Automata in Mathematics

Rational subsets of groups

L. Bartholdi[1] *P. V. Silva*[2,*]

[1]Mathematisches Institut
Georg-August Universität zu Göttingen
Bunsenstraße 3–5
D-37073 Göttingen, Germany
email: laurent.bartholdi@gmail.com

[2]Centro de Matemática, Faculdade de Ciências
Universidade do Porto
R. Campo Alegre 687
4169-007 Porto, Portugal
email: pvsilva@fc.up.pt

2010 Mathematics Subject Classification: 20F10, 20E05, 68Q45, 68Q70

Key words: Free groups, inverse automata, Stallings automata, rational subsets.

Contents

[*]The second author acknowledges support by Project ASA (PTDC/MAT/65481/2006) and C.M.U.P., financed by F.C.T. (Portugal) through the programmes POCTI and POSI, with national and E.U. structural funds.

Over the years, finite automata have been used effectively in the theory of infinite groups to represent rational subsets. This includes the important particular case of finitely generated subgroups (and the beautiful theory of Stallings automata for the free group case), but goes far beyond that: certain inductive procedures need a more general setting than mere subgroups, and rational subsets constitute the natural generalization. The connections between automata theory and group theory are rich and deep, and many are portrayed in Sims' book [58].

This chapter is divided into three parts: in Section 1 we introduce basic concepts, terminology and notation for finitely generated groups, devoting special attention to free groups. These will also be used in Chapter 24.

Section 2 describes the use of finite inverse automata to study finitely generated subgroups of free groups. The automaton recognizes elements of a subgroup, represented as words in the ambient free group.

Section 3 considers, more generally, rational subsets of groups, when good closure and decidability properties of these subsets are satisfied.

The authors are grateful to Stuart Margolis, Benjamin Steinberg and Pascal Weil for their remarks on a preliminary version of this text.

1 Finitely generated groups

Let G be a group. Given $A \subseteq G$, let $\langle A\rangle = (A \cup A^{-1})^*$ denote the subgroup of G *generated* by A. We say that $H \leqslant G$ is *finitely generated* and write $H \leqslant_{f.g.} G$ if $H = \langle A\rangle$ for some finite subset A of H.

Given $H \leqslant G$, we denote by $[G : H]$ the *index* of H in G, that is, the number of right cosets Hg for all $g \in G$; or, equivalently, the number of left cosets. If $[G : H]$ is finite, we write $H \leqslant_{f.i.} G$. It is well known that every finite index subgroup of a finitely generated group is finitely generated.

We denote by $\mathbb{1}$ the identity of G. An element $g \in G$ has *finite order* if $\langle g\rangle$ is finite. Elements $g, h \in G$ are *conjugate* if $h = x^{-1}gx$ for some $x \in G$. We use the notation $g^h = h^{-1}gh$ and $[g, h] = g^{-1}g^h$ to denote, respectively, conjugates and commutators.

Given an alphabet A, we denote by A^{-1} a set of *formal inverses* of A, and write $\widetilde{A} = A \cup A^{-1}$. We say that $\widetilde{A}$ is an *involutive alphabet*. We extend $^{-1} : A \to A^{-1} : a \mapsto a^{-1}$ to an involution on $\widetilde{A}^*$ through

$$(a^{-1})^{-1} = a, \quad (uv)^{-1} = v^{-1}u^{-1} \quad (a \in A,\ u, v \in \widetilde{A}^*)\,.$$

If $G = \langle A\rangle$, we have a canonical epimorphism $\rho : \widetilde{A}^* \twoheadrightarrow G$, mapping $a^{\pm 1} \in \widetilde{A}$ to $a^{\pm 1} \in G$. We present next some classical decidability problems:

Definition 1.1. Let $G = \langle A\rangle$ be a finitely generated group.

word problem: is there an algorithm that, upon receiving as input a word $u \in \widetilde{A}^*$, determines whether or not $\rho(u) = \mathbb{1}$?

conjugacy problem: is there an algorithm that, upon receiving as input words $u, v \in \widetilde{A}^*$, determines whether or not $\rho(u)$ and $\rho(v)$ are conjugate in G?

membership problem for $\mathcal{K} \subseteq 2^G$: is there for every $X \in \mathcal{K}$ an algorithm that, upon receiving as input a word $u \in \widetilde{A}^*$, determines whether or not $\rho(u) \in X$?

generalized word problem: is the membership problem for the class of finitely generated subgroups of G solvable?

order problem: is there an algorithm that, upon receiving as input a word $u \in \widetilde{A}^*$, determines whether $\rho(u)$ has finite or infinite order?

isomorphism problem for a class $\mathcal{G}$ of groups: is there an algorithm that, upon receiving as input a description of groups $G, H \in \mathcal{G}$, decides whether or not $G \cong H$? Typically, $\mathcal{G}$ may be a subclass of finitely presented groups (given by their presentation), or automata groups (see Chapter 24) given by automata.

We can also require complexity bounds on the algorithms; more precisely, we may ask with which complexity bound an answer to the problem may be obtained, and also with which complexity bound a witness (a normal form for the word problem, an element conjugating $\rho(u)$ to $\rho(v)$ in case they are conjugate, an expression of u in the generators of X in the generalized word problem) may be constructed.

1.1 Free groups

We recall that an equivalence relation $\sim$ on a semigroup S is a *congruence* if $a \sim b$ implies $ac \sim bc$ and $ca \sim cb$ for all $a, b, c \in S$.

Definition 1.2. Given an alphabet A, let $\sim$ denote the congruence on $\widetilde{A}^*$ generated by the relation

$$\{(aa^{-1}, 1) \mid a \in \widetilde{A}\}\,. \tag{1.1}$$

The quotient $F_A = \widetilde{A}^*/{\sim}$ is the *free group on A*. We denote by $\theta : \widetilde{A}^* \to F_A$ the canonical morphism $u \mapsto [u]_\sim$.

Free groups admit the following universal property: for every map $f : A \to G$, there is a unique group morphism $F_A \to G$ that extends f.

Alternatively, we can view (1.1) as a *confluent* length-reducing rewriting system on $\widetilde{A}^*$, where each word $w \in \widetilde{A}^*$ can be transformed into a unique *reduced* word $\overline{w}$ with no factor of the form aa^{-1}, see [10]. As a consequence, the equivalence

$$u \sim v \qquad \Leftrightarrow \qquad \overline{u} = \overline{v} \qquad\qquad (u, v \in \widetilde{A}^*)$$

solves the word problem for F_A.

We shall use the notation $R_A = \overline{\widetilde{A}^*}$. It is well known that F_A is isomorphic to R_A under the binary operation

$$u \star v = \overline{uv} \qquad\qquad (u, v \in R_A)\,.$$

We recall that the *length* $|g|$ of $g \in F_A$ is the length of the reduced form of g, also denoted by $\overline{g}$.

The letters of A provide a natural *basis* for F_A: they generate F_A and satisfy no nontrivial relations, that is, all reduced words on these generators represent distinct elements of F_A. A group is free if and only if it has a basis.

Throughout this chapter, we assume A to be a finite alphabet. It is well known that free groups F_A and F_B are isomorphic if and only if $\operatorname{Card} A = \operatorname{Card} B$. This leads to the concept of *rank* of a free group F: the *cardinality* of a basis of F, denoted by $\operatorname{rk} F$. It is common to use the notation F_n to denote a free group of rank n.

We recall that a reduced word u is *cyclically reduced* if uu is also reduced. Any reduced word $u \in R_A$ admits a unique decomposition of the form $u = vwv^{-1}$ with w cyclically reduced. A solution for the conjugacy problem follows easily from this: first reduce the words cyclically; then two cyclically reduced words in R_A are conjugate if and only if they are cyclic permutations of each other. On the other hand, the order problem admits a trivial solution: only the identity has finite order. Finally, the generalized word problem shall be discussed in the following section.

2 Inverse automata and Stallings' construction

The study of finitely generated subgroups of free groups entered a new era in the early eighties when Stallings made explicit and effective a construction [59] that can be traced back to the early part of the twentieth century in Schreier's coset graphs (see [58] and §24.1) and to Serre's work [50]. Stallings' seminal paper was built over *immersions of finite graphs*, but the alternative approach using finite inverse automata became much more popular over the years; for more on their link, see [30]. An extensive survey has been written by Kapovich and Miasnikov [24].

Stallings' construction for $H \leqslant_{f.g.} F_A$ consists in taking a finite set of generators for H in reduced form, building the so-called *flower automaton* and then proceeding to make this automaton deterministic through the operation known as *Stallings foldings*. This is clearly a terminating procedure, but the key fact is that the construction is independent from both the given finite generating set and the chosen folding sequence. A short simple automata-theoretic proof of this claim will be given. The finite inverse automaton $\mathcal{S}(H)$ thus obtained is usually called the *Stallings automaton* of H. Over the years, Stallings automata became the standard representation for finitely generated subgroups of free groups and are involved in many of the algorithmic results presently obtained.

Several of these algorithms are implemented in computer software, see e.g. CRAG [2], or the packages AUTOMATA and FGA in GAP [17].

2.1 Inverse automata

An automaton $\mathcal{A}$ over an involutive alphabet $\widetilde{A}$ is *involutive* if, whenever (p, a, q) is an edge of $\mathcal{A}$, so is (q, a^{-1}, p). Therefore it suffices to depict just the *positively labelled* edges (having label in A) in their graphical representation.

Definition 2.1. An involutive automaton is *inverse* if it is deterministic, trim and has a single final state.

If the latter happens to be the initial state, it is called the *basepoint*. It follows easily from the computation of the Nerode equivalence (see §10.2) that every inverse automaton is a minimal automaton.

Finite inverse automata capture the idea of an action (of a finite *inverse monoid*, their *transition monoid*) on a finite set (the vertex set) through partial bijections. We recall that a monoid M is inverse if, for every $x \in M$, there exists a unique $y \in M$ such that $xyx = x$ and $y = yxy$; then M acts by partial bijections on itself.
The next result is easily proven, but is quite useful.

Proposition 2.1. *Let $\mathcal{A}$ be an inverse automaton and let $p \xrightarrow{uvv^{-1}w} q$ be a path in $\mathcal{A}$. Then there exists also a path $p \xrightarrow{uw} q$ in $\mathcal{A}$.*

Another important property relates languages to morphisms. For us, a *morphism* between deterministic automata $\mathcal{A}$ and $\mathcal{A}'$ is a mapping φ between their respective vertex sets which preserves initial vertices, final vertices and edges, in the sense that $(\varphi(p), a, \varphi(q))$ is an edge of $\mathcal{A}'$ whenever (p, a, q) is an edge of $\mathcal{A}$.

