Enumeration and Decidable Properties of Automatic Sequences

Émilie Charlier
University of Waterloo
echarlier@uwaterloo.ca

This is joint work with Narad Rampersad and Jeffrey Shallit.

Abstract. We show that various aspects of k-automatic sequences — such as having an unbordered factor of length n — are both decidable and effectively enumerable. As a consequence it follows that many related sequences are either k-automatic or k-regular. These include many sequences previously studied in the literature, such as the recurrence function, the appearance function, and the repetitivity index. We also give a new characterization of the class of k-regular sequences. Many results extend to other sequences defined in terms of Pisot numeration systems.

1. Introduction

An infinite sequence $\mathbf{x} = (a(n))_{n\geq 0}$ is said to be *k-automatic* if it is computable by a finite automaton taking as input the base-k representation of n, and having $a(n)$ as the output associated with the last state encountered [AS03a].

Honkala [Hon86] showed that, given an automaton, it is decidable if the sequence it generates is ultimately periodic. Later, Leroux [Ler05] gave a polynomial-time algorithm for the problem.

Recently, Allouche, Rampersad, and Shallit [ARS09] found a different proof of Honkala's result using a more general technique. They showed that their technique suffices to show that the following properties (and many more) are decidable for k-automatic sequences:

(a) Given a rational number $r > 1$, whether $\mathbf{x}$ is r-power-free;
(b) Given a rational number $r > 1$, whether $\mathbf{x}$ contains infinitely many occurrences of r-powers;
(c) Given a rational number $r > 1$, whether $\mathbf{x}$ contains infinitely many distinct r-powers;
(d) Given a length l, whether $\mathbf{x}$ avoids palindromes of length $\geq l$.

In this paper we first show that many additional properties of automatic sequences are decidable using the same general technique. More significantly, we also show that related enumeration questions on automatic sequences (such as counting the number of distinct factors of length n) can be solved using a similar technique, in an entirely effective manner. As a consequence, we recover or improve results due to Mossé [Mos96]; Allouche, Baake, Cassaigne, and Damanik [ABCD03]; Currie and Saari [CS09]; Garel [Gar97]; Fagnot [Fag97]; and Brown, Rampersad, Shallit, and Vasiga [BRSV06].

2. Connection with logic

The technique used in [ARS09] was, at its core, very similar to previous techniques developed by Büchi, Bruyère, Michaux, Villemaire, and others, involving formal logic; see, e.g., [BHMV94]. As it turns out, the properties (a)–(d) above are decidable because they are expressible as predicates in the first-order structure $\langle \mathbb{N}, +, V_k \rangle$, where $V_k(n)$ is the largest power of k dividing n.

We briefly recall the technique discussed in [ARS09] in the context of a particular example. Suppose we want to decide if an automatic sequence $\mathbf{x}$ is squarefree (contains no nonempty square factor). Given an automaton M generating a k-automatic sequence $\mathbf{x}$, we create, via a series of transformations, a new automaton M' that accepts the base-k representations of integers corresponding to the squares in $\mathbf{x}$. For example, M' could accept those integers corresponding to the starting position of each square, or those integers corresponding to the lengths of the squares. The operations we can use in constructing M' include digit-by-digit addition or subtraction (with carry, if necessary), comparison, and lookup of the corresponding term in $\mathbf{x}$ (which comes from simulation of M). Nondeterminism can be used to implement "$\exists$", and "$\forall$" can be implemented by nondeterminism combined with suitable negations. Ultimately, then, deciding if $\mathbf{x}$ is squarefree corresponds to verifying that $L(M') = \emptyset$ for the M' we construct. Deciding whether $\mathbf{x}$ contains only finitely many square occurrences corresponds to verifying that $L(M')$ is finite. Both can easily be done by the standard methods for automata.

