, v_2 \in \Sigma_k^*$ be two distinct words with $|v_1| = |v_2| = l$, and let $u_0, u_1 \in \Sigma_k^*$ be arbitrary. Consider the set

$$E = \{[u_0 v_{j_1} v_{j_2} \cdots v_{j_t} u_1] \mid j_i \in \{1, 2\} \text{ for } 1 \leq i \leq t \text{ and } t \geq 0\}.$$

Then E is an IPS set. Indeed, $E = \mathrm{FS}(n_i; N_t)$, where $N_t = [u_0 v_1^t u_1]_k$ and $n_i = ([v_2]_k - [v_1]_k)k^{(i-1)l+|u_1|}$ (assuming, as we may, that $[v_2]_k > [v_1]_k$). If $[u_0]_k = [u_1]_k = [v_1]_k = 0$, then E is an IP set.

IPS sets occur in our work due to the following result.

Theorem 3.13. *Let $E \subset \mathbb{N}_0$ be an automatic set. Then either E is arid or it is* IPS.

Proof. Assume that $E \subset \mathbb{N}_0$ is automatic but not arid; we need to show that E is IPS. Let $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \{0, 1\}, \tau)$ be a k-automaton with output $\{0, 1\}$ which produces the characteristic sequence of E when reading digits starting from the most significant one and ignoring the initial zeros.

Since E is not arid, neither is the set $A = \{w \in \Sigma_k^* \mid \tau(\delta(s_0, w)) = 1\}$. Hence, by Proposition 3.4, there exists an accessible and coaccessible state $s \in S$ and $v_1, v_2 \in \Sigma_k^*$ such that $v_1 v_2 \neq v_2 v_1$ and $\delta(s, v_1) = \delta(s, v_2) = s$. Replacing v_1 and v_2 by their powers and interchanging them if necessary, we may assume that v_1 and v_2 are of equal length $l = |v_1| = |v_2|$ and $[v_1]_k < [v_2]_k$. Pick $u_0, u_1 \in \Sigma_k^*$ so that $s = \delta(s_0, u_0)$, and $\tau(\delta(s, u_1)) = 1$.

The set A contains all words of the form $w = u_0 v_{j_1} v_{j_2} \dots v_{j_t} u_1$, where $j_i \in \{1, 2\}$ and $t \in \mathbb{N}_0$. It follows that A is IPS (cf. Example 3.12). □

In order to prove Theorem 3.7, we need to recall one of the main results of [BK18b] (Theorem A), whose proof uses ergodic theory and the machinery of ultrafilters.

Theorem 3.14. *Let $E \subset \mathbb{Z}$ be a sparse generalised polynomial set. Then E is not* IPS*.*

Theorem 3.7 and Theorem C now follow quite easily.

Proof of Theorem 3.7. Let E be the set in Theorem 3.7. By Theorem 3.14, E is not IPS. Hence, by Theorem 3.13, it is arid. □

Proof of Theorem C. Suppose that $f\colon \mathbb{N}_0 \to \mathbb{R}$ is automatic and generalised polynomial. Let $b(n)$ be the periodic function such that the set $Z = \{n \in \mathbb{N}_0 \mid f(n) \neq b(n)\}$ has $d^*(Z) = 0$. (The existence of $b(n)$ is guaranteed by Theorem B.)

Note that Z is generalised polynomial and automatic (automaticity is clear; to see that Z is generalised polynomial, compose $f - b$ with a polynomial p such that $p(0) = 1$ and $p(x - y) = 0$ for $x \in f(\mathbb{N}_0)$, $y \in b(\mathbb{N}_0)$, $x \neq y$).

By Theorem 3.7, Z is arid. Hence, by Lemma 3.6 below, we have

$$|Z \cap [M, M + N)| = O\left(\log^r(N)\right)$$

for some $r \in \mathbb{N}_0$ as $N \to \infty$. □

If Conjecture A is true then there are no nontrivial examples of arid generalised polynomial sets (indeed, by Theorem D non-existence of such sets is precisely equivalent to Conjecture A; see also Proposition 5.3). However, there are examples of generalised polynomial sets which exhibit some properties reminiscent of arid sets. We have already mentioned in this context that the set of Fibonacci numbers is a generalised polynomial set, and in [BK18b, Theorems B & C] we have extended this to certain linear recurrences of order 2 and 3 as well as arbitrary sets whose size grows at a sublogarithmic rate.

It is important to note that in the statement of Theorem 3.13 it is not possible to replace IPS sets with IP sets or their translates (cf. Example 4.3). We discuss this question further in the next section.

4. Examples and properties of automatic sets

$\mathcal{B}$-free sets. In this subsection, we will discuss a simple class of examples of automatic sets, the $\mathcal{B}$-free sets, which will allow us to show that in the statement of Theorem 3.13 it is in general not possible to replace IPS sets with translates of IP sets (Example 4.3).

Example 4.1. Let $k \geq 2$, and let $\mathcal{B} \subset \Sigma_k^*$ be a finite set of 'prohibited' words of length $\leq t$. A word $u \in \Sigma_k^*$ is $\mathcal{B}$-free if u contains no $b \in \mathcal{B}$ as a factor. Accordingly, $n \in \mathbb{N}_0$ is $\mathcal{B}$-free if its base-k expansion $(n)_k$ is $\mathcal{B}$-free. Denote the set of $\mathcal{B}$-free integers by $F_{\mathcal{B}}$.

(*i*) The set $F_{\mathcal{B}}$ is k-automatic.
(*ii*) If $\mathcal{B} \neq \emptyset$, then $F_{\mathcal{B}}$ is sparse.
(*iii*) If $\sum_{b \in \mathcal{B}} k^{-|b|} \leq \frac{1}{16t}$, then $F_{\mathcal{B}}$ is not arid.
(*iv*) If each $b \in \mathcal{B}$ contains at least two non-zero digits, then $F_{\mathcal{B}}$ is IP.
(*v*) If some $b \in \mathcal{B}$ consists only of 0's, then $F_{\mathcal{B}} - m$ is not IP for any $m \in \mathbb{Z}$.

Proof. (*i*) It is not difficult to explicitly describe a k-automaton which computes the characteristic function of $F_{\mathcal{B}}$; alternatively, the claim follows immediately from Kleene's theorem.

(*ii*) We may assume that $\mathcal{B}$ consists of a single string of length t. Then the probability that a randomly chosen word of length m does not contain b is at most $(1-k^{-t})^{\lfloor m/t \rfloor}$. The claim easily follows from this.