Proposition 2.2. *Given inverse automata $\mathcal{A}$ and $\mathcal{A}'$, then $L(\mathcal{A}) \subseteq L(\mathcal{A}')$ if and only if there exists a morphism $\varphi : \mathcal{A} \to \mathcal{A}'$. Moreover, such a morphism is unique.*

Proof. ($\Rightarrow$): Given a vertex q of $\mathcal{A}$, take a successful path

$$\to q_0 \xrightarrow{u} q \xrightarrow{v} t \to$$

in $\mathcal{A}$, for some $u, v \in \widetilde{A}^*$. Since $L(\mathcal{A}) \subseteq L(\mathcal{A}')$, there exists a successful path

$$\to q_0' \xrightarrow{u} q' \xrightarrow{v} t' \to$$

in $\mathcal{A}'$. We take $\varphi(q) = q'$.

To show that φ is well defined, suppose that

$$\to q_0 \xrightarrow{u'} q \xrightarrow{v'} t \to$$

is an alternative successful path in $\mathcal{A}$. Since $u'v \in L(\mathcal{A}) \subseteq L(\mathcal{A}')$, there exists a successful path

$$\to q_0' \xrightarrow{u'} q'' \xrightarrow{v} t' \to$$

in $\mathcal{A}'$ and it follows that $q' = q''$ since $\mathcal{A}'$ is inverse. Thus φ is well defined.

It is now routine to check that φ is a morphism from $\mathcal{A}$ to $\mathcal{A}'$ and that it is unique.

($\Leftarrow$): Immediate from the definition of morphism. □

2.2 Stallings' construction

Let X be a finite subset of R_A. We build an involutive automaton $\mathcal{F}(X)$ by fixing a basepoint q_0 and gluing to it a *petal* labelled by every word in X as follows: if $x = a_1 \dots a_k \in X$, with $a_i \in \widetilde{A}$, the petal consists of a closed path of the form

$$q_0 \xrightarrow{a_1} \bullet \xrightarrow{a_2} \cdots \xrightarrow{a_k} q_0$$

and the respective inverse edges. All such intermediate vertices $\bullet$ are assumed to be distinct in the automaton. For obvious reasons, $\mathcal{F}(X)$ is called the *flower automaton* of X.

The automaton $\mathcal{F}(X)$ is almost an inverse automaton – except that it need not be deterministic. We can fix it by performing a sequence of so-called *Stallings foldings*. Assume that $\mathcal{A}$ is a trim involutive automaton with a basepoint, possessing two distinct edges of the form

$$p \xrightarrow{a} q, \quad p \xrightarrow{a} r \tag{2.1}$$

for $a \in \widetilde{A}$. The *folding* is performed by identifying these two edges, as well as the two respective inverse edges. In particular, the vertices q and r are also identified (if they were distinct).

The number of edges is certain to decrease through foldings. Therefore, if we perform enough of them, we are sure to turn $\mathcal{F}(X)$ into a finite inverse automaton.

Definition 2.2. The *Stallings automaton* of X is the finite inverse automaton $\mathcal{S}(X)$ obtained through folding $\mathcal{F}(X)$.

We shall see that $\mathcal{S}(X)$ depends only on the finitely generated subgroup $\langle X \rangle$ of F_A generated by X, being in particular independent from the choice of foldings taken to reach it.

Since inverse automata are minimal, it suffices to characterize $L(\mathcal{S}(X))$ in terms of H to prove uniqueness (up to isomorphism):

Proposition 2.3. *Fix $H \leqslant_{f.g.} F_A$ and let $X \subseteq R_A$ be a finite generating set for H. Then*

$$L(\mathcal{S}(X)) = \bigcap \{L \subseteq \widetilde{A}^* \mid L \text{ is recognized by a finite inverse automaton} \\ \text{with a basepoint and } \overline{H} \subseteq L\}\,.$$

Proof. ($\supseteq$): Clearly, $\mathcal{S}(X)$ is a finite inverse automaton with a basepoint. Since $X \cup X^{-1} \subseteq L(\mathcal{F}(X)) \subseteq L(\mathcal{S}(X))$, it follows easily from Proposition 2.1 that

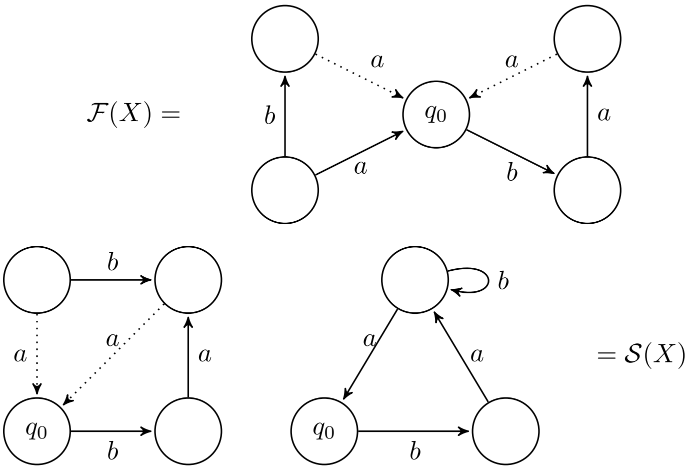
$$\overline{H} \subseteq L(\mathcal{S}(X))\,. \tag{2.2}$$

($\subseteq$): Let $L \subseteq \widetilde{A}^*$ be recognized by a finite inverse automaton $\mathcal{A}$ with a basepoint, with $\overline{H} \subseteq L$. Since $X \subseteq \overline{H}$, we have an automaton morphism from $\mathcal{F}(X)$ to $\mathcal{A}$, hence $L(\mathcal{F}(X)) \subseteq L$. To prove that $L(\mathcal{S}(X)) \subseteq L$, it suffices to show that inclusion in L is preserved through foldings.

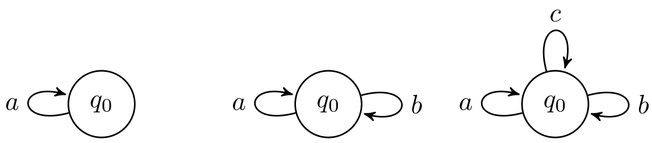
Indeed, assume that $L(\mathcal{B}) \subseteq L$ and $\mathcal{B}'$ is obtained from $\mathcal{B}$ by folding the two edges in (2.1). It is immediate that every successful path $q_0 \xrightarrow{u} t$ in $\mathcal{B}'$ can be lifted to a successful path $q_0 \xrightarrow{v} t$ in $\mathcal{B}$ by successively inserting the word $a^{-1}a$ into u. Now $v \in L = L(\mathcal{A})$ implies $u \in L$ in view of Proposition 2.1. □

Now, given $H \leqslant F_A$ finitely generated, we take a finite set X of generators. Without loss of generality, we may assume that X consists of reduced words, and we may define $\mathcal{S}(H) = \mathcal{S}(X)$ to be the *Stallings automaton* of H.

Example 2.1. Stallings' construction for $X = \{a^{-1}ba, ba^2\}$, where the next edges to be identified are depicted by dotted lines, is



A simple, yet important example is given by applying the construction to F_n itself, when we obtain the so-called *bouquet* of n circles:



$\mathcal{S}(F_1)$ $\mathcal{S}(F_2)$ $\mathcal{S}(F_3)$

In terms of complexity, the best known algorithm for the construction of $\mathcal{S}(X)$ is due to Touikan [61]. Its time complexity is $O(n \log^* n)$, where n is the sum of the lengths of the elements of X.

2.3 Basic applications

The most fundamental application of Stallings' construction is an elegant and efficient solution to the generalized word problem:

Theorem 2.4. *The generalized word problem in F_A is solvable.*

We will see many groups in Chapter 24 that have solvable word problem; however, few of them have solvable generalized word problem. The proof of Theorem 2.4 relies on

Proposition 2.5. *Consider $H \leqslant_{f.g.} F_A$ and $u \in F_A$. Then $u \in H$ if and only if $\overline{u} \in L(\mathcal{S}(H))$.*

Proof. $(\Rightarrow)$: Follows from (2.2).

$(\Leftarrow)$: It follows easily from the last paragraph of the proof of Proposition 2.3 that, if $\mathcal{B}'$ is obtained from $\mathcal{B}$ by performing Stallings foldings, then $\overline{L(\mathcal{B}')} = \overline{L(\mathcal{B})}$. Hence, if

$H = \langle X \rangle$, we get

$$\overline{L(\mathcal{S}(H))} = \overline{L(\mathcal{F}(X))} = \overline{(X \cup X^{-1})^*} = \overline{H}$$

and the implication follows. □

It follows from our previous remark that the complexity of the generalized word problem is $O(n \log^* n + m)$, where n is the sum of the lengths of the elements of X and m is the length of the input word. In particular, once the subgroup X has been fixed, complexity is linear in m.

Example 2.2. We may use the Stallings automaton constructed in Example 2.1 to check that $baba^{-1}b^{-1} \in H = \langle a^{-1}ba, ba^2 \rangle$ but $ab \notin H$.

Stallings automata also provide an effective construction for bases of finitely generated subgroups. Consider $H \leqslant_{f.g.} F_A$, and let m be the number of vertices of $\mathcal{S}(H)$. A *spanning tree* T for $\mathcal{S}(H)$ consists of $m-1$ edges and their inverses which, together, connect all the vertices of $\mathcal{S}(H)$. Given a vertex p of $\mathcal{S}(H)$, we denote by g_p the T-*geodesic* connecting the basepoint q_0 to p, that is, $q_0 \xrightarrow{g_p} p$ is the shortest path contained in T connecting q_0 to p.

Proposition 2.6. *Let $H \leqslant_{f.g.} F_A$ and let T be a spanning tree for $\mathcal{S}(H)$. Let E_+ be the set of positively labelled edges of $\mathcal{S}(H)$. Then H is free with basis*

$$Y = \{g_p a g_q^{-1} \mid (p,a,q) \in E_+ \setminus T\}\,.$$

Proof. It follows from Proposition 2.5 that $L(\mathcal{S}(H)) \subseteq H$, hence $Y \subseteq H$. To show that $H = \langle Y \rangle$, take $h = a_1 \cdots a_k \in H$ in reduced form ($a_i \in \widetilde{A}$). By Proposition 2.5, there exists a successful path

$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \cdots \xrightarrow{a_k} q_k = q_0$$

in $\mathcal{S}(H)$. For $i = 1, \ldots, k$, we have either $g_{q_{i-1}} a_i g_{q_i}^{-1} \in Y \cup Y^{-1}$ or $\overline{g_{q_{i-1}} a_i g_{q_i}^{-1}} = 1$, the latter occurring if $(q_{i-1}, a_i, q_i) \in T$. In any case, we get

$$h = a_1 \cdots a_k = \overline{(g_{q_0} a_1 g_{q_1}^{-1})(g_{q_1} a_2 g_{q_2}^{-1}) \cdots (g_{q_{k-1}} a_k g_{q_0}^{-1})} \in \langle Y \rangle$$

and so $H = \langle Y \rangle$.