Thus, the main idea of [ARS09] can be restated as follows:

Theorem 1. *If we can express a property of a k-automatic sequence $\mathbf{x}$ using quantifiers, logical operations, integer variables, the operations of addition, subtraction, indexing into $\mathbf{x}$, and comparison of integers or elements of $\mathbf{x}$, then this property is decidable.*

We obtain the following new result. A word w is *bordered* if it begins and ends with the same word x with $0 < |x| \leq |w|/2$. (An example in English is `ingoing`.) Otherwise it is unbordered.

Theorem 2. *Let $\mathbf{x} = a(0)a(1)a(2)\cdots$ be a k-automatic sequence. Then the associated infinite sequence $\mathbf{b} = b(0)b(1)b(2)\cdots$ defined by*

$$b(n) = \begin{cases} 1, & \text{if } \mathbf{x} \text{ has an unbordered factor of length } n; \\ 0, & \text{otherwise;} \end{cases}$$

is k-automatic.

We now turn to deciding if a given automatic sequence has infinite *critical exponent* (e.g., [KS07]).

Theorem 3. *The following question is decidable: given a k-automatic sequence, does it contain powers of arbitrarily large exponent?*

In a similar fashion we can show

Theorem 4. *The following question is decidable: given a k-automatic sequence $\mathbf{x}$, does $\mathbf{x}$ contain arbitrarily large unbordered factors?*

Now we turn to questions of recurrence. An infinite word $\mathbf{a} = (a(n))_{n\geq 0}$ is said to be *recurrent* if every factor that occurs at least once in $\mathbf{a}$ occurs infinitely often. Equivalently, $\forall n \geq 0$, $\forall r \geq 1$, $\exists m > n$ such that $a(n+j) = a(m+j)$ for $0 \leq j < r$. Similarly, an infinite word $\mathbf{a} = (a(n))_{n\geq 0}$ is said to be *uniformly recurrent* if every factor that occurs at least once in $\mathbf{a}$ occurs infinitely often, with bounded gaps between consecutive occurrences. Equivalently, $\forall r \geq 1$, $\exists t > 0$, $\forall n \geq 0$, $\exists m \geq 0$ with $n < m < n + t$ such that $a(n + i) = a(m + i)$ for $0 \leq i < r$. Thus we recover the following recent result of Nicolas and Pritykin [NP09]:

Theorem 5. *It is decidable if a k-automatic sequence is recurrent or linearly recurrent.*

We now turn to questions of factors shared by two k-automatic sequences. Fagnot [Fag97] showed that it is decidable whether two such sequences $\mathbf{x} = a(0)a(1)\cdots$ and $\mathbf{y} = b(0)b(1)\cdots$ have exactly the same set of factors. This is also decidable by our methods. In a similar fashion, the question of whether the set of factors of one k-automatic word form a subset of the set of factors of another k-automatic word is decidable.

3. Enumeration

In this section we show that many sequences counting aspects of k-automatic sequences are k-regular. Recall that a sequence $(a(n))_{n\geq 0}$ is *k-regular* if the module generated by all sequences of the set $\{(a(k^e n + c))_{n\geq 0} \ : \ e \geq 0, \ 0 \leq c < k^e\}$ is finitely generated [AS92, AS03a, AS03b]. Alternatively, $(a(n))_{n\geq 0}$ is k-regular if $\sum_{n\geq 0} a(n)(n)_k$ is a noncommutative rational series [BR11], where $(n)_k$ is the canonical base-k encoding of n. The k-regular sequences play the same role for integer-valued sequences as the k-automatic sequences play for sequences over a finite alphabet. Classical examples of k-regular sequences include polynomials in n with non-negative coefficients and $s_k(n)$, the sum of the base-k digits of n.

Theorem 6. *Let $\mathbf{x} = a(0)a(1)a(2)\ldots$ be a k-automatic sequence. Let $b(n)$ be the number of distinct factors of length n in $\mathbf{x}$. Then $(b(n))_{n\geq 0}$ is a k-regular sequence.*

Remark 7. Mossé [Mos96] proved, among other things, that a sequence that is the fixed point of a k-uniform morphism has a k-regular subword complexity function. With our technique, we obtain her result for these sequences and also the slightly more general case of k-automatic sequence.