(*iii*) We may assume $\mathcal{B} \neq \emptyset$. Construct an undirected graph $G = (V, E)$ (we allow G to have loops), where $V = \Sigma_k^t$, and $\{u, v\} \in E$ if uv and vu are both $\mathcal{B}$-free. If $u_1, u_2, \dots, u_r$ is a walk in G, then $u_1 u_2 \cdots u_r$ is $\mathcal{B}$-free. Assume that G contains a walk u_1, w, u_2 of length 2 with $u_1 \neq u_2$. With loss of generality, we may assume that $u_1 \neq 0^t$ (otherwise, switch u_1 and u_2). Then for any $i_1, \dots, i_r \in \{1, 2\}$ the word $v = u_1 w u_{i_1} w u_{i_2} w \cdots u_{i_r} w$ is $\mathcal{B}$-free. Hence, $[v]_k \in F_{\mathcal{B}}$ and we can see either directly or from Proposition 3.4 that $F_{\mathcal{B}}$ is not arid. Thus, it remains to check that G contains a length 2 walk with distinct endpoints; for the sake of contradiction suppose that this is not the case.

Since each vertex has at most one neighbour (including itself if $\{u, u\}$ is an edge), the graph is a disjoint union of paths of length 1, loops, and vertices, and hence $|E| \leq |V| = k^t$. On the other hand, given $b \in \mathcal{B}$, the number of pairs $(u, v) \in V^2$ such that b appears in uv or vu is $< 4tk^{2t-|b|}$, so

$$|E| \geq \frac{k^t(k^t+1)}{2} - 4tk^{2t} \sum_{b \in \mathcal{B}} k^{-|b|} > \frac{k^{2t}}{4} \geq k^t,$$

(note that the assumption implies that $k^t \geq 16$), which gives a contradiction.

(*iv*) Let $n_i = k^{it}$. Then $\mathrm{FS}(n_i) \subset F_{\mathcal{B}}$.

(*v*) Suppose that $F_{\mathcal{B}}$ contains $E+m$ for some IP set E and integer m. Replacing E with a smaller IP set if necessary, we may assume that $m > 0$. Since E is IP, for any $l \geq 0$ there exists $n \in E$ which is divisible by k^l. If l is large enough (it suffices that $l > t + \lfloor \log N / \log k \rfloor$) then $n + m$ is an element of $F_{\mathcal{B}}$ whose base-k expansion contains t consecutive zeros, contradicting the assumption on $\mathcal{B}$. □

Remark 4.2. A similar example was considered by Miller [Mil12], who gave sufficient conditions for $F_{\mathcal{B}}$ to be infinite.

Example 4.3. The set

$$F_{00} = \{n \in \mathbb{N}_0 \mid \text{the binary expansion of } n \text{ does not contain } 00\}$$

is 2-automatic, sparse, not arid, and does not contain a translate of an IP set.

Proof. We see that F_{00} is not arid by Proposition 3.4 or by a simple modification of the proof of 4.1.(iii). The remaining claims follow directly from Example 4.1. □

The following two examples can be verified similarly.

Example 4.4. The set

$$F_{11} = \{n \in \mathbb{N}_0 \mid \text{the binary expansion of } n \text{ does not contain } 11\}$$

is 2-automatic, sparse, not arid, and IP.

Example 4.5. The Baum–Sweet sequence ([BS76]) given by

$$f_{\mathrm{BS}}(n) = [\![\text{the binary expansion of } n \text{ does not contain } 10^l1 \text{ for an odd integer } l]\!].$$

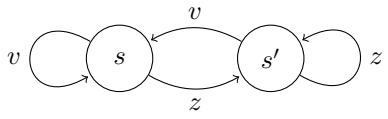
It takes the value 1 on a set which is 2-automatic, sparse, not arid, and IP.

Translates of IP sets. Even though in general non-arid automatic sets need not contain translates of IP sets, this is nevertheless the case under certain stronger assumptions on the set.

Proposition 4.6. *Let $E \subset \mathbb{N}_0$ be a k-automatic set. Assume that for every $w \in \Sigma_k^*$ there is an integer $n \in E$ such that w is a factor of $(n)_k$. Then the set $E - m = \{n - m \mid n \in E\}$ is* IP *for some $m \in \mathbb{N}_0$.*

Proof of Proposition 4.6. Let $\mathcal{A} = (S, \Sigma_k, \delta, s_0, \{0,1\}, \tau)$ be a k-automaton that produces the characteristic sequence of E by reading the digits of n *starting with the least significant one*, allowing for leading zeros. We will denote the word $0 \cdots 0 \in \Sigma_k^*$ with n zeros by 0^n. We begin by proving the following claim.

Claim. There exist states $s, s' \in S$ with $\tau(s) = 1$, an integer $l \in \mathbb{N}$, and a word $v \in \Sigma_k^l$ that is not a power of 0 such that for $z = 0^l$ we have $\delta(s, z) = s'$, $\delta(s, v) = s$, $\delta(s', z) = s'$, $\delta(s', v) = s$. This is portrayed below:



Proof of the claim. Let $n = |S|$ be the number of states in $\mathcal{A}$. We first show a weaker statement, namely that there is a state s with $\tau(s) = 1$ such that if $\tilde{s} = \delta(s, 0^n)$ denotes the state reached from s after reading n zeros, then we can return from $\tilde{s}$ to s along a path not consisting only of zeros, that is $\delta(\tilde{s}, \tilde{v}) = s$ for some $\tilde{v} \in \Sigma_k^*$ that is not a power of 0.

To prove this, we construct a word $w = w_1 w_2 \cdots w_{n^2}$ as follows. Enumerate all pairs in $S \times S$ as (s_i, s_i') for $1 \le i \le n^2$. In the first step, if $\tilde{s}_1$ is reachable from s_1, let w_1 describe any path between the two, so that $\delta(s_1, w_1) = s_1'$; otherwise, let $w_1 = \epsilon$. In general, if $w_1, \ldots, w_{i-1}$ have been defined, choose w_i so that $\delta(s_i, w_1 w_2 \cdots w_{i-1} w_i) = s_i'$ if possible (i.e., if s_i' is reachable from $\delta(s_i, w_1 w_2 \cdots w_{i-1})$), and $w_i = \epsilon$ otherwise.

By the assumption on the set E, there exists some $x, y \in \Sigma_k^*$ such that for $s = \delta(s_0, xwy)$ we have $\tau(s) = 1$. Applying the same assumption with $w1$ in place of w, we may ensure that y is not a power of 0. It remains to show that we can return from $\tilde{s} = \delta(s, 0^n)$ to s. For $0 \le i \le n^2$, let $r_i = \delta(s_0, x w_1 w_2 \cdots w_i)$ denote the intermediate states on the path from s_0 to s labelled xwy, in particular $r_0 = \delta(s_0, x)$. The construction of w is arranged so that for any i with $s_i = r_0$, we have $r_i = \delta(r_{i-1}, w_i) = \tilde{s}_i$, provided that s_i' is reachable from r_{i-1}.