It remains to show that the elements of Y satisfy no nontrivial relations. Let $y_1, \ldots, y_k \in Y \cup Y^{-1}$ with $y_i \neq y_{i-1}^{-1}$ for $i = 2, \ldots, k$. Write $y_i = g_{p_i} a_i g_{r_i}^{-1}$, where $a_i \in \widetilde{A}$ labels the edge not in T. It follows easily from $y_i \neq y_{i-1}^{-1}$ and the definition of spanning tree that

$$\overline{y_1 \cdots y_k} = g_{p_1} a_1 \overline{g_{r_1}^{-1} g_{p_2}} a_2 \cdots a_{k-1} \overline{g_{r_{k-1}}^{-1} g_{p_k}} a_k g_{r_k}\,,$$

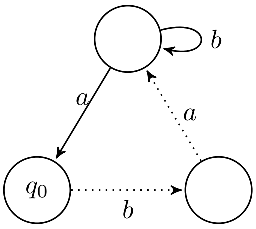
a nonempty reduced word if $k \geqslant 1$. Therefore Y is a basis of H as claimed. □

In the process, we also obtain a proof of the Nielsen-Schreier Theorem, in the case of finitely generated subgroups. A simple topological proof may be found in [40]:

Theorem 2.7 (Nielsen-Schreier). *Every subgroup of a free group is itself free.*

Example 2.3. We use the Stallings automaton constructed in Example 2.1 to construct a basis of $H = \langle a^{-1}ba, ba^2\rangle$.

If we take the spanning tree T defined by the dotted lines in



then $\operatorname{Card} E_+ \setminus T = 2$ and the corresponding basis is $\{ba^2, baba^{-1}b^{-1}\}$. Another choice of spanning tree actually proves that the original generating set is also a basis.

We remark that Proposition 2.6 can be extended to the case of infinitely generated subgroups, proving the general case of Theorem 2.7. However, in this case there is no effective construction such as Stallings', and the (infinite) inverse automaton $\mathcal{S}(H)$ remains a theoretical object, using appropriate cosets as vertices.

Another classical application of Stallings' construction regards the identification of finite index subgroups.

Proposition 2.8. *Consider $H \leqslant_{f.g.} F_A$.*

(i) *H is a finite index subgroup of F_A if and only if $\mathcal{S}(H)$ is a complete automaton.*

(ii) *If H is a finite index subgroup of F_A, then its index is the number of vertices of $\mathcal{S}(H)$.*

Proof. (i) $(\Rightarrow)$: Suppose that $\mathcal{S}(H)$ is not complete. Then there exist some vertex q and some $a \in \widetilde{A}$ such that $q \cdot a$ is undefined. Let g be a geodesic connecting the basepoint q_0 to q in $\mathcal{S}(H)$. We claim that

$$Hga^m \neq Hga^n \quad \text{if} \quad m - n > |g| \,. \tag{2.3}$$

Indeed, $Hga^m = Hga^n$ implies $ga^{m-n}g^{-1} \in H$ and so $\overline{ga^{m-n}g^{-1}} \in L(\mathcal{S}(H))$ by Proposition 2.5. Since ga is reduced due to $\mathcal{S}(H)$ being inverse, it follows from $m - n > |g|$ that $\overline{gaa^{m-n-1}g^{-1}} = \overline{ga^{m-n}g^{-1}} \in L(\mathcal{S}(H))$: indeed, g^{-1} is not long enough to erase all the a's. Since $\mathcal{S}(H)$ is deterministic, $q \cdot a$ must be defined, a contradiction. Therefore (2.3) holds and so H has infinite index.

$(\Leftarrow)$: Let Q be the vertex set of $\mathcal{S}(H)$ and fix a geodesic $q_0 \xrightarrow{g_q} q$ for each $q \in Q$. Take $u \in F_A$. Since $\mathcal{S}(H)$ is complete, we have a path $q_0 \xrightarrow{u} q$ for some $q \in Q$. Hence $ug_q^{-1} \in H$ and so $u = ug_q^{-1}g_q \in Hg_q$. Therefore $F_A = \bigcup_{q \in Q} Hg_q$ and so $H \leqslant_{f.i.} F_A$.

(ii) In view of $F_A = \bigcup_{q \in Q} Hg_q$, it suffices to show that the cosets Hg_q are all distinct. Indeed, assume that $Hg_p = Hg_q$ for some vertices $p, q \in Q$. Then $g_pg_q^{-1} \in H$ and so $\overline{g_pg_q^{-1}} \in L(\mathcal{S}(H))$ by Proposition 2.5. On the other hand, since $\mathcal{S}(H)$ is complete, we have a path

$$q_0 \xrightarrow{g_pg_q^{-1}} r$$

for some $r \in Q$. In view of Proposition 2.1, and by determinism, we get $r = q_0$. Hence we have paths

$$p \xrightarrow{g_q^{-1}} q_0, \quad q \xrightarrow{g_q^{-1}} q_0 .$$

Since $\mathcal{S}(H)$ is inverse, we get $p = q$ as required. □

Example 2.4. Since the Stallings automaton constructed in Example 2.1 is not complete, it follows that $\langle a^{-1}ba, ba^2\rangle$ is not a finite index subgroup of F_2.

Corollary 2.9. *If $H \leqslant F_A$ has index n, then* $\operatorname{rk} H = 1 + n(\operatorname{Card} A - 1)$.

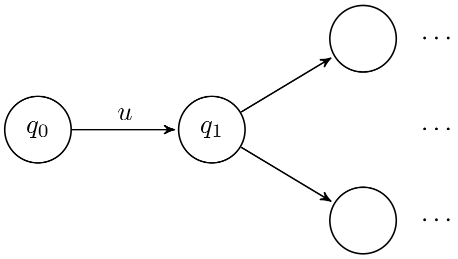
Proof. By Proposition 2.8, the automaton $\mathcal{S}(H)$ has n vertices and $n \operatorname{Card} A$ positive edges. A spanning tree has $n-1$ positive edges, so $\operatorname{rk} H = n \operatorname{Card} A - (n-1) = 1 + n(\operatorname{Card} A - 1)$ by Proposition 2.6. □

Beautiful connections between finite index subgroups and certain classes of *bifix codes* — set of words none of which is a prefix or a suffix of another — have recently been unveiled by Berstel, De Felice, Perrin, Reutenauer and Rindone [6].

2.4 Conjugacy

We start now a brief discussion of conjugacy. Recall that the *outdegree* of a vertex q is the number of edges starting at q and the *geodesic distance* in a connected graph is the length of the shortest undirected path connecting two vertices.

Since the original generating set is always taken in reduced form, it follows easily that there is at most one vertex in a Stallings automaton having outdegree < 2: the basepoint q_0. Assuming that H is nontrivial, $\mathcal{S}(H)$ must always be of the form



where q_1 is the closest vertex to q_0 (in terms of geodesic distance) having outdegree > 2 (since there is at least one vertex having such outdegree). Note that $q_1 = q_0$ if q_0 has outdegree > 2 itself. We call $q_0 \xrightarrow{u}$ the *tail* (which is empty if $q_1 = q_0$) and the remaining subgraph the *core* of $\mathcal{S}(H)$.

Note that $\mathcal{S}(H)$, and its core, may be understood as follows. Consider the graph with vertex set $F_A/H = \{gH \mid g \in F_A\}$, with an edge from gH to agH for each generator $a \in A$. Then this graph, called the *Schreier graph* (see §24.1) of $H\backslash F_A$, consists of finitely many trees attached to the core of $\mathcal{S}(H)$.

Theorem 2.10. *There is an algorithm that decides whether or not two finitely generated subgroups of F_A are conjugate.*

Proof. Finitely generated subgroups G, H are conjugate if and only if the cores of $\mathcal{S}(G)$ and $\mathcal{S}(H)$ are equal (up to their basepoints). □

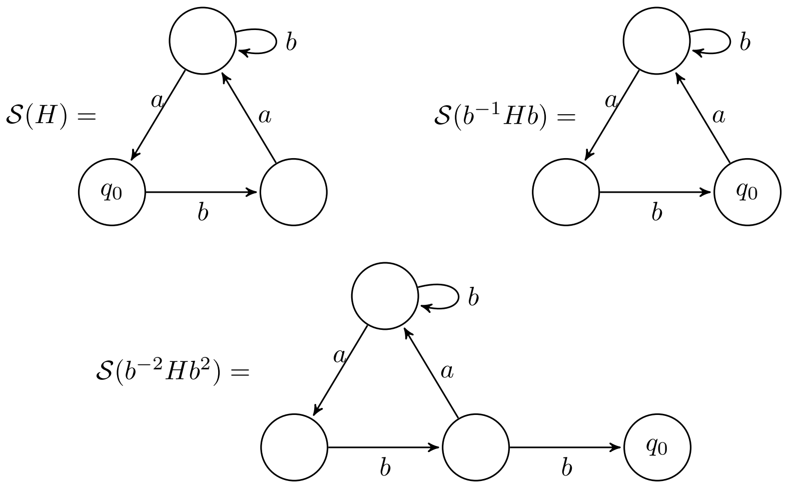
The Stallings automata of the conjugates of H can be obtained in the following ways: (1) declaring a vertex in the core C to be the basepoint; (2) gluing a tail to some vertex in the core C and taking its other endpoint to be the basepoint.

Note that the tail must be glued in some way that keeps the automaton inverse, so in particular this second type of operation can only be performed if the automaton is not complete, or equivalently, if H has infinite index. An immediate consequence is the following classical

Proposition 2.11. *A finite rank normal subgroup of a free group is either trivial or has finite index.*

Moreover, a finite index subgroup H is normal if and only if its Stallings automaton is *vertex-transitive*, that is, if all choices of basepoint yield the same automaton.

Example 2.5. Stallings automata of some conjugates of $H = \langle a^{-1}ba, ba^2 \rangle$:



We can also use the previous discussion on the structure of (finite) Stallings automata to provide them with an abstract characterization.

Proposition 2.12. *A finite inverse automaton with a basepoint is a Stallings automaton if and only if it has at most one vertex of outdegree 1: the basepoint.*

Proof. Indeed, for any such automaton we can take a spanning tree and use it to construct a basis for the subgroup as in the proof of Proposition 2.6. □

2.5 Further algebraic properties

The study of intersections of finitely generated subgroups of F_A provides further applications of Stallings automata. Howson's classical theorem admits a simple proof using the *direct product* of two Stallings automata; it is also an immediate consequence of Theorem 3.1 and Corollary 3.4(ii).