Theorem 8. *The sequence counting the number of palindromic factors of length n is k-regular.*

Remark 9. Allouche, Baake, Cassaigne, and Damanik [ABCD03, Theorem 10] proved that the palindrome complexity of the fixed point of a primitive k-uniform morphism is k-automatic. Our result is more general: it shows that the palindrome complexity of a k-automatic sequence is k-regular, and hence is k-automatic iff it is bounded. Jean-Paul Allouche kindly informs us that our result has just been obtained independently by Carpi and D'Alonzo [CD10].

Theorem 10. *Let $\mathbf{x} = a(0)a(1)a(2)\cdots$ be a k-automatic sequence. Then the following sequences are also k-automatic:*

(a) $b(i) = 1$ if there is a square beginning at position i; 0 otherwise
(b) $c(i) = 1$ if there is a square centered at position i; 0 otherwise
(c) $d(i) = 1$ if there is an overlap beginning at position i; 0 otherwise
(d) $e(i) = 1$ if there is a palindrome beginning at position i; 0 otherwise
(e) $f(i) = 1$ if there is a palindrome centered at position i; 0 otherwise

Theorem 11. *Let $\mathbf{x}$ and $\mathbf{y}$ be k-automatic sequences. Then the following are k-regular:*

(a) the number of distinct square factors in $\mathbf{x}$ of length n;
(b) the number of squares in $\mathbf{x}$ beginning at (centered at, ending at) position n;
(c) the length of the longest square in $\mathbf{x}$ beginning at (centered at, ending at) position n;
(d) the number of palindromes in $\mathbf{x}$ beginning at (centered at, ending at) position n;
(e) the length of the longest palindrome in $\mathbf{x}$ beginning at (centered at, ending at) position n;
(f) the length of the longest fractional power in $\mathbf{x}$ beginning at (ending at) position n;
(g) the number of distinct recurrent factors in $\mathbf{x}$ of length n;
(h) the number of factors of length n that occur in $\mathbf{x}$ but not in $\mathbf{y}$.
(i) the number of factors of length n that occur in both $\mathbf{x}$ and $\mathbf{y}$.

We now turn to some other measures that have received much attention. If an infinite word $\mathbf{x}$ is recurrent, then its recurrence function $R_{\mathbf{x}}(n) = R(n)$ is the smallest integer t such that every factor of length t of $\mathbf{x}$ contains as a factor every factor of length n. Said otherwise, it is the size of the smallest "window" one can slide along $\mathbf{x}$ and always contain all length-n factors.

Theorem 12. *If $\mathbf{x}$ is k-automatic, then $R_{\mathbf{x}}(n)$ is k-regular.*

Another measure is called "appearance" [AS03a, §10.10]. The appearance function $A_{\mathbf{x}}(n) = A(n)$ is the smallest integer t such that every factor of length n appears in a prefix of length t of $\mathbf{x}$. This can be proved in an analogous manner.

Theorem 13. *If $\mathbf{x}$ is k-automatic, then $A_{\mathbf{x}}(n)$ is k-regular.*

Next, we consider a measure due to Garel [Gar97]. The separator length $S_{\mathbf{x}}(n)$ is the length of the smallest factor that begins at position n of $\mathbf{x}$ and does not occur previously.

Theorem 14. *If $\mathbf{x}$ is k-automatic, then $S_{\mathbf{x}}(n)$ is k-regular.*

Remark 15. Garel proved this for the case of a fixed point of a uniform circular morphism; our proof works for the more general case of an arbitrary k-automatic sequence.