Choose $1 \le j \le n^2$ such that $s_j = r_0$ and $s_j' = \tilde{s}$. Since s is reachable from r_{j-1} and $\tilde{s}$ is reachable from s, $\tilde{s}$ is reachable from r_{j-1}. Hence, the construction of w guarantees that $r_j = \delta(r_{j-1}, w_j) = \tilde{s}$. In particular, $\delta(\tilde{s}, \tilde{v}) = s$, where $\tilde{v} = w_{j+1} \ldots w_{n^2} y$. Note that $\tilde{v}$ is not a power of 0 since neither is y. This proves the weaker version of the claim.

To prove the stronger statement, note first that since S has only n states, there exist $0 \le i < j \le n$ such that $\delta(s, 0^i) = \delta(s, 0^j)$. Let $m > i$ be any integer divisible by $(j - i)$ and put $s' = \delta(s, 0^m)$. Since m is divisible by $(j - i)$, we have $s' = \delta(s', 0^m)$. Because $\tilde{s}$ is reachable from s' (actually, $\delta(s', 0^n) = \tilde{s}$), there is a word u (equal to $0^n v$, hence not a power of 0) such that $\delta(s', u) = s$. Take $v = (0^m u)^m$ and $l = m(|u| + m)$. The states s, s' and the word v (of length l)

satisfy all the required conditions, namely $\delta(s,0^l)=s'$, $\delta(s,v)=s$, $\delta(s',0^l)=s'$, $\delta(s',v)=s$, and $\tau(s)=1$. □

To finish the proof of Proposition 4.6, we may assume that all states in $\mathcal{A}$ are accessible. Choose states s and s' and words v and $z=0^l$ as in the statement of the claim. Let $u\in\Sigma_k^*$ be such that $\delta(s_0,u)=s$. For any word $w=uv_1v_2\cdots v_r$, where $v_i\in\{v,z\}$ for $1\le i<r$ and $v_r=v$, we have $\delta(s_0,w)=s$, whence $[w^{\mathrm{R}}]_k\in E$. It follows that E contains $\mathrm{FS}(n_i;N)$, where $N=[u^{\mathrm{R}}]_k$ and $n_i=k^{(i-1)l+|u|}[v^{\mathrm{R}}]_k$, $i\in\mathbb{N}$. □

Proposition 4.6 has the following amusing application which, however, does not require the full strength of Theorem 3.14. (Similar results can be shown in greater generality.)

Example 4.7. There exists a constant $c>0$ such that for any sequence $\varepsilon(n)$ which is a rational power of a generalised polynomial such that $\varepsilon(n)\ll n^{-c}$ as $n\to\infty$, the set

$$E=\left\{n\in\mathbb{N}_0 \;\middle|\; \left\|\sqrt{2}n\left\lfloor\sqrt{3}n\right\rfloor\right\|<\varepsilon(n)\right\}$$

is not automatic.

Proof. It is shown in [BK18b, Propositions 4.6 & 4.8] that E is generalised polynomial, E contains no translate of an IP set, and that $E\cap(a\mathbb{N}+b)\neq\emptyset$ for any $a\in\mathbb{N}$, $b\in\mathbb{N}_0$.

Suppose that E were k-automatic. Since E intersects nontrivially any arithmetic progression, it would satisfy the assumptions of Proposition 4.6, and thus would contain a translate of an IP set, contradicting the previously mentioned results. □

Densities of symbols. In this subsection, we prove a lemma on densities of occurrences of symbols in automatic sequences. As a corollary, we obtain the claim that sparse automatic sequences take non-zero value at a set of *Banach* density 0.

The density of symbols for an automatic sequence is often uniform. A set $E\subset\mathbb{N}_0$ has *uniform density* $d=d(E)$ if $|E\cap[M,M+N)|/N\to d$ as $N\to\infty$ uniformly in M. For an automaton $\mathcal{A}=(S,\Sigma_k,\delta,s_0,\Omega,\tau)$, a *strongly connected component* is an automaton $\mathcal{A}'=(S',\Sigma_k,\delta',s_0',\Omega,\tau')$, where $S'\subset S$ is non-empty, preserved under $\delta(\cdot,j)$ for all $j\in\Sigma_k$ and minimal with respect to these properties, $s_0'\in S'$, and δ', τ' are the restrictions of δ and τ to S', respectively.

Lemma 4.8. *Let $a\colon\mathbb{N}_0\to\Omega$ be a k-automatic sequence generated by an automaton $\mathcal{A}=(S,\Sigma_k,\delta,s_0,\Omega,\tau)$ reading input starting with the most significant digit, ignoring the initial zeros, and such that all the states are accessible. For $y\in\Omega$, let $\rho_y\ge0$. Then the following conditions are equivalent:*

- *(i) For any $y\in\Omega$, the set $\{n\in\mathbb{N}_0\mid a(n)=y\}$ has density ρ_y;*
- *(ii) For any $y\in\Omega$, the set $\{n\in\mathbb{N}_0\mid a(n)=y\}$ has uniform density ρ_y;*
- *(iii) For any sequence $\tilde{a}'\colon\Sigma_k^*\to\Omega$ produced by a strongly connected component $\mathcal{A}'$ of $\mathcal{A}$ and for any $y\in\Omega$ we have*

$$\left|\left\{u\in\Sigma_k^L \;\middle|\; \tilde{a}'(u)=y\right\}\right|/k^L\to\rho_y \text{ as } L\to\infty.$$

Proof. It is clear that (ii) implies (i). We will show that (i) implies (iii) and (iii) implies (ii). Throughout, it will be convenient to assume that $\Omega=\{0,1\}$, which we may do without loss of generality. We then write ρ for ρ_1.

Suppose that (i) holds, and take some $\tilde{a}'$ as in (iii). There is some $v \in \Sigma_k^*$ such that $\tilde{a}'(u) = a([vu]_k)$, whence

$$\frac{1}{k^L} \sum_{u \in \Sigma_k^L} \tilde{a}'(u) = \frac{1}{k^L} \sum_{n=[v]_k k^L}^{([v]_k+1)k^L-1} a(n) \to \rho \tag{13}$$

as $L \to \infty$.

Now suppose that (iii) holds. For any N, M and L, we have

$$\frac{1}{N} \sum_{n=M}^{M+N-1} a(n) = \frac{1}{N} \sum_{m=\lfloor M/k^L \rfloor}^{\lfloor (M+N)/k^L \rfloor} \sum_{n=0}^{k^L-1} a(mk^L + n) + O(k^L/N) \tag{14}$$

uniformly in M. For any $m \in \mathbb{N}_0$, consider the sequence $\tilde{a}'_m \colon \Sigma_k^* \to \Omega$ given by $\tilde{a}'_m(u) = a([(m)_k u]_k)$, so that if $u \in \Sigma_k^L$, then $a(mk^L + [u]_k) = \tilde{a}'_m(u)$. Note that $\tilde{a}'_m$ is produced by the automaton $\mathcal{A}'$ that is obtained from $\mathcal{A}$ by changing the initial state to $s'_0 = \delta(s_0, (m)_k)$.