Theorem 2.13 (Howson). *If $H, K \leqslant_{f.g.} F_A$, then also $H \cap K \leqslant_{f.g.} F_A$.*

Stallings automata are also naturally related to the famous Hanna Neumann Conjecture, recently proved by Mineyev [36] and Friedman [16]: given $H, K \leqslant_{f.g.} F_A$, then $\operatorname{rk}(H \cap K) - 1 \leqslant (\operatorname{rk} H - 1)(\operatorname{rk} K - 1)$. The conjecture arose in a paper of Hanna Neumann [38], where the inequality $\operatorname{rk}(H \cap K) - 1 \leqslant 2(\operatorname{rk} H - 1)(\operatorname{rk} K - 1)$ was also proved. In one of the early applications of Stallings' approach, Gersten provided an alternative geometric proof of Hanna Neumann's inequality [18].

A *free factor* of a free group F_A can be defined as a subgroup H generated by a subset of a basis of F_A. This is equivalent to saying that there exists a *free product decomposition* $F_A = H * K$ for some $K \leqslant F_A$.

Since the rank of a free factor never exceeds the rank of the ambient free group, it is easy to construct examples of subgroups which are not free factors: it follows easily from Proposition 2.6 that any free group of rank $\geqslant 2$ can have subgroups of arbitrary finite rank (and even infinite countable).

The problem of identifying free factors has a simple solution based on Stallings automata [55]: one must check whether or not a prescribed number of vertex identifications in the Stallings automaton can lead to a bouquet. However, the most efficient solution, due to Roig, Ventura and Weil [44], involves Whitehead automorphisms and will therefore be postponed to §23.2.7.

Given a morphism $\varphi : \mathcal{A} \to \mathcal{B}$ of inverse automata, let the *morphic image* $\varphi(\mathcal{A})$ be the subautomaton of $\mathcal{B}$ induced by the image by φ of all the successful paths of $\mathcal{A}$.

The following classical result characterizes the extensions of $H \leqslant_{f.g.} F_A$ contained in F_A. We present the proof from [35]:

Theorem 2.14 (Takahasi [60]). *Given $H \leqslant_{f.g.} F_A$, one can effectively compute finitely many extensions $K_1, \dots, K_m \leqslant_{f.g.} F_A$ of H such that the following conditions are equivalent for every $K \leqslant_{f.g.} F_A$:*

(i) *$H \leqslant K$;*
(ii) *K_i is a free factor of K for some $i \in \{1, \dots, m\}$.*

Proof. Let $\mathcal{A}_1, \dots, \mathcal{A}_m$ denote all the morphic images of $\mathcal{S}(H)$, up to isomorphism. Since a morphic image cannot have more vertices than the original automaton, there are only finitely many isomorphism classes. Moreover, it follows from Proposition 2.12 that, for $i = 1, \dots, m$, $\mathcal{A}_i = \mathcal{S}(K_i)$ for some $K_i \leqslant_{f.g.} F_A$. Since $L(\mathcal{S}(H)) \subseteq L(\mathcal{A}_i) = L(\mathcal{S}(K_i))$, it follows from Proposition 2.5 that $H \leqslant K_i$. Clearly, we can construct all $\mathcal{A}_i$ and therefore all K_i.

(i) $\Rightarrow$ (ii). If $H \leqslant K$, it follows from Stallings' construction that $L(\mathcal{S}(H)) \subseteq L(\mathcal{S}(K))$ and so there is a morphism $\varphi : \mathcal{S}(H) \to \mathcal{S}(K)$ by Proposition 2.2. Let $\mathcal{A}_i$

be, up to isomorphism, the morphic image of $\mathcal{S}(H)$ through φ. Since $\mathcal{A}_i = \mathcal{S}(K_i)$ is a subautomaton of $\mathcal{S}(K)$, it follows easily from Proposition 2.6 that K_i is a free factor of K: it suffices to take a spanning tree for $\mathcal{S}(K_i)$, extend it to a spanning tree for $\mathcal{S}(K)$, and the induced basis of K_i will be contained in the induced basis of K.

(ii) ⇒ (i) is immediate. □

An interesting research line related to this result is built on the concept of algebraic extension, introduced by Kapovich and Miasnikov [24], and inspired by the homonymous field-theoretical classical notion. Given $H \leqslant K \leqslant F_A$, we say that K is an *algebraic* extension of H if no proper free factor of K contains H. Miasnikov, Ventura and Weil [35] proved that the set of algebraic extensions of H is finite and effectively computable, and it constitutes the minimum set of extensions $K_1, \dots, K_m$ satisfying the conditions of Theorem 2.14.

Consider a subgroup H of a group G. The *commensurator* of H in G, is

$$\mathrm{Comm}_G(H) = \{g \in G \mid H \cap H^g \text{ has finite index in } H \text{ and } H^g\}. \tag{2.4}$$

For example, the commensurator of $\mathrm{GL}_n(\mathbb{Z})$ in $\mathrm{GL}_n(\mathbb{R})$ is $\mathrm{GL}_n(\mathbb{Q})$.

The special case of finite-index extensions, $H \leqslant_{f.i.} K \leqslant F_A$ is of special interest, and can be interpreted in terms of commensurators. It can be proved (see [24, Lemma 8.7] and [57]) that every $H \leqslant_{f.g.} F_A$ has a maximum finite-index extension inside F_A, denoted by H_{fi}; and $H_{fi} = \mathrm{Comm}_{F_A}(H)$. Silva and Weil [57] proved that $\mathcal{S}(H_{fi})$ can be constructed from $\mathcal{S}(H)$ using a simple automata-theoretic algorithm:

(1) The standard minimization algorithm is applied to the core of $\mathcal{S}(H)$, *taking all vertices as final*.

(2) The original tail of $\mathcal{S}(H)$ is subsequently reinstated in this new automaton, at the appropriate vertex.

We present now an application of different type, involving transition monoids. It follows easily from the definitions that the transition monoid of a finite inverse automaton is always a *finite inverse monoid*. Given a group G, we say that a subgroup $H \leqslant G$ is *pure* if the implication

$$g^n \in H \Rightarrow g \in H \tag{2.5}$$

holds for all $g \in F_A$ and $n \geqslant 1$. If p is a prime, we say that H is *p-pure* if (2.5) holds when $(n, p) = 1$.

The next result is due to Birget, Margolis, Meakin and Weil, and is the only natural problem among applications of Stallings automata that is known so far to be PSPACE-complete [8].

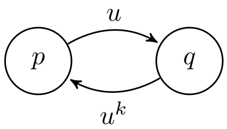
Proposition 2.15. *For every $H \leqslant_{f.g.} F_A$, the following conditions hold:*

(i) *H is pure if and only if the transition monoid of $\mathcal{S}(H)$ is aperiodic.*

(ii) *H is p-pure if and only if the transition monoid of $\mathcal{S}(H)$ has no subgroups of order p.*

Proof. Both conditions in (i) are easily proved to be equivalent to the nonexistence in

$\mathcal{S}(H)$ of a cycle of the form



$(k \geqslant 1, p \neq q)$

where u can be assumed to be cyclically reduced. The proof of (ii) runs similarly. □

2.6 Topological properties

We require for this subsection some basic topological concepts, which the reader can recover from Chapter 17.

For all $u, v \in F_A$, written in reduced form as elements of R_A, let $u \wedge v$ denote the longest common prefix of u and v. The *prefix metric* d on F_A is defined, for all $u, v \in F_A$, by

$$d(u,v) = \begin{cases} 2^{-|u \wedge v|-1} & \text{if } u \neq v \\ 0 & \text{if } u = v \end{cases}$$

It follows easily from the definition that d is an ultrametric on F_A, satisfying in particular the axiom

$$d(u,v) \leqslant \max\{d(u,w), d(w,v)\}\,.$$

The *completion* of this metric space is compact; its extra elements are *infinite reduced words* $a_1a_2a_3\dots$, with all $a_i \in \widetilde{A}$, and constitute the *hyperbolic boundary* ∂F_A of F_A, see §24.1.5. Extending the operator $\wedge$ to $F_A \cup \partial F_A$ in the obvious way, it follows easily from the definitions that, for every infinite reduced word α and every sequence $(u_n)_n$ in F_A,

$$\alpha = \lim_{n\to+\infty} u_n \qquad \text{if and only if} \qquad \lim_{n\to+\infty} |\alpha \wedge u_n| = +\infty\,. \tag{2.6}$$

The next result shows that Stallings automata are given a new role in connection with the prefix metric. We denote by $\operatorname{cl} H$ the closure of H in the completion of F_A.

Proposition 2.16. *If $H \leqslant_{f.g.} F_A$, then $\operatorname{cl} H$ is the union of H with the set of all $\alpha \in \partial F_A$ that label paths in $\mathcal{S}(H)$ out of the basepoint.*

Proof. Since the topology of F_A is discrete, we have $\operatorname{cl} H \cap F_A = H$.

($\subseteq$): If $\alpha \in \partial F_A$ does not label a path in $\mathcal{S}(H)$ out of the basepoint, then $\{|\alpha \wedge h| : h \in H\}$ is finite and so no sequence of H can converge to α by (2.6).

($\supseteq$): Let $\alpha = a_1a_2a_3\cdots \in \partial F_A$, with $a_i \in \widetilde{A}$, label a path in $\mathcal{S}(H)$ out of the basepoint. Let m be the number of vertices of $\mathcal{S}(H)$. For every $n \geqslant 1$, there exists some word w_n of length $< m$ such that $a_1 \cdots a_n w_n \in H$. Now $\alpha = \lim_{n\to+\infty} a_1 \cdots a_n w_n$ by (2.6) and so $\alpha \in \operatorname{cl} H$. □

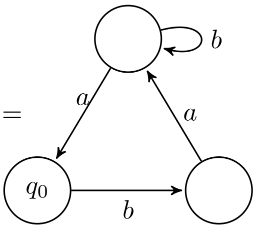
The *profinite topology* on F_A is defined in Chapter 17: for every $u \in F_A$, the collection $\{Ku \mid K \leqslant_{f.i.} F_A\}$ constitutes a basis of clopen neighbourhoods of u. In his seminal 1983 paper [59], Stallings gave an alternative proof of Marshall Hall's Theorem:

Theorem 2.17 (M. Hall). *Every finitely generated subgroup of F_A is closed for the profinite topology.*