Finally, Carpi and D'Alonzo have introduced a measure they called repetitivity index [CD09]. This measure $I_{\mathbf{x}}(n)$ is the minimum distance between two consecutive occurrences of the same length-n factor in $\mathbf{x}$. But "$I_{\mathbf{x}}(n) > t$" is the same as saying for all $i, j \geq 0$ with $i \neq j$, the equality $\mathbf{x}[i..i+n-1] = \mathbf{x}[j..j+n-1]$ implies that $j - i > t$. Hence we get

Theorem 16. *If $\mathbf{x}$ is k-automatic, then its repetitivity index is k-regular.*

4. A new characterization of k-regular sequences

Carpi and Maggi [CM01] defined the class of k-synchronized sequences, a class which contains the k-automatic sequences and is properly contained in the class of k-regular sequences. A sequence $(u_n)_{n\geq 0}$ is k-synchronized if the relation $\{((n)_k, (u_n)_k) : n \geq 0\}$ is a right-synchronized rational relation. Roughly speaking, this means that the relation is realized by a length-preserving rational transduction, except that we also permit the presence of "padding" symbols at the end of one or the other component of the input. Here we give a similar transducer-based characterization of the more general class of k-regular sequences.

For us, a *j-uniform transducer* is a nondeterministic finite state machine $T = (Q, \Sigma, \delta, q_0, \tau, \Delta, F)$ where $\delta : Q \times \Sigma \to 2^Q$ and $\tau : \Sigma \to \Delta^j$ is a j-uniform morphism. An accepting path P begins at q_0 and ends at a state of F. The output associated with P is the concatenation of the outputs associated with the transitions. The output of T on an input x is the union of outputs associated with all accepting paths labeled x.

We work with strings over the alphabet $\Sigma' = \Sigma_k \times \Delta$, where $\Sigma_k = \{0, 1, \ldots, k-1\}$. For $x \in (\Sigma')^*$ we let $\pi_i(x)$ denote projection onto the i'th coordinate. For $x \in \Sigma_k^*$, $y \in \Delta^*$ with $|x| = |y|$ we let $x \times y$ denote the element of $(\Sigma')^*$ with $\pi_1(x \times y) = x$ and $\pi_2(x \times y) = y$.

Theorem 17. *Let $(b(n))_{n\geq 0}$ be a sequence taking values in $\mathbb{N} \cup \{+\infty\}$. Let $(n)_k \in \Sigma_k^*$ denote the canonical base-k encoding of n in base k, starting with the least significant digit.*

Then the following are equivalent:

(1) $(b(n))_{n\geq 0}$ is k-regular;

(2) there exist an integer m and vectors $\lambda \in \mathbb{N}^{1\times m}$, $\gamma \in \mathbb{N}^{m\times 1}$, and a matrix-valued morphism $\mu : \Sigma_k^* \to \mathbb{N}^{m\times m}$ such that $b(n) = \lambda\mu((n)_k)\gamma$ for all $n \geq 1$;

(3) there exist an alphabet Δ and a DFA $M = (Q, \Sigma_k \times \Delta, \delta, q_0, F)$ such that, for all $n \geq 1$,

$$b(n) = |\{x \in (\Sigma_k \times \Delta)^* \ : \ \pi_1(x) = (n)_k\}|;$$

(4) there exist an integer $j \geq 1$ and a j-uniform transducer T with inputs and outputs in Σ_k^* such that $b(n) = |T((n)_k)|$ for all $n \geq 1$. □

As an application we have:

Theorem 18. *Let E be any finite set of integers, and consider $b(n)$, the sequence that counts the number of base-k representations where the digits are chosen only from E. Then $b(n)$ is k-regular.*

5. Linear bounds

Yet another application of our method allows us to obtain linear bounds on many quantities associated with automatic sequences. As a first example, we recover an old result of Cobham [Cob72] on "subword" complexity.

Theorem 19. *The number of distinct factors of length n of an automatic sequence is $O(n)$.*

In a similar manner we can prove that all the quantities in Theorem 11 are either linearly bounded, or unbounded.

6. Other numeration systems

All our results transfer, *mutatis mutandis*, to the setting of other numeration systems where addition can be performed on numbers using a transducer that processes numbers starting with the least significant digit.