If $\delta(s_0, (m)_k)$ lies in a strongly connected component of $\mathcal{A}$, then we may use (iii) to estimate the inner sums in (14):

$$\sum_{n=0}^{k^L-1} a(mk^L + n) = k^L \rho + o(k^L)$$

as $L \to \infty$ (where the error term is uniform with respect to m, since there are only finitely many possible sequences $\tilde{a}'_m$). It is an easy exercise to check that the set of $m \in \mathbb{N}_0$ such that $\delta(s_0, (m)_k)$ does not lie in a strongly connected component of $\mathcal{A}$ has upper Banach density 0. Estimating the inner sums in (14) corresponding to such m trivially by $O(k^L)$, and letting $L \to \infty$ slowly enough so that $k^L/N \to 0$, we conclude that

$$\frac{1}{N} \sum_{n=M}^{M+N-1} a(n) = \rho + o(1)$$

as $N \to \infty$ uniformly in M. Hence, (ii) holds. □

Linear recurrence sequences. We have already noted that the set of values of a linear recurrence sequence can be a generalised polynomial set. This is the case for the Fibonacci sequence; for more information, see [BK18b, Theorem B]. In contrast, we show that the set of values of a linear recurrence sequence is not automatic, except for trivial examples. In the proof, we apply Theorem 3.13.

Proposition 4.9. *Let $(a_m)_{m \geq 0}$ be an $\mathbb{N}$-valued sequence satisfying a linear recurrence of the form*

$$a_{m+n} = \sum_{i=1}^{n} c_i a_{m+n-i}, \quad m \geq 0 \tag{15}$$

with integer coefficients c_i. Suppose that for some k the set $E = \{a_m \mid m \in \mathbb{N}_0\}$ is k-automatic. Then E is a finite union of the following standard sets: linear progressions $\{am + b \mid m \in \mathbb{N}_0\}$ with $a, b \in \mathbb{N}_0$; exponential progressions $\{ak^{tm} + b \mid m \in \mathbb{N}_0\}$ with $a, b \in \mathbb{Q}$ and $t \in \mathbb{N}$; and finite sets.

Proof. We first claim that there exists a representation of E as a finite union

$$E = \bigcup_{i=1}^{K_{\text{lin}}} L_i \cup \bigcup_{i=1}^{K_{\text{poly}}} P_i \cup \bigcup_{i=1}^{K_{\text{exp}}} E_i \cup F, \tag{16}$$

where F is finite, $L_i = \{a_i m + b_i \mid m \in \mathbb{N}_0\}$ are arithmetic progressions, $P_i = \{p_i(m) \mid m \in \mathbb{N}_0\}$ are value sets of polynomials $p_i(x) \in \mathbb{Z}[x]$ with $\deg p_i \geq 2$, and E_i have exponential growth in the sense that $|E_i \cap [N]| \ll \log N$.

In order to prove this claim, we begin by noting that any restriction of (a_m) to an arithmetic progression $a_m^{(h,r)} = a_{hm+r}$ obeys some (minimal length) linear recurrence

$$a_{m+n'}^{(h,r)} = \sum_{i=1}^{n'} c_i^{(h,r)} a_{m+n'-i}^{(h,r)}, \quad m \geq 0$$

with $n' = n'(h,r) \leq n$. Moreover, there exists a choice of h such that each of that each $a_m^{(h,r)}$ is either identically zero or non-degenerate, in the sense that the associated characteristic polynomial $q^{(h,r)}(x) = x^{n'} - \sum_{i=1}^{n'} c_i^{(h,r)} x^{n'-i}$ has no pair of roots $\lambda, \mu \in \mathbb{C}$ such that λ/μ is a root of unity (see, e.g., [EvdPSW03, Theorem 1.2] for a much stronger statement). Hence, for the purpose of showing the existence of a representation of the form (16), we may assume that (a_m) is non-degenerate. Suppose also that n is minimal, and let $\lambda_1, \dots, \lambda_r$ be the roots of $q(x) = x^n - \sum_{i=1}^{n} c_i x^{n-i}$ with $|\lambda_1| \geq |\lambda_2| \geq \dots$. Note that either E is finite or $|\lambda_1| \geq 1$.

If $|\lambda_1| > 1$, then by the result of Evertse [Eve84] and van der Poorten and Schlickewei [vdPS91] (see [EvdPSW03, Theorem 2.3]), we have $a_m = |\lambda_1|^{m+o(m)}$ as $m \to \infty$. Hence, E has exponential growth, and we are done.

Otherwise, if $|\lambda_1| = 1$, then for all j we have $|\lambda_j| = 1$ or $\lambda_j = 0$. Kronecker's theorem [Kro57] (or a standard Galois theory argument) shows that if λ is an algebraic integer all of whose conjugates have absolute value 1, then λ is a root of unity. Using the general formula for the solution of a linear recurrence, we may write for sufficiently large m

$$a_m = \sum_{j=1}^{r} \lambda_j^m p_j(m) = \sum_{j=1}^{r} b_j(m) p_j(m),$$

where $p_j(x)$ are polynomials and $b_j(m)$ are periodic. Splitting $\mathbb{N}_0$ into arithmetic progressions where $b_j(m)$ are constant, we conclude that E is a finite union of value sets of polynomials. This again produces a representation of the form (16).

Such a representation is not unique. Splitting P_i into a finite number of subprogressions and discarding those which are redundant, we may assume that $P_i \cap L_j = \emptyset$ for any i, j. Likewise, we may assume that $E_i \cap L_j = F \cap L_j = \emptyset$ for any i, j. Fix one such representation subject to these restrictions. The set

$$E' = \bigcup_{i=1}^{K_{\text{poly}}} P_i \cup \bigcup_{i=1}^{K_{\text{exp}}} E_i \cup F = E \setminus \bigcup_{i=1}^{K_{\text{lin}}} L_i$$

is again k-automatic; it will suffice to show that E' is a union of the standard sets mentioned above.