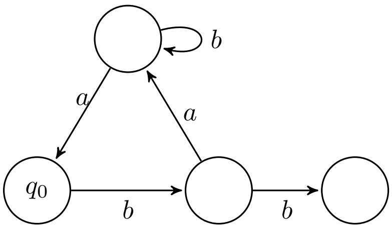
Proof. Fix $H \leqslant_{f.g.} F_A$ and let $u \in F_A \setminus H$ be written in reduced form as an element of R_A. In view of Proposition 2.5, u does not label a loop at the basepoint q_0 of $\mathcal{S}(H)$. If there is no path $q_0 \xrightarrow{u} \cdots$ in $\mathcal{S}(H)$, we add new edges to $\mathcal{S}(H)$ to get a finite inverse automaton $\mathcal{A}$ having a path $q_0 \xrightarrow{u} q \neq q_0$. Otherwise just take $\mathcal{A} = \mathcal{S}(H)$. Next add new edges to $\mathcal{A}$ to get a finite complete inverse automaton $\mathcal{B}$. In view of Propositions 2.8 and 2.12, we have $\mathcal{B} = \mathcal{S}(K)$ for some $K \leqslant_{f.i.} F_A$. Hence Ku is open and contains u. Since $H \cap Ku \neq \emptyset$ yields $u \in K^{-1}H = K$, contradicting Proposition 2.5, it follows that $H \cap Ku = \emptyset$ and so H is closed as claimed. □

Example 2.6. We consider the above construction for $H = \langle a^{-1}ba, ba^2 \rangle$ and $u = b^2$:

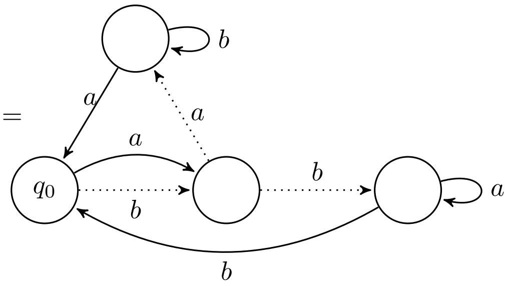
$\mathcal{S}(H) =$



$\mathcal{A} =$



$\mathcal{B} =$



If we take the spanning tree defined by the dotted lines in $\mathcal{B}$, it follows from Proposition 2.6 that

$$K = \langle ba^{-1}, b^3, b^2ab^{-2}, ba^2, baba^{-1}b^{-1} \rangle$$

is a finite index subgroup of F_2 such that $H \cap Kb^2 = \emptyset$.

We recall that a group G is *residually finite* if its finite index subgroups have trivial intersection. Considering the trivial subgroup in Theorem 2.17, we deduce

Corollary 2.18. *F_A is residually finite.*

We remark that Ribes and Zalessky extended Theorem 2.17 to products of finitely many finitely generated subgroups of F_A, see [42]. This result is deeply connected to the solution of Rhodes' Type II conjecture, see [41, Chapter 4].

If $\mathbf{V}$ denotes a pseudovariety of finite groups (see Chapter 16), the *pro-$\mathbf{V}$ topology* on F_A is defined by considering that each $u \in F_A$ has

$$\{Ku \mid K \trianglelefteq_{f.i.} F_A,\ F_A/K \in \mathbf{V}\}$$

as a basis of clopen neighbourhoods. The closure for the pro-**V** topology of $H \leqslant_{f.g} F_A$ can be related to an extension property of $\mathcal{S}(H)$, and Margolis, Sapir and Weil used automata to prove that efficient computation can be achieved for the pseudovarieties of finite p-groups and finite nilpotent groups [32]. The original computability proof for the p-group case is due to Ribes and Zalessky [43].

2.7 Dynamical properties

We shall mention briefly some examples of applications of Stallings automata to the study of endomorphism dynamics, starting with Gersten's solution of the subgroup orbit problem [19].

The subgroup orbit problem consists in finding an algorithm to decide, for given $H, K \leqslant_{f.g.} F_A$, whether or not $K = \varphi(H)$ for some automorphism φ of F_A. Equivalently, this can be described as deciding whether or not the automorphic orbit of a finitely generated subgroup is recursive.

Gersten's solution adapts to the context of Stallings automata Whitehead's idea to solve the orbit problem for words [64]. Whitehead's proof relies on a suitable decomposition of automorphisms as products of elementary factors (which became known as *Whitehead automorphisms*), and on using these as a tool to compute the elements of minimum length in the automorphic orbit of the word. In the subgroup case, word length is replaced by the number of vertices of the Stallings automaton.

The most efficient solution to the problem of identifying free factors [44], mentioned in §23.2.5, also relies on this approach: $H \leqslant_{f.g.} F_A$ is a free factor if and only if the Stallings automaton of some automorphic image of H has a single vertex (that is, a bouquet).

Another very nice application is given by the following theorem of Goldstein and Turner [20]:

Theorem 2.19. *The fixed point subgroup of an endomorphism of F_A is finitely generated.*

Proof. Let φ be an endomorphism of F_A. For every $u \in F_A$, define $Q(u) = \varphi(u)u^{-1}$. We define a potentially infinite automaton $\mathcal{A}$ by taking

$$\{Q(u) \mid u \in F_A\} \subseteq F_A$$

as the vertex set, all edges of the form $Q(u)\xrightarrow{a}Q(au)$ with $u \in F_A,\ a \in \widetilde{A}$, and fixing $\mathbb{1}$ as the basepoint. Then $\mathcal{A}$ is a well-defined inverse automaton.

Next we take $\mathcal{B}$ to be the subautomaton of $\mathcal{A}$ obtained by retaining only those vertices and edges that lie in successful paths labelled by reduced words. Clearly, $\mathcal{B}$ is still an inverse automaton, and it is easy to check that it must be the Stallings automaton of the fixed point subgroup of φ.

It remains to be proved that $\mathcal{B}$ is finite. We define a subautomaton $\mathcal{C}$ of $\mathcal{B}$ by removing exactly one edge among each inverse pair

$$Q(u)\xrightarrow{a}Q(au), \quad Q(au)\xrightarrow{a^{-1}}Q(u)$$

with $a \in A$ as follows: if a^{-1} is the last letter of $Q(au)$, we remove $Q(u)\xrightarrow{a}Q(au)$; otherwise, we remove $Q(au)\xrightarrow{a^{-1}}Q(u)$.

Let M denote the maximum length of the image of a letter by φ. We claim that, whenever $|Q(v)| > 2M$, the vertex $Q(v)$ has outdegree at most 1.

Indeed, if $Q(v)\xrightarrow{a^{-1}}Q(a^{-1}v)$ is an edge in $\mathcal{C}$ for $a \in A$, then a^{-1} is the last letter of $Q(v)$. On the other hand, if $Q(v)\xrightarrow{b}Q(bv)$ is an edge in $\mathcal{C}$ for $b \in A$, then b^{-1} is not the last letter of $Q(bv)$. Since $Q(bv) = \varphi(b)Q(v)b^{-1}$ and $|Q(v)| > 2|\varphi(b)|$, then b must be the last letter of $Q(v)$ in this case. Since $Q(v)$ has at most one last letter, it follows that its outdegree is at most 1.

Let $\mathcal{D}$ be a finite subautomaton of $\mathcal{C}$ containing all vertices $Q(v)$ such that $|Q(v)| \leqslant 2M$. Suppose that $p\longrightarrow q$ is an edge in $\mathcal{C}$ not belonging to $\mathcal{D}$. Since $p\longrightarrow q$, being an edge of $\mathcal{B}$, must lie in some reduced path, and by the outdegree property of $\mathcal{C}$, it is easy to see that there exists some path in $\mathcal{C}$ of the form

$$p'\longrightarrow p\longrightarrow q\longrightarrow r\longleftarrow r'$$

where p', r' are vertices in $\mathcal{D}$. Since there are only finitely many directed paths out of $\mathcal{D}$, it follows that $\mathcal{C}$ is finite and so is $\mathcal{B}$. Therefore the fixed point subgroup of φ is finitely generated. □

Note that this proof is not by any means constructive. Bogopolski and Maslakova give in [9] an algorithm that computes the fixed point subgroup of a free group automorphism; it relies on the sophisticated *train track* theory of Bestvina and Handel [7] and other algebraic geometry tools. The general endomorphism case remains open.

Stallings automata were also used by Ventura in the study of various properties of fixed subgroups, considering in particular arbitrary families of endomorphisms [62, 34] (see also [63]).

Automata also play a part in the study of *infinite fixed points*. In [53], these are considered for the continuous extension of a virtually injective endomorphism to the hyperbolic boundary of a virtually free group.

3 Rational and recognizable subsets

Rational subsets generalize the notion of finitely generated from subgroups to arbitrary subsets of a group, and can be quite useful in establishing inductive procedures that need to go beyond the territory of subgroups. Similarly, recognizable subsets extend the notion of finite index subgroups. Basic properties and results can be found in [5] or [47].

We consider a finitely generated group $G = \langle A \rangle$, with the canonical map $\pi : F_A \to G$. A subset of G is *rational* if it is the image by $\rho = \pi\theta$ of a rational subset of $\widetilde{A}^*$, and is *recognizable* if its full preimage under ρ is rational in $\widetilde{A}^*$.

For every group G, the classes $\operatorname{Rat} G$ and $\operatorname{Rec} G$ satisfy the following closure properties:

- $\operatorname{Rat} G$ is (effectively) closed under union, product, star, morphisms, inversion, subgroup generating.
- $\operatorname{Rec} G$ is (effectively) closed under boolean operations, translation, product, star, inverse morphisms, inversion, subgroup generating.

Kleene's Theorem is not valid for groups: $\operatorname{Rat} G = \operatorname{Rec} G$ if and only if G is finite. However, if the class of rational subsets of G possesses some extra algorithmic properties, then many decidability/constructibility results can be deduced for G. Two properties are particularly coveted for $\operatorname{Rat} G$:

- (effective) closure under complement (yielding closure under all the boolean operations);
- decidable membership problem for arbitrary rational subsets.

In these cases, one may often solve problems (e.g. equations, or systems of equations) whose statement lies far out of the rational universe, by proving that the solution is a rational set.

3.1 Rational and recognizable subgroups

We start by some basic, general facts. The following result is essential to connect language theory to group theory.

Theorem 3.1 (Anisimov and Seifert). *A subgroup H of a group G is rational if and only if H is finitely generated.*

Proof. ($\Rightarrow$): Let H be a rational subgroup of G and let $\pi : F_A \to G$ denote a morphism. Then there exists a finite $\widetilde{A}$-automaton $\mathcal{A}$ such that $H = \rho(L(\mathcal{A}))$. Assume that $\mathcal{A}$ has m vertices and let X consist of all the words in $\rho^{-1}(H)$ of length $< 2m$. Since A is finite, so is X. We claim that $H = \langle \rho(X) \rangle$. To prove it, it suffices to show that

$$u \in L(\mathcal{A}) \Rightarrow \rho(u) \in \langle \rho(X) \rangle \tag{3.1}$$

holds for every $u \in \widetilde{A}^*$. We use induction on $|u|$. By definition of X, (3.1) holds for words of length $< 2m$. Assume now that $|u| \geqslant 2m$ and (3.1) holds for shorter words. Write $u = vw$ with $|w| = m$. Then there exists a path

$$\to q_0 \xrightarrow{v} q \xrightarrow{z} t \to$$

in $\mathcal{A}$ with $|z| < m$. Thus $vz \in L(\mathcal{A})$ and by the induction hypothesis $\rho(vz) \in \langle \rho(X) \rangle$. On the other hand, $|z^{-1}w| < 2m$ and $\rho(z^{-1}w) = \rho(z^{-1}v^{-1})\rho(vw) \in H$, hence $z^{-1}w \in X$ and so $\rho(u) = \rho(vz)\rho(z^{-1}w) \in \langle \rho(X) \rangle$, proving (3.1) as required.