A (generalized) numeration system is given by an increasing sequence of integers $U = (U_i)_{i\geq 0}$ such that $U_0 = 1$ and $C_U := \lim_{i\to+\infty} U_{i+1}/U_i$ exists and is finite. Then the canonical U-representation of n (with least significant digit first), which is denoted by $(n)_U$, is the unique finite word w over the alphabet $\Sigma_U = \{0, \ldots, C_U - 1\}$ not ending with 0 and satisfying $n = \sum_{i=0}^{|w|-1} w[i]\, U_i$ and $\forall t \in \{0, \ldots, |w| - 1\}$, $\sum_{i=0}^{t} w[i]\, U_i < U_{t+1}$. The notion of k-automatic sequence extends naturally to this context: an infinite sequence $\mathbf{x}$ is said to be U-automatic if it is computable by a finite automaton taking as input the U-representation $(n)_U$ of n, and having $\mathbf{x}[n]$ as the output associated with the last state encountered.

A numeration system U is called *linear* if U satisfies a linear recurrence relation over $\mathbb{Z}$. A Pisot system is a linear numeration system U whose characteristic polynomial is the minimal polynomial of a Pisot number. Recall that a Pisot number is an algebraic integer greater than 1, all of whose conjugates have moduli less than 1. For example, all integer base numeration systems and the Fibonacci numeration system are Pisot systems. Frougny and Solomyak [FS96] proved that addition is U-recognizable within all Pisot systems U, i.e., it can be performed by a finite letter-to-letter transducer reading U-representations with least significant digit first. Bruyère and Hansel [BH97] then proved the following logical characterization of U-automatic sequences for Pisot systems: a sequence is U-automatic if and only if it is U-definable, i.e., it is expressible as a predicate of $\langle \mathbb{N}, +, V_U\rangle$, where $V_U(n)$ is the smallest U_i occurring in $(n)_U$ with a nonzero coefficient. Therefore, if U is a Pisot system, any combinatorial property of U-automatic words that can be described by a predicate of $\langle \mathbb{N}, +, V_U\rangle$ is decidable.

The notion of k-regular sequences extends to Pisot numeration systems: an infinite sequence $\mathbf{x}$ is said to be *U-regular* if the series $\sum_{n\geq 0} \mathbf{x}[n](n)_U$ is a noncommutative rational series. Thus we obtain

Theorem 20. *Let U be a Pisot numeration system and let $\mathbf{x}$ be any U-automatic word. The following sequences are U-automatic:*

(a) $a(n) = 1$ if there is a square beginning at (centered at, ending at) position n of $\mathbf{x}$, 0 otherwise;
(b) $b(n) = 1$ if there is a palindrome beginning at (centered at, ending at) position n of $\mathbf{x}$, 0 otherwise;
(c) $c(n) = 1$ if there is an unbordered factor beginning at (centered at, ending at) position n of $\mathbf{x}$, 0 otherwise.

The following sequences are U-regular:

(a) The number of distinct square factors beginning at (centered at, ending at) position n of $\mathbf{x}$;
(b) The number of distinct palindromic factors beginning at (centered at, ending at) position n of $\mathbf{x}$, 0 otherwise;
(c) The number of distinct unbordered factors beginning at (centered at, ending at) position n of $\mathbf{x}$, 0 otherwise.

Berstel showed that the cardinality of the set of unnormalized Fibonacci representations is Fibonacci-regular [Ber01], a result also obtained (but not published) by the third author about the same time. In analogy with Theorem 18 we have

Theorem 21. *The number of unnormalized representations of n in a Pisot numeration system U is U-regular.*

7. Closing remarks

It may be worth noting that the explicit constructions of automata we have given also imply bounds on the smallest example of (or counterexample to) the properties we consider. The bounds are essentially given by a tower of exponents whose height is related to the number of alternating quantifiers. For example,

Theorem 22. *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *are* k*-automatic sequences generated by automata with at most* q *states. If the set of factors of* $\mathbf{x}$ *differs from the set of factors of* $\mathbf{y}$*, then there exists a factor of length at most* $2^{2^{2^{2q^2}}}$ *that occurs in one word but not the other.*

References

[ABCD03] J.-P. Allouche, M. Baake, J. Cassaigne, and D. Damanik. Palindrome complexity. *Theoret. Comput. Sci.*, 292(1):9–31, 2003.