We claim that $K_{\text{poly}} = 0$, i.e., the representation of E uses no polynomial progressions of degree ≥ 2. Suppose for the sake of contradiction that $P = \{p(m) \mid m \in \mathbb{N}_0\}$ appears in one of the sets P_i, and write $p(m) = \sum_{i=0}^{d} c_i m^i$,

where $c_i \in \mathbb{Z}$. Replacing $p(m)$ with $p(m+r)$ for a suitably chosen $r \in \mathbb{N}_0$, we may assume that $c_i > 0$ for $0 \leq i \leq d$. For sufficiently large t, we have $p(k^t) = [u_d 0^{t-t_0} u_{d-1} 0^{t-t_0} u_{d-2} \cdots u_1 0^{t-t_0} u_0]_k$, where t_0 is a constant and u_i is the base-k expansion of c_i, padded by 0's so as to have $|u_i| = t_0$. Since $p(k^t) \in E'$, from the pumping lemma 1.9 it follows that there is $l \in \mathbb{N}$ such that for any $s_1, \ldots, s_d \in \mathbb{N}$ it holds that

$$n(s_1, \ldots, s_d) := [u_d 0^{ls_d} u_{d-1} 0^{ls_{d-1}} \cdots u_1 0^{ls_1} u_0]_k \in E'.$$

For sufficiently large S and a small absolute constant δ to be determined later, consider the set

$$Q(S) = \{n(s_1, \ldots, s_d) \mid s_i \in \mathbb{N}, s_1 + \ldots + s_d = S, s_d \geq (1-\delta)S\},$$

and put $N(S) := n(1, \ldots, 1, S-d+1) = \min Q(S)$ (for large S). Note that $N(S) = k^{lS+O(1)}$ and that $\max Q(S) = N(S) + O(N(S)^\delta)$. For a fixed T_0 and $T \to \infty$, we shall consider the cardinality of the set $Q(T_0, T) = \bigcup_{T_0 \leq S \leq T} Q(S)$. By an elementary counting argument, we find

$$|Q(T_0, T)| \gg T^d \gg T^2. \tag{17}$$

To obtain an upper bound, we separately estimate $|Q(S) \cap P_i|$ and $|Q(T_0, T) \cap E_j|$ for each i, j.

Suppose that $n, n' \in Q(S) \cap P_i$ with $n' > n$, so in particular $n = p_i(m)$ and $n' = p_i(m')$ for some $m, m' \gg N(S)^{1/\deg p_i}$. We then have the chain of inequalities:

$$N(S)^\delta \gg n' - n = p_i(m') - p_i(m) \geq \min_{x \in [m, m']} |p_i'(x)| \gg N(S)^{\frac{\deg p_i - 1}{\deg p_i}},$$

which is a contradiction for sufficiently large S, provided that $\delta < \frac{\deg p_i - 1}{\deg p_i}$ (which will hold if we put $\delta = \frac{1}{3}$). Thus, $|Q(S) \cap P_i| \leq 1$.

As for $Q(T_0, T) \cap E_j$, from the bounds on growth of E_j we immediately have

$$\left| \bigcup_{T_0 \leq S \leq T} Q(T_0, T) \cap E_i \right| \ll |E_i \cap [2N(T)]| \ll T. \tag{18}$$

In total, using (17) and (18) we find that

$$|Q(T_0, T)| \leq \sum_{S=T_0}^{T} \sum_{i=1}^{K_{\mathrm{poly}}} |Q(S) \cap P_i| + \sum_{i=1}^{K_{\mathrm{exp}}} |Q(T_0, T) \cap E_i| + O(1) \ll T, \tag{19}$$

contradicting the previously obtained bound $|Q(T_0, T)| \gg T^2$. It follows that indeed $K_{\mathrm{poly}} = 0$.

Since E' contains no polynomial or linear progressions, we have $|E' \cap [N]| \ll \log N$. It follows from Proposition 3.4 that E' must be k-arid of rank 1. Since all basic arid sets of rank 1 are of the form described in the statement of the theorem, we are done. □

5. Proof of Theorem D

In this section, we derive Theorem D from Theorem C. Our argument is purely combinatorial and can be entirely phrased in terms of finite automata with no further recourse to dynamics.

Proposition 5.1. *Let $A \subset \Sigma_k^*$ be an infinite arid set. Then there exists $v \in \Sigma_k^*$ such that $A \cap v\Sigma_k^*$ takes the form*

$$A \cap v\Sigma_k^* = \bigcup_{i=1}^{p} \{vw^l u_i \mid l \in \mathbb{N}_0\},$$

where $p \geq 1$, $v, w, u_i \in \Sigma_k^*$ and $w \neq \epsilon$. In particular, $A \cap v\Sigma_k^*$ is arid of rank 1.

Likewise, there exists $\tilde{v} \in \Sigma_k^*$ such that $A \cap \Sigma_k^*\tilde{u}$ takes the form

$$A \cap \Sigma_k^*\tilde{u} = \bigcup_{i=1}^{\tilde{p}} \{\tilde{v}_i(\tilde{w})^l \tilde{u} \mid l \in \mathbb{N}_0\},$$

where $\tilde{p} \geq 1$, $\tilde{v}_i, \tilde{w}, \tilde{u} \in \Sigma_k^*$ and $\tilde{w} \neq \epsilon$.

Proof. Since the notion of an arid set is preserved under the reversal operation, it is sufficient to prove the former statement. For $B \subset \Sigma_k^*$ and $v \in \Sigma_k^*$, put $v^{-1}B = \{u \in \Sigma_k^* \mid vu \in B\}$. If B is arid of rank $\leq r$, then so is $v^{-1}B$.

Claim. Let $B \subset \Sigma_k^*$ be arid of rank r, and let $x_1, x_2, y \in \Sigma_k^*$ be such that $a = |y| = |x_2|$ and $y \neq x_2$. Then for sufficiently large m (depending on B, x_1, x_2, y), $(x_1 y^m x_2)^{-1}B$ is arid of rank $\leq (r-1)$.

Proof. Replacing B with $x_1^{-1}B$, we may assume that $x_1 = \epsilon$.

Let $a = |y| = |x_2|$. In analogy with Remark 3.5, note that there is a natural way to identify $\Sigma_{k^a}^*$ with a subset of Σ_k^*, and any arid set $B \subset \Sigma_k^*$ is a finite union of translates $B_i v_i$ with $v_i \in \Sigma_k^*$ of arid sets $B_i \subset \Sigma_{k^a}^*$. Hence, it will suffice to show that if $B \subset \Sigma_{k^a}^*$ is arid of rank r, then for sufficiently large m, $B \cap y^m x_2 \Sigma_k^*$ is arid of rank $\leq (r-1)$. We may now replace k with k^a and assume that $|y| = |x_2| = 1$.

It will suffice to prove the claim for B of the form

$$B = \left\{ v_0 w_1^{l_1} v_1 w_2^{l_2} \cdots w_r^{l_r} v_r \;\middle|\; l_1, \ldots, l_r \in \mathbb{N} \right\}$$

where $w_i \neq \epsilon$ for all i (note that l_i here are required to be strictly positive; any arid set of rank r is a union of such sets and an arid set of rank $\leq (r-1)$). Now, if $m > |v_0 w_1|$ then either $B \cap y^m x_2 \Sigma_k^* = \emptyset$ (in which case we are trivially done) or $B \cap y^m x_2 \Sigma_k^* \neq \emptyset$ and both v_0 and w_1 is a power of y. In the latter case, we further conclude that x_2 appears in $v_1 w_2$ (else B would have rank $\leq (r-1)$), which is necessarily of the form $y^b x_2 v_1'$ with $b \in \mathbb{N}_0$. Hence

$$(y^m x_2)^{-1} B = \left\{ v_1' w_2^{l_2 - 1} v_2 w_3^{l_3} \cdots w_r^{l_r} v_r \;\middle|\; l_2, \ldots, l_r \in \mathbb{N} \right\}$$

is arid of rank $\leq (r-1)$. □

The proof of the proposition is now a simple induction on the rank r of A. Since A is infinite, we have $r \geq 1$.