($\Leftarrow$) is trivial. □

It is an easier task to characterize recognizable subgroups:

Proposition 3.2. *A subgroup H of a group G is recognizable if and only if it has finite index.*

Proof. ($\Rightarrow$): In general, a recognizable subset of G is of the form NX, where $N \trianglelefteq_{f.i.} G$ and $X \subseteq G$ is finite. If $H = NX$ is a subgroup of G, then $N \subseteq H$ and so H has finite index as well.

($\Leftarrow$): This follows from the well-known fact that every finite index subgroup H of G contains a finite index normal subgroup N of G, namely $N = \bigcap_{g \in G} gHg^{-1}$. Since N has finite index, H must be of the form NX for some finite $X \subseteq G$. □

3.2 Benois' Theorem

The central result in this subsection is Benois' Theorem, the cornerstone of the whole theory of rational subsets of free groups:

Theorem 3.3 (Benois).

(i) *If $L \subseteq \widetilde{A}^*$ is rational, then $\overline{L}$ is also rational, and can be effectively constructed from L.*

(ii) *A subset of R_A is a rational language as a subset of $\widetilde{A}^*$ if and only if it is rational as a subset of F_A.*

We illustrate this in the case of finitely generated subgroups: temporarily calling "Benois automata" those automata recognizing rational subsets of R_A, we may convert them to Stallings automata by "folding" them, at the same time making sure they are inverse automata. Given a Stallings automaton, one intersects it with R_A to obtain a Benois automaton.

Proof. (i) Let $\mathcal{A} = (Q, \widetilde{A}, E, I, T)$ be a finite automaton recognizing L. We define a sequence $(\mathcal{A}_n)_n$ of finite automata with ε-transitions as follows. Let $\mathcal{A}_0 = \mathcal{A}$. Assuming that $\mathcal{A}_n = (Q, \widetilde{A}, E_n, I, T)$ is defined, we consider all instances of ordered pairs $(p, q) \in Q \times Q$ such that

$$\text{there exists a path } p \xrightarrow{aa^{-1}} q \text{ in } \mathcal{A}_n \text{ for some } a \in \widetilde{A}, \text{ but no path } p \xrightarrow{1} q. \tag{P}$$

Clearly, there are only finitely many instances of (P) in $\mathcal{A}_n$. We define E_{n+1} to be the union of E_n with all the new edges $(p, 1, q)$, where $(p, q) \in Q \times Q$ is an instance of (P). Finally, we define $\mathcal{A}_{n+1} = (Q, \widetilde{A}, E_{n+1}, I, T)$. In particular, note that $\mathcal{A}_n = \mathcal{A}_{n+k}$ for every $k \geqslant 1$ if there are no instances of (P) in $\mathcal{A}_n$.

Since Q is finite, the sequence $(\mathcal{A}_n)_n$ is ultimately constant, say after reaching $\mathcal{A}_m$. We claim that

$$\overline{L} = L(\mathcal{A}_m) \cap R_A\,. \tag{3.2}$$

Indeed, take $u \in L$. There exists a sequence of words $u = u_0, u_1, \ldots, u_{k-1}, u_k = \overline{u}$ where each term is obtained from the preceding one by erasing a factor of the form aa^{-1} for some $a \in \widetilde{A}$. A straightforward induction shows that $u_i \in L(\mathcal{A}_i)$ for $i = 0, \ldots, k$, since the existence of a path $p \xrightarrow{aa^{-1}} q$ in $\mathcal{A}_i$ implies the existence of a path $p \xrightarrow{1} q$ in $\mathcal{A}_{i+1}$. Hence $\overline{u} = u_k \in L(\mathcal{A}_k) \subseteq L(\mathcal{A}_m)$ and it follows that $\overline{L} \subseteq L(\mathcal{A}_m) \cap R_A$.

For the opposite inclusion, we start by noting that any path $p \xrightarrow{u} q$ in $\mathcal{A}_{i+1}$ can be lifted to a path $p \xrightarrow{v} q$ in $\mathcal{A}_i$, where v is obtained from u by inserting finitely many factors of the form aa^{-1}. It follows that

$$\overline{L(\mathcal{A}_m)} = \overline{L(\mathcal{A}_{m-1})} = \cdots = \overline{L(\mathcal{A}_0)} = \overline{L}$$

and so $L(\mathcal{A}_m) \cap R_A \subseteq \overline{L(\mathcal{A}_m)} = \overline{L}$. Thus (3.2) holds.

Since

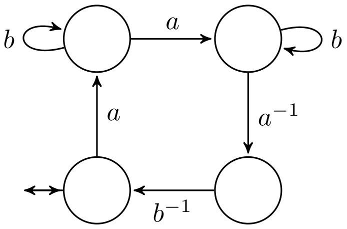
$$R_A = \widetilde{A}^* \setminus \bigcup_{a \in \widetilde{A}} \widetilde{A}^* aa^{-1} \widetilde{A}^*$$

is obviously rational, and the class of rational languages is closed under intersection, it follows that $\overline{L}$ is rational. Moreover, we can effectively compute the automaton $\mathcal{A}_m$ and a finite automaton recognizing R_A, hence the direct product construction can be used to construct a finite automaton recognizing the intersection $\overline{L} = L(\mathcal{A}_m) \cap R_A$.

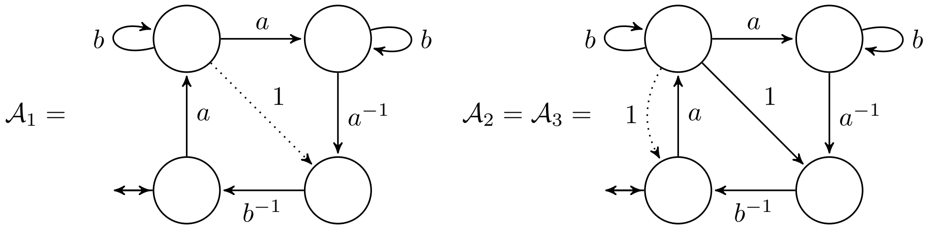
(ii) Consider $X \subseteq R_A$. If $X \in \operatorname{Rat} \widetilde{A}^*$, then $\theta(X) \in \operatorname{Rat} F_A$ and so X is rational as a subset of F_A.

Conversely, if X is rational as a subset of F_A, then $X = \theta(L)$ for some $L \in \operatorname{Rat} \widetilde{A}^*$. Since $X \subseteq R_A$, we get $X = \overline{L}$. Now part (i) yields $\overline{L} \in \operatorname{Rat} \widetilde{A}^*$ and so $X \in \operatorname{Rat} \widetilde{A}^*$ as required. □

Example 3.1. Let $\mathcal{A} = \mathcal{A}_0$ be depicted by



We get



and we can then proceed to compute $\overline{L} = L(\mathcal{A}_2) \cap R_2$.

The following result summarizes some of the most direct consequences of Benois' Theorem:

Corollary 3.4.

(i) *F_A has decidable rational subset membership problem.*

(ii) *$\operatorname{Rat} F_A$ is closed under the boolean operations.*

Proof. (i) Given $X \in \operatorname{Rat} F_A$ and $u \in F_A$, write $X = \theta(L)$ for some $L \in \operatorname{Rat} \widetilde{A}^*$. Then $u \in X$ if and only if $\overline{u} \in \overline{X} = \overline{L}$. By Theorem 3.3(i), we may construct a finite automaton recognizing $\overline{L}$ and therefore decide whether or not $\overline{u} \in \overline{L}$.

(ii) Given $X \in \operatorname{Rat} F_A$, we have $\overline{F_A \setminus X} = R_A \setminus \overline{X}$ and so $F_A \setminus X \in \operatorname{Rat} F_A$ by Theorem 3.3. Therefore $\operatorname{Rat} F_A$ is closed under complement.

Since $\operatorname{Rat} F_A$ is trivially closed under union, it follows from De Morgan's laws that it is closed under intersection as well. □

Note that we can associate algorithms to these boolean closure properties of $\operatorname{Rat} F_A$ in a constructive way. We remark also that the proof of Theorem 3.3 can be clearly adapted to more general classes of rewriting systems (see [10]). Theorem 3.3 and Corollary 3.4 have been generalized several times by Benois herself [4] and by Sénizergues, who obtained the most general versions. Sénizergues' results [48] hold for *rational length-reducing left basic confluent* rewriting systems and remain valid for the more general notion of *controlled* rewriting system.

3.3 Rational versus recognizable

Since F_A is a finitely generated monoid, it follows that every recognizable subset of F_A is rational [5, Proposition III.2.4]. We turn to the problem of deciding which rational subsets of F_A are recognizable. The first proof, using rewriting systems, is due to Sénizergues [49] but we follow the shorter alternative proof from [52], where a third alternative proof, of a more combinatorial nature, was also given.

Given a subset X of a group G, we define the *right stabilizer* of X to be the submonoid of G defined by

$$R(X) = \{g \in G \mid Xg \subseteq X\}\,.$$

Next let

$$K(X) = R(X) \cap (R(X))^{-1} = \{g \in G \mid Xg = X\}$$

be the largest subgroup of G contained in $R(X)$ and let

$$N(X) = \bigcap_{g \in G} gK(X)g^{-1}$$

be the largest normal subgroup of G contained in $K(X)$, and therefore in $R(X)$.

Lemma 3.5. *A subset X of a group G is recognizable if and only if $K(X)$ is a finite index subgroup of G.*

In fact, the Schreier graph (see §24.1) of $K(X)\backslash G$ is the underlying graph of an automaton recognizing X, and $G/N(X)$ is the syntactic monoid of X.

Proof. ($\Rightarrow$): If $X \subseteq G$ is recognizable, then $X = NF$ for some $N \trianglelefteq_{f.i.} G$ and $F \subseteq G$ finite. Hence $N \subseteq R(X)$ and so $N \subseteq K(X)$ since $N \leqslant G$. Since N has finite index in G, so does $K(X)$.

($\Leftarrow$): If $K(X)$ is a finite index subgroup of G, so is $N = N(X)$. Indeed, a finite index subgroup has only finitely many conjugates (having also finite index) and a finite intersection of finite index subgroups is easily checked to have finite index itself.