[ARS09] J.-P. Allouche, N. Rampersad, and J. Shallit. Periodicity, repetitions, and orbits of an automatic sequence. *Theoret. Comput. Sci.*, 410:2795–2803, 2009.

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

[AS03a] J.-P. Allouche and J. Shallit. *Automatic Sequences. Theory, Applications, Generalizations.* 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. Words.

[Ber01] J. Berstel. An exercise on Fibonacci representations. *Theor. Inform. Appl.*, 35(6):491–498 (2002), 2001. A tribute to Aldo de Luca.

[BH97] V. Bruyère and G. Hansel. Bertrand numeration systems and recognizability. *Theoret. Comput. Sci.*, 181(1):17–43, 1997. Latin American Theoretical Informatics (Valparaíso, 1995).

[BHMV94] V. Bruyère, G. Hansel, Ch. Michaux, and R. Villemaire. Logic and p-recognizable sets of integers. *Bull. Belg. Math. Soc. Simon Stevin*, 1(2):191–238, 1994. Journées Montoises (Mons, 1992).

[BR11] Jean Berstel and Christophe Reutenauer. *Noncommutative rational series with applications*, volume 137 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2011.

[BRSV06] S. Brown, N. Rampersad, J. Shallit, and T. Vasiga. Squares and overlaps in the Thue-Morse sequence and some variants. *Theor. Inform. Appl.*, 40(3):473–484, 2006.

[CD09] A. Carpi and V. D'Alonzo. On the repetitivity index of infinite words. *Internat. J. Algebra Comput.*, 19(2):145–158, 2009.

[CD10] A. Carpi and V. D'Alonzo. On factors of synchronized sequences. *Theoret. Comput. Sci.*, 411(44-46):3932–3937, 2010.

[CM01] A. Carpi and C. Maggi. On synchronized sequences and their separators. *Theor. Inform. Appl.*, 35(6):513–524 (2002), 2001. A tribute to Aldo de Luca.

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

[CS09] J. D. Currie and K. Saari. Least periods of factors of infinite words. *Theor. Inform. Appl.*, 43(1):165–178, 2009.

[Fag97] I. Fagnot. Sur les facteurs des mots automatiques. *Theoret. Comput. Sci.*, 172(1-2):67–89, 1997.

[FS96] Ch. Frougny and B. Solomyak. On the representation of integers in linear numeration systems. In *Ergodic theory of Z_d actions (Warwick, 1993–1994)*, volume 228 of *London Math. Soc. Lecture Note Ser.*, pages 345–368. Cambridge Univ. Press, Cambridge, 1996.

[Gar97] E. Garel. Séparateurs dans les mots infinis engendrés par morphismes. *Theoret. Comput. Sci.*, 180(1-2):81–113, 1997.

[Hon86] J. Honkala. A decision method for the recognizability of sets defined by number systems. *RAIRO Inform. Theor. Appl.*, 20(4):395–403, 1986.

[KS07] D. Krieger and J. Shallit. Every real number greater than 1 is a critical exponent. *Theoret. Comput. Sci.*, 381(1-3):177–182, 2007.

[Ler05] J. Leroux. A polynomial time Presburger criterion and synthesis for number decision diagrams. In *20th IEEE Symposium on Logic in Computer Science*, pages 147–156. IEEE Computer Society, Chicago, IL, USA, 2005.

[Mos96] B. Mossé. Reconnaissabilité des substitutions et complexité des suites automatiques. *Bull. Soc. Math. France*, 124(2):329–346, 1996.

[NP09] F. Nicolas and Y. Pritykin. On uniformly recurrent morphic sequences. *Internat. J. Found. Comput. Sci.*, 20(5):919–940, 2009.