If $r = 1$, then A takes the form $\bigcup_{i=1}^{r} \left\{ v_i w_i^l u_i \;\middle|\; l \in \mathbb{N}_0 \right\}$, where $w_i \neq \epsilon$ for at least one i, say $i = 1$. Then $A \cap v\Sigma_k^*$ takes the required form for $v = v_1 w_1^m$ for m large enough.

If $r > 1$, then we may find a rank 2 basic arid set

$$B = \left\{ v_0 w_1^{l_1} v_1 w_2^{l_2} v_2 \;\middle|\; l_1, l_2 \in \mathbb{N}_0 \right\}$$

contained in A. Without loss of generality, we may assume that $|w_1| = |w_2| > |v_1|$. Apply the above Claim with $x_1 = v_0$, $y = w_1^{l_1}$ and x_2 equal to the first $|y|$ symbols

of $v_1 w_2^{l_2}$, where $l_2 \geq l_1 \geq 2$. Note that $y \neq x_2$, because otherwise by an elementary computation one could show that the rank of B is 1. Then for m large enough $A' = (x_1 y^m x_2)^{-1} A$ is arid of rank $\leq (r-1)$ and infinite. By the inductive assumption, there exists $v' \in \Sigma_k^*$ such that $A' \cap v' \Sigma_k^*$ takes the required form. It remains to take $v = x_1 y^m x_2 v'$. □

Corollary 5.2. *Let E be an infinite k-arid set. Then there exist integers $n \geq 1$, $r \geq 0$, $p \geq 1$, and words $v_1, \ldots, v_p, w, u \in \Sigma_k^*$, $w \neq \epsilon$ such that*

$$E \cap (n\mathbb{Z} + r) = \bigcup_{i=1}^{p} \{[v_i w^l u]_k \mid l \in \mathbb{N}_0\}.$$

Proof. Follows immediately from the second part of Proposition 5.1. □

Proposition 5.3. *If the set $\{k^l \mid l \geq 0\}$ is not generalised polynomial, then neither is any infinite k-arid set.*

Proof. Assume we know that $P = \{k^l \mid l \geq 0\}$ is not generalised polynomial. Then neither is any set of the form $P_t = \{k^{tl} \mid l \geq 0\}$ for $t \geq 1$ since $P = \bigcup_{j=0}^{t-1} k^j P_t$.

Suppose that there exists an infinite k-arid set which is generalised polynomial. Since the class of generalised polynomial sets contains all arithmetic progressions and is closed under finite intersections, Corollary 5.2 allows us to assume that

$$E = \bigcup_{i=1}^{p} \{[v_i w^l u]_k \mid l \geq 0\}$$

for some $p \geq 1, v_1, \ldots, v_p, w, u \in \Sigma_k^*$, $w \neq \epsilon$. Let $s = |u|, t = |w|$ and note that

$$[v_i w^l u]_k = [u]_k + k^s [w]_k \frac{k^{tl} - 1}{k^t - 1} + [v_i]_k k^{tl+s}.$$

Let g be a generalised polynomial such that $E = \{n \in \mathbb{N}_0 \mid g(n) = 0\}$ and assume further that g is a restriction of a generalised polynomial of a real variable that has no further zeros in $\mathbb{R}_{>0} \setminus \mathbb{N}$. (To this end, replace $g(n)$ by $g(n)^2 + \|n\|^2$.) Then an easy computation shows that the polynomial

$$h(n) = g\left(k^s \frac{n - [w]_k}{k^t - 1} + [u]_k\right)$$

has as its zero set

$$B = \{n \in \mathbb{N} \mid h(n) = 0\} = \bigcup_{i=1}^{p} \{b_i k^{tl} \mid l \geq 0\}$$

where $b_i = [w]_k + (k^t - 1)[v_i]_k$, $i = 1, \ldots, p$.

The set $C = \{n \in \mathbb{N}_0 \mid b_1 n \in B\}$ is also generalised polynomial and it has the form

$$C = \bigcup_{i=1}^{p} \{c_i k^{tl} \mid l \geq 0\}$$

with $c_1 = 1$ and $c_i = b_i k^{tl_i}/b_1$, where $l_i \geq 0$ is the smallest integer such that b_1 divides $b_i k^{tl_i}$. (If there is no such integer, the corresponding term is not present.)

Let $m \geq 1$ be such that $c_i < k^{tm}$ for $i = 1, \dots, p$. Replacing the set $\{c_i k^{tl} \mid l \geq 0\}$ by the union

$$\{c_i k^{tl} \mid l \geq 0\} = \bigcup_{j=0}^{m-1} \{c_i k^{tj} k^{mtl} \mid l \geq 0\}$$

and replacing k by k^{mt}, we may assume that

$$C = \bigcup_{i=1}^{p} \{c_i k^l \mid l \geq 0\}$$

with $c_1 = 1$ and $1 \leq c_i < k^2$.

Consider the set $D = \{n \in C \mid n \equiv 1 \pmod{k^2 - 1}\}$. The set D is generalised polynomial and an integer $c_i k^l \in C$ can be an element of D only if $c_i \equiv 1 \pmod{k^2 - 1}$ or $c_i \equiv k \pmod{k^2 - 1}$. Since $1 \leq c_i \leq k^2 - 1$, this gives $c_i = 1$ or $c_i = k$ and whether the latter possibility is realised or not, we have $D = \{k^{2l} \mid l \geq 0\}$. This is a contradiction with our remark that no set of the form $P_t = \{k^{tl} \mid l \geq 0\}$, $t \geq 1$, is generalised polynomial (note that during the proof we have replaced k by its power). □

We are now ready to finish the proof of Theorem D.

Proof of Theorem D. The two statements in Theorem D are of course mutually exclusive. Now assume that there exists a sequence (a_n) which is k-automatic, generalised polynomial, and not ultimately periodic. By Theorem C, it nevertheless coincides with a periodic sequence (b_n) except at a set of density zero. Consider the set $C = \{n \in \mathbb{N}_0 \mid a_n \neq b_n\}$. This set is k-automatic, generalised polynomial, sparse, and infinite. By Theorem 3.7, C is then arid and hence by Proposition 5.3 the set $\{k^l \mid l \geq 0\}$ is generalised polynomial as well. □

6. Concluding remarks

In this section, we gather some remarks and questions which arise naturally. The question with which we begin was already alluded to in the introduction and in [BK18b]. As previously discussed, its resolution would suffice to decide if Conjecture A is true.