Therefore it suffices to show that $X = FN$ for some finite subset F of G. Since N has finite index, the claim follows from $XN = X$, in turn an immediate consequence of $N \subseteq R(X)$. □

Proposition 3.6. *It is decidable whether or not a rational subset of F_A is recognizable.*

Proof. Take $X \in \operatorname{Rat} F_A$. In view of Lemma 3.5 and Proposition 2.8, it suffices to show that $K(X)$ is finitely generated and effectively computable.

Given $u \in F_A$, we have

$$u \notin R(X) \Leftrightarrow Xu \nsubseteq X \Leftrightarrow Xu \cap (F_A \setminus X) \neq \emptyset \Leftrightarrow u \in X^{-1}(F_A \setminus X),$$

hence

$$R(X) = F_A \setminus (X^{-1}(F_A \setminus X)) .$$

It follows easily from the fact that the class of rational languages is closed under reversal and morphisms, combined with Theorem 3.3(ii), that $X^{-1} \in \operatorname{Rat} F_A$. Since $\operatorname{Rat} F_A$ is trivially closed under product, it follows from Corollary 3.4 that $R(X)$ is rational and effectively computable, and so is $K(X) = R(X) \cap (R(X))^{-1}$. By Theorem 3.1, the subgroup $K(X)$ is finitely generated and the proof is complete. □

These results are related to the Sakarovitch Conjecture [46], which states that every rational subset of F_A must be either recognizable or *disjunctive*: a subset X of a monoid M is disjunctive if it has trivial syntactic congruence, or equivalently, if any morphism $\varphi : M \to M'$ recognizing X is necessarily injective.

In the group case, it follows easily from the proof of the direct implication of Lemma 3.5 that the projection $G \to G/N$ recognizes $X \subseteq G$ if and only if $N \subseteq N(X)$. Thus X is disjunctive if and only if $N(X)$ is the trivial subgroup.

The Sakarovitch Conjecture was first proved in [49], but once again we follow the shorter alternative proof from [52]:

Theorem 3.7 (Sénizergues). *A rational subset of F_A is either recognizable or disjunctive.*

Proof. Since the only subgroups of $\mathbb{Z}$ are the trivial subgroup and finite index subgroups, we may assume that $\operatorname{Card} A > 1$.

Take $X \in \operatorname{Rat} F_A$. By the proof of Proposition 3.6, the subgroup $K(X)$ is finitely generated. In view of Lemma 3.5, we may assume that $K(X)$ is not a finite index subgroup. Thus $\mathcal{S}(K(X))$ is not complete by Proposition 2.8. Let q_0 denote the basepoint of $\mathcal{S}(K(X))$. Since $\mathcal{S}(K(X))$ is not complete, $q_0 \cdot u$ is undefined for some reduced word u.

Let w be an arbitrary nonempty reduced word. We must show that $w \notin N(X)$. Suppose otherwise. Since u, w are reduced and $\operatorname{Card} A > 1$, there exist enough letters to make sure that there is some word $v \in R_A$ such that $uvwv^{-1}u^{-1}$ is reduced. Now $w \in N(X)$, hence $uvwv^{-1}u^{-1} \in N(X) \subseteq K(X)$ by normality. Since $uvwv^{-1}u^{-1}$ is reduced, it follows from Proposition 2.5 that $uvwv^{-1}u^{-1}$ labels a loop at q_0 in $\mathcal{S}(K(X))$, contradicting $q_0 \cdot u$ being undefined. Thus $w \notin N(X)$ and so $N(X) = 1$. Therefore X is disjunctive as required. □

3.4 Beyond free groups

Let $\pi : F_A \twoheadrightarrow G$ be a morphism onto a group G. We consider the *word problem submonoid* of a group G, defined as

$$W_\pi(G) = (\pi\theta)^{-1}(\mathbb{1}). \tag{3.3}$$

Proposition 3.8 (Anisimov). *The language $W_\pi(G)$ is rational if and only if G is finite.*

Proof. If G is finite, it is easy to check that $W_\pi(G)$ is rational by viewing the Cayley graph of G (see §24.1) as an automaton. Conversely, if $W_\pi(G)$ is rational, then $\pi^{-1}(\mathbb{1})$ is a finitely generated normal subgroup of F_A, either finite index or trivial by the proof of Theorem 3.7. It is well known that the *Dyck language* $D_A = \theta^{-1}(\mathbb{1})$ is not rational if $\mathrm{Card}\,A > 0$, thus it follows easily that $\pi^{-1}(\mathbb{1})$ has finite index and therefore G must be finite. □

How about groups with context-free $W_\pi(G)$? A celebrated result by Muller and Schupp [37], with a contribution by Dunwoody [15], relates them to *virtually free groups*: these are groups with a free subgroup of finite index.

As usual, we focus on the case of G being finitely generated. We claim that G has a *normal* free subgroup F_A of finite index, with A finite. Indeed, letting F be a finite-index free subgroup of G, it suffices to take $F' = \bigcap_{g\in G} gFg^{-1}$. Since F has finite index, so does F', see the proof of Lemma 3.5. Taking a morphism $\pi : F_B \to G$ with B finite, we get from Corollary 2.9 that $\pi^{-1}(F') \leqslant_{f.i.} F_B$ is finitely generated, so F' is itself finitely generated. Finally, F' is a subgroup of F, so F' is still free by Theorem 2.7, and we can write $F' = F_A$.

We may therefore decompose G as a finite disjoint union of the form

$$G = F_A b_0 \cup F_A b_1 \cup \cdots \cup F_A b_m, \qquad \text{with } b_0 = 1. \tag{3.4}$$

Theorem 3.9 (Muller & Schupp). *The language $W_\pi(G)$ is context-free if and only if G is virtually free.*

Sketch of proof. If G is virtually free, the rewriting system implicit in (3.4) provides a rational transduction between $W_\pi(G)$ and D_A.

The converse implication can be proved by arguing geometrical properties of the Cayley graph of G such as in Chapter 24; briefly said, one deduces from the context-freeness of $W_\pi(G)$ that the Cayley graph of G is close (more precisely, quasi-isometric) to a tree. □

It follows that virtually free groups have decidable word problem. In Chapter 24, we shall discuss the word problem for more general classes of groups using other techniques.

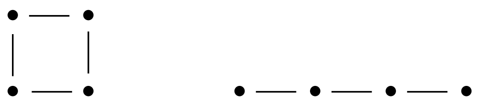
Grunschlag proved that every rational (respectively recognizable) subset of a virtually free group G decomposed as in (3.4) admits a decomposition as a finite union $X_0 b_0 \cup \cdots \cup X_m b_m$, where the X_i are rational (respectively recognizable) subsets of F_A, see [21]. Thus basic results such as Corollary 3.4 or Proposition 3.6 can be extended to virtually free groups (see [21, 51]). Similar generalizations can be obtained for free abelian groups of finite rank [51].

The fact that the strong properties of Corollary 3.4 do hold for both free groups and free abelian groups suggests considering the case of graph groups (also known as free partially abelian groups or right angled Artin groups), where we admit partial commutation between letters.

An *independence graph* is a finite undirected graph (A, I) with no loops, that is, I is a symmetric anti-reflexive relation on A. The *graph group* $G(A, I)$ is the quotient $F_A/\sim$, where $\sim$ denotes the congruence generated by the relation

$$\{(ab, ba) \mid (a, b) \in I\}.$$

On both extremes, we have $F_A = G(A, \emptyset)$ and the free abelian group on A, which corresponds to the complete graph on A. These turn out to be particular cases of *transitive* *forests*. We can say that (A, I) is a transitive forest if it has no induced subgraph of either of the following forms:



C_4 P_4

We recall that an induced subgraph of (A, I) is formed by a subset of vertices $A' \subseteq A$ and all the edges in I connecting vertices from A'.

The following difficult theorem, a group-theoretic version of a result on trace monoids by Aalbersberg and Hoogeboom [1], was proved in [27]:

Theorem 3.10 (Lohrey & Steinberg). *Let (A, I) be an independence graph. Then $G(A, I)$ has decidable rational subset membership problem if and only if (A, I) is a transitive forest.*

They also proved that these conditions are equivalent to decidability of the membership problem for finitely generated submonoids. Such a 'bad' $G(A, I)$ gives an example of a finitely presented group with a decidable generalized word problem that does not have a decidable membership problem for finitely generated submonoids.

It follows from Theorem 3.10 that any group containing a direct product of two free monoids has undecidable rational subset membership problem, a fact that can be directly deduced from the undecidability of the Post correspondence problem.

Other positive results on rational subsets have been obtained for graphs of groups, HNN extensions and amalgamated free products by Kambites, Silva and Steinberg [23], or Lohrey and Sénizergues [26]. Lohrey and Steinberg proved recently that the rational subset membership problem is recursively equivalent to the finitely generated submonoid membership problem for groups with two or more ends [28].

With respect to closure under complement, Lohrey and Sénizergues [26] proved that the class of groups for which the rational subsets form a boolean algebra is closed under HNN extension and amalgamated products over finite groups.

On the negative side, Bazhenova proved that rational subsets of finitely generated nilpotent groups do not form a boolean algebra, unless the group is virtually abelian [3]. Moreover, Roman′kov proved in [45], via a reduction from Hilbert's 10th problem, that the rational subset membership problem is undecidable for free nilpotent groups of any class $\geqslant 2$ of sufficiently large rank.

Last but not least, we should mention that Stallings' construction was successfully generalized in various directions: to fundamental groups of certain classes of graphs of groups (by Kapovich, Miasnikov and Weidmann [25]); to amalgamated free products of finite groups (by Markus-Epstein [33]); and to virtually free groups (by Silva, Soler-Escrivà and Ventura [54]).

3.5 Rational solution sets and rational constraints

In this final subsection we make a brief incursion in the brave new world of rational constraints. Rational subsets provide group theorists with two main assets:

- A concept which generalizes finite generation for subgroups and is much more fit to stand most induction procedures.
- A systematic way of looking for solutions of the *right type* in the context of equations of many sorts.

This second feature leads us to the notion of *rational constraint*, when we restrict the set of potential solutions to some rational subset. And there is a particular combination of circumstances that can ensure the success of this strategy: if $\mathrm{Rat}\, G$ is closed under intersection and we can prove that the solution set of problem P is an effectively computable rational subset of G, then we can solve problem P with any rational constraint.

An early example is the adaptation by Margolis and Meakin of Rabin's language and Rabin's tree theorem to free groups, where first-order formulae provide rational solution sets [31]. The logic language considered here is meant to be applied to words, seen as models, and consists basically of unary predicates that associate letters to positions in each word, as well as a binary predicate for position ordering. Margolis and Meakin used this construction to solve problems in combinatorial inverse semigroup theory [31].