Question 1. Let $k \geq 2$ be an integer. Is the set $\left\{k^i \;\middle|\; i \geq 0\right\}$ generalised polynomial?

We find this question exceptionally pertinent because of its simple formulation.

Morphic words. The class of morphic words is a natural extension of the class of automatic sequences. Let Ω be a finite set. Any morphism φ of the monoid Ω^* extends naturally to $\Omega^{\mathbb{N}_0}$. A word $w \in \Omega^{\mathbb{N}_0}$ (which we identify with a function $\mathbb{N}_0 \to \Omega$) is a *pure morphic word* if it is a fixed point of a non-trivial morphism of Ω^*. A morphic word is the image $\pi \circ w \colon \mathbb{N}_0 \to \Omega'$ of a pure morphic word w under a coding $\pi \colon \Omega \to \Omega'$ (i.e., any set-theoretic map, not necessarily injective). Morphic words are connected with automatic sequences via the fact that k-automatic sequences are precisely the morphic words coming from k-uniform morphisms. Here, a morphism $\varphi \colon \Omega^* \to \Omega^*$ is k-uniform if $|\varphi(u)| = k$ for all $u \in \Omega$.

We have already encountered possibly the most famous example of a non-uniform morphic word, the Fibonacci word. Recall from the introduction that the Fibonacci word w_{Fib} was defined as the limit of the words $w_0 := 0$, $w_1 := 01$, and $w_{i+2} :=$

$w_{i+1}w_i$. Directly from this definition, it is easy to see that w_{Fib} is fixed by the morphism $\varphi\colon \Omega^{\mathbb{N}_0} \to \Omega^{\mathbb{N}_0}$ given by $\varphi(0) = 01$ and $\varphi(1) = 0$.

Recall also that w_{Fib} is a Sturmian word. Here, a *Sturmian word* is one of the form $f(n) = \lfloor \alpha(n+1) + \rho \rfloor - \lfloor \alpha n + \rho \rfloor - \lfloor \alpha \rfloor$, where $\alpha, \rho \in \mathbb{R}$ and $\alpha \notin \mathbb{Q}$ (for w_{Fib} we may take $\alpha = \rho = 2 - \varphi$). Some (but not all) of these sequences give rise to morphic words; see [BS93] for details (cf. also [Yas99, Fag06, BEIR07]).

In analogy with Conjecture A, one could ask about a classification of all morphic words which are given by generalised polynomials. We believe that examples such as the Fibonacci word are essentially the only possible ones.

Question 2. Assume that a sequence $f\colon \mathbb{N}_0 \to \Omega \subset \mathbb{R}$ is both a morphic word and a generalised polynomial. Is it true that f is a linear combination of a number of Sturmian morphic words and an eventually periodic sequence?

Regular sequences. We finish by presenting a generalisation of Conjecture A to regular sequences. We call a function $f\colon \mathbb{N}_0 \to \mathbb{Z}$ a quasi-polynomial if there exists an integer $m \geq 1$ such that the sequences f_j given by $f_j(n) = f(mn + j)$, $0 \leq j \leq m-1$, are polynomials in n. We say that a function $f\colon \mathbb{N}_0 \to \mathbb{Z}$ is ultimately a quasi-polynomial if it coincides with a quasi-polynomial except on a finite set.

Question 3. Assume that a sequence $f\colon \mathbb{N}_0 \to \mathbb{Z}$ is both regular and generalised polynomial. Is it then true that f is ultimately a quasi-polynomial?

If f takes only finitely many values, then all the polynomials inducing f_j are necessarily constant, and so in this case the question coincides with Conjecture A.

References

[AB08] B. Adamczewski and J. Bell. Function fields in positive characteristic: expansions and Cobham's theorem. *J. Algebra*, 319(6):2337–2350, 2008.

[AB12] B. Adamczewski and J. P. Bell. On vanishing coefficients of algebraic power series over fields of positive characteristic. *Invent. Math.*, 187(2):343–393, 2012.

[AS92] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. *Theoret. Comput. Sci.*, 98(2):163–197, 1992.

[AS03a] J.-P. Allouche and J. Shallit. *Automatic sequences*. Cambridge University Press, Cambridge, 2003.

[AS03b] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. II. *Theoret. Comput. Sci.*, 307(1):3–29, 2003.

[BS76] L. E. Baum and M. M. Sweet. Continued fractions of algebraic power series in characteristic 2. *Ann. of Math. (2)*, 103(3):593–610, 1976.

[BHS18] J. Bell, K. Hare, and J. Shallit. When is an automatic set an additive basis? *Proc. Amer. Math. Soc. Ser. B*, 5:50–63, 2018.

[Bel07] J. P. Bell. p-adic valuations and k-regular sequences. *Discrete Math.*, 307(23):3070–3075, 2007.

[BL07] V. Bergelson and A. Leibman. Distribution of values of bounded generalized polynomials. *Acta Math.*, 198(2):155–230, 2007.

[BL16] V. Bergelson and A. Leibman. Sets of large values of correlation functions for polynomial cubic configurations. *Ergodic Theory and Dynamical Systems*, pages 1–24, 2016.

[Ber81] J. Berstel. Mots de Fibonacci. In *Séminaire d'Informatique Théorique*, pages 57–78, Paris, 1980–1981.

[Ber85] J. Berstel. Fibonacci words, a survey. In G. Rozenberg and A. Salomaa, editors, *The Book of L*, pages 11–25. Springer-Verlag, 1985.

[BS93] J. Berstel and P. Séébold. A characterization of Sturmian morphisms. In *Mathematical foundations of computer science 1993 (Gdańsk, 1993)*, volume 711 of *Lecture Notes in Comput. Sci.*, pages 281–290. Springer, Berlin, 1993.

[BEIR07] V. Berthé, H. Ei, S. Ito, and H. Rao. On substitution invariant Sturmian words: an application of Rauzy fractals. *Theor. Inform. Appl.*, 41(3):329–349, 2007.

[BK18a] J. Byszewski and J. Konieczny. Factors of generalised polynomials and automatic sequences. *Indag. Math. (N.S.)*, 29(3):981–985, 2018.

[BK18b] J. Byszewski and J. Konieczny. Sparse generalised polynomials. *Trans. Amer. Math. Soc.*, 370(11):8081–8109, 2018.

[Cho59] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.

[Cob69] A. Cobham. On the base-dependence of sets of numbers recognizable by finite automata. *Math. Systems Theory*, 3:186–192, 1969.

[Cob72] A. Cobham. Uniform tag sequences. *Math. Systems Theory*, 6:164–192, 1972.

[Der07] H. Derksen. A Skolem-Mahler-Lech theorem in positive characteristic and finite automata. *Invent. Math.*, 168(1):175–224, 2007.

[DM15] H. Derksen and D. Masser. Linear equations over multiplicative groups, recurrences, and mixing II. *Indag. Math. (N.S.)*, 26(1):113–136, 2015.

[Eil74] S. Eilenberg. *Automata, languages, and machines. Vol. A*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York, 1974. Pure and Applied Mathematics, Vol. 58.

[EW11] M. Einsiedler and T. Ward. *Ergodic theory with a view towards number theory*, volume 259 of *Graduate Texts in Mathematics*. Springer-Verlag London, Ltd., London, 2011.

[EvdPSW03] G. Everest, A. van der Poorten, I. Shparlinski, and T. Ward. *Recurrence sequences*, volume 104 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2003.

[Eve84] J.-H. Evertse. On sums of S-units and linear recurrences. *Compositio Math.*, 53(2):225–244, 1984.

[Fag06] I. Fagnot. A little more about morphic Sturmian words. *Theor. Inform. Appl.*, 40(3):511–518, 2006.

[FK18] A. Fan and J. Konieczny. On uniformity of q-multiplicative sequences. 2018. Preprint. arXiv:1806.04267 [math.NT].

[Fur61] H. Furstenberg. Strict ergodicity and transformation of the torus. *Amer. J. Math.*, 83:573–601, 1961.

[Fur81] H. Furstenberg. *Recurrence in ergodic theory and combinatorial number theory*. Princeton University Press, Princeton, N.J., 1981. M. B. Porter Lectures.

[GKRS10] P. Gawrychowski, D. Krieger, N. Rampersad, and J. Shallit. Finding the growth rate of a regular or context-free language in polynomial time. *Internat. J. Found. Comput. Sci.*, 21(4):597–618, 2010.

[Gow01] W. T. Gowers. A new proof of Szemerédi's theorem. *Geom. Funct. Anal.*, 11(3):465–588, 2001.

[GT10] B. Green and T. Tao. An arithmetic regularity lemma, an associated counting lemma, and applications. In *An irregular mind*, volume 21 of *Bolyai Soc. Math. Stud.*, pages 261–334. János Bolyai Math. Soc., Budapest, 2010.

[GT12] B. Green and T. Tao. The quantitative behaviour of polynomial orbits on nilmanifolds. *Ann. of Math. (2)*, 175(2):465–540, 2012.

[GTZ12] B. Green, T. Tao, and T. Ziegler. An inverse theorem for the Gowers $U^{s+1}[N]$-norm. *Ann. of Math. (2)*, 176(2):1231–1372, 2012.

[Hål93] I. J. Håland. Uniform distribution of generalized polynomials. *J. Number Theory*, 45(3):327–366, 1993.

[Hål94] I. J. Håland. Uniform distribution of generalized polynomials of the product type. *Acta Arith.*, 67(1):13–27, 1994.

[HK95] I. J. Håland and D. E. Knuth. Polynomials involving the floor function. *Math. Scand.*, 76(2):194–200, 1995.

[Ked06] K. S. Kedlaya. Finite automata and algebraic extensions of function fields. *J. Théor. Nombres Bordeaux*, 18(2):379–420, 2006.

[Kle56] S. C. Kleene. Representation of events in nerve nets and finite automata. In *Automata studies*, Annals of mathematics studies, no. 34, pages 3–41. Princeton University Press, Princeton, N. J., 1956.

[Kon16] J. Konieczny. Gowers norms for the Thue–Morse and Rudin–Shapiro sequences. 2016. To appear in Annales de l'Institut Fourier. arXiv:1611.09985 [math.NT].

[Kro57] L. Kronecker. Zwei Sätze über Gleichungen mit ganzzahligen Coefficienten. *J. Reine Angew. Math.*, 53:173–175, 1857.

[Lei12] A. Leibman. A canonical form and the distribution of values of generalized polynomials. *Israel J. Math.*, 188:131–176, 2012.

[Lot02] M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2002.

[MR15] L. A. Medina and E. Rowland. p-regularity of the p-adic valuation of the Fibonacci sequence. *Fibonacci Quart.*, 53(3):265–271, 2015.

[Mil12] J. S. Miller. Two notes on subshifts. *Proc. Amer. Math. Soc.*, 140(5):1617–1622, 2012.

[MS02] R. Moosa and T. Scanlon. The Mordell-Lang conjecture in positive characteristic revisited. In *Model theory and applications*, volume 11 of *Quad. Mat.*, pages 273–296. Aracne, Rome, 2002.

[Mos08] Y. Moshe. On some questions regarding k-regular and k-context-free sequences. *Theoret. Comput. Sci.*, 400(1-3):62–69, 2008.

[Rig00] M. Rigo. Generalization of automatic sequences for numeration systems on a regular language. *Theoret. Comput. Sci.*, 244(1-2):271–281, 2000.

[Row10] E. S. Rowland. Non-regularity of $\lfloor \alpha + \log_k n \rfloor$. *Integers*, 10:A3, 19–23, 2010.

[SP11] J.-C. Schlage-Puchta. Regularity of a function related to the 2-adic logarithm. *Bull. Belg. Math. Soc. Simon Stevin*, 18(2):375–377, 2011.

[Sha88] J. Shallit. A generalization of automatic sequences. *Theoret. Comput. Sci.*, 61(1):1–16, 1988.

[SY11] Z. Shu and J.-Y. Yao. Analytic functions over $\mathbb{Z}_p$ and p-regular sequences. *C. R. Math. Acad. Sci. Paris*, 349(17-18):947–952, 2011.

[SYZS92] A. Szilard, S. Yu, K. Zhang, and J. Shallit. Characterizing regular languages with polynomial densities. In *Mathematical foundations of computer science 1992 (Prague, 1992)*, volume 629 of *Lecture Notes in Comput. Sci.*, pages 494–503. Springer, Berlin, 1992.

[vdPS91] A. J. van der Poorten and H. P. Schlickewei. Additive relations in fields. *J. Austral. Math. Soc. Ser. A*, 51(1):154–170, 1991.

[Wey16] H. Weyl. Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.*, 77(3):313–352, 1916.

[Yas99] S.-I. Yasutomi. On Sturmian sequences which are invariant under some substitutions. In *Number theory and its applications (Kyoto, 1997)*, volume 2 of *Dev. Math.*, pages 347–373. Kluwer Acad. Publ., Dordrecht, 1999.

(JB) Department of Mathematics and Computer Science, Institute of Mathematics, Jagiellonian University, ul. prof. Stanisława Łojasiewicza 6, 30-348 Kraków

E-mail address: jakub.byszewski@gmail.com

(JK) Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG

(JK, current address) Einstein Institute of Mathematics, Edmond J. Safra Campus, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, 9190401, Israel

E-mail address: jakub.konieczny@gmail.com