Diekert, Gutierrez and Hagenah proved that the existential theory of systems of equations with rational constraints is solvable over a free group [12]. Working basically on a free monoid with involution, and adapting Plandowski's approach [39] in the process, they extended the classical result of Makanin [29] to include rational constraints, with much lower complexity as well. The complexity of their results has been recently improved in a paper by Diekert, Jeż and Plandowski [13], using the compression techniques developed by Jeż [22] for word equations.

The proof of this deep result is well out of scope here, but its potential applications are immense. Group theorists are only starting to discover its full strength.

The results in [26] can be used to extend the existential theory of equations with rational constraints to virtually free groups, a result that follows also from Dahmani and Guirardel's recent paper on equations over hyperbolic groups with quasi-convex rational constraints [11]. Equations over graph groups with a restricted class of rational constraints were also successfully considered by Diekert and Lohrey [14].

A somewhat exotic example of computation of a rational solution set arises in the problem of determining which automorphisms of F_2 (if any) carry a given word into a given finitely generated subgroup. The full solution set is recognized by a finite automaton; its vertices are themselves structures named "finite truncated automata" [56].

References

[1] I. J. Aalbersberg and H. J. Hoogeboom. Characterizations of the decidability of some problems for regular trace languages. *Math. Systems Theory*, 22:1–19, 1989. 782

[2] Algebraic Cryptography Center. *CRAG – the Cryptography and Groups Software Library*, 2010. 762

[3] G. A. Bazhenova. On rational sets in finitely generated nilpotent groups. *Algebra and Logic*, 39(4):215–223, 2000. Translated from *Algebra i Logika*, 39:379–394, 2000. 782

[4] M. Benois. Descendants of regular language in a class of rewriting systems: algorithm and complexity of an automata construction. In *Rewriting techniques and applications*, volume 256 of *Lecture Notes in Comput. Sci.*, pages 121–132. Springer-Verlag, 1987. 779

[5] J. Berstel. *Transductions and context-free languages*. B. G. Teubner, 1979. 775, 779

[6] J. Berstel, C. De Felice, D. Perrin, C. Reutenauer, and G. Rindone. Bifix codes and sturmian words. preprint, 2010. arXiv.org/pdf/1011.5369v2. 768

[7] M. Bestvina and M. Handel. Train tracks and automorphisms of free groups. *Ann. Math.*, 135:1–51, 1992. 775

[8] J.-C. Birget, S. W. Margolis, J. C. Meakin, and P. Weil. PSPACE-complete problems for subgroups of free groups and inverse finite automata. *Theoret. Comput. Sci.*, 242(1-2):247–281, 2000. 771

[9] O. Bogopolski and O. Maslakova. An algorithm for finding a basis of the fixed point subgroup of an automorphism of a free group. 2012. arXiv.org/pdf/1204.6728. 775

[10] R. V. Book and F. Otto. *String-Rewriting Systems*. Springer-Verlag, 1993. 761, 778

[11] F. Dahmani and V. Guirardel. Foliations for solving equations in groups: free, virtually free, and hyperbolic groups. *J. Topology*, 3(2):343–404, 2010. 783

[12] V. Diekert, C. Gutierrez, and C. Hagenah. The existential theory of equations with rational constraints in free groups is PSPACE-complete. *Inform. Comput.*, 202(2):105–140, 2005. 783

[13] V. Diekert, A. Jeż, and W. Plandowski. Finding all solutions of equations in free groups and monoids with involution. In E. A. Hirsch and J. E. Pin, editors, *Computer Science Symposium in Russia 2014, Conference Proceedings*, volume 8476 of *Lecture Notes in Computer Science*, pages 1–15. Springer-Verlag, 2014. 783

[14] V. Diekert and M. Lohrey. Word equations over graph products. *Internat. J. Algebra Comput.*, 18(3):493–533, 2008. 783

[15] M. J. Dunwoody. The accessibility of finitely presented groups. *Invent. Math.*, 81(3):449–457, 1985. 781

[16] J. Friedman. Sheaves on graphs, their homological invariants, and a proof of the Hanna Neumann conjecture: with an appendix by warren dicks. *Memoirs Amer. Math. Soc.*, 233(1100):xii + 106 pp, 2014. 770

[17] The GAP Group. *GAP – Groups, Algorithms, and Programming, Version 4.4.12*, 2008. 762

[18] S. M. Gersten. Intersections of finitely generated subgroups of free groups and resolutions of graphs. *Inventiones Math.*, 71(3):567–591, 1983. 770

[19] S. M. Gersten. On whitehead's algorithm. *Bull. Amer. Math. Soc.*, 10(2):281–284, 1984. 774

[20] R. Z. Goldstein and E. C. Turner. Fixed subgroups of homomorphisms of free groups. *Bull. Lond. Math. Soc.*, 18(5):468–470, 1986. 774

[21] Z. Grunschlag. *Algorithms in geometric group theory*. PhD thesis, University of California at Berkeley, 1999. 781

[22] A. Jeż. Recompression: a simple and powerful technique for word equations. In N. Portier and T. Wilke, editors, *STACS*, volume 20 of *LIPIcs*, pages 233–244, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 783

[23] M. Kambites, P. V. Silva, and B. Steinberg. On the rational subset problem for groups. *J. Algebra*, 309(2):622–639, 2007. 782

[24] I. Kapovich and A. Myasnikov. Stallings foldings and subgroups of free groups. *J. Algebra*, 248(2):608–668, 2002. 762, 771

[25] I. Kapovich, R. Weidmann, and A. Miasnikov. Foldings, graphs of groups and the membership problem. *Internat. J. Algebra Comput.*, 15(1):95–128, 2005. 782

[26] M. Lohrey and G. Sénizergues. Rational subsets in HNN-extensions and amalgamated products. *Internat. J. Algebra Comput.*, 18(1):111–163, 2008. 782, 783

[27] M. Lohrey and B. Steinberg. The submonoid and rational subset membership problems for graph groups. *J. Algebra*, 320(2):728–755, 2008. 782

[28] M. Lohrey and B. Steinberg. Submonoids and rational subsets of groups with infinitely many ends. To appear in *J. Algebra*, 2009. 782

[29] G. S. Makanin. Equations in a free group. *Math. USSR Izv.*, 21:483–546, 1983. Translated from *Izv. Akad. Nauk. SSR, Ser. Math.*, 46:1199–1273, 1983. 783

[30] S. W. Margolis and J. C. Meakin. Free inverse monoids and graph immersions. *Internat. J. Algebra Comput.*, 3(1):79–99, 1993. 762

[31] S. W. Margolis and J. C. Meakin. Inverse monoids, trees and context-free languages. *Trans. Amer. Math. Soc.*, 335(1):259–276, 1993. 783

[32] S. W. Margolis, M. V. Sapir, and P. Weil. Closed subgroups in pro-V topologies and the extension problem for inverse automata. *Internat. J. Algebra Comput.*, 11(4):405–446, 2001. 774

[33] L. Markus-Epstein. Stallings foldings and subgroups of amalgams of finite groups. *Internat. J. Algebra Comput.*, 17(8):1493–1535, 2007. 782

[34] A. Martino and E. Ventura. Fixed subgroups are compressed in free groups. *Commun. Algebra*, 32(10):3921–3935, 2004. 775

[35] A. Miasnikov, E. Ventura, and P. Weil. Algebraic extensions in free groups. In *Geometric group theory*, Trends Math., pages 225–253. Birkhäuser, 2007. 770, 771

[36] I. Mineyev. Groups, graphs and the Hanna Neumann Conjecture. *J. Topol. Anal.*, 4(1):1–12, 2012. 770

[37] D. E. Muller and P. E. Schupp. Groups, the theory of ends, and context-free languages. *J. Comput. System Sci.*, 26(3):295–310, 1983. 781

[38] H. Neumann. On the intersection of finitely generated free groups. addendum. *Publ. Math. (Debrecen)*, 5:128, 1957. 770

[39] W. Plandowski. Satisfiability of word equations with constants is in PSPACE. In *Proc. 40th Ann. Symp. Found. Comput. Sci.*, pages 495–500. IEEE Press, 1999. 783

[40] K. Reidemeister. Fundamentalgruppe und Überlagerungsräume. *J. Nachrichten Göttingen*, pages 69–76, 1928. 766

[41] J. Rhodes and B. Steinberg. *The q-theory of finite semigroups*. Springer-Verlag, 2009. 773

[42] L. Ribes and P. A. Zalesskii. On the profinite topology on a free group. *Bull. Lond. Math. Soc.*, 25:37–43, 1993. 773

[43] L. Ribes and P. A. Zalesskii. The pro-p topology of a free group and algorithmic problems in semigroups. *Internat. J. Algebra Comput.*, 4(3):359–374, 1994. 774

[44] A. Roig, E. Ventura, and P. Weil. On the complexity of the Whitehead minimization problem. *Internat. J. Algebra Comput.*, 17(8):1611–1634, 2007. 770, 774

[45] V. Roman'kov. On the occurrence problem for rational subsets of a group. In V. Roman'kov, editor, *International Conference on Combinatorial and Computational Methods in Mathematics*, pages 76–81, 1999. 782

[46] J. Sakarovitch. *Syntaxe des langages de Chomsky, essai sur le déterminisme*. PhD thesis, Université Paris VII, 1979. 780

[47] J. Sakarovitch. *Eléments de théorie des automates*. Vuibert, 2003. 775

[48] G. Sénizergues. Some decision problems about controlled rewriting systems. *Theoret. Comput. Sci.*, 71(3):281–346, 1990. 779

[49] G. Sénizergues. On the rational subsets of the free group. *Acta Informatica*, 33(3):281–296, 1996. 779, 780

[50] J.-P. Serre. *Arbres, amalgames,* SL_2. Société Mathématique de France, 1977. Avec un sommaire anglais; rédigé avec la collaboration de Hyman Bass; *Astérisque* 46. 762

[51] P. V. Silva. Recognizable subsets of a group: finite extensions and the abelian case. *Bull. European Assoc. Theor. Comput. Sci.*, 77:195–215, 2002. 781

[52] P. V. Silva. Free group languages: rational versus recognizable. *RAIRO Inform. Théor. App.*, 38(1):49–67, 2004. 779, 780

[53] P. V. Silva. Fixed points of endomorphisms of virtually free groups. *Pacific J. Math.*, 263(1):207–240, 2013. 775

[54] P. V. Silva, X. Soler-Escrivà, and E. Ventura. Finite automata for schreier graphs of virtually free groups. 2011. arXiv.org/pdf/1112.5709. 782

[55] P. V. Silva and P. Weil. On an algorithm to decide whether a free group is a free factor of another. *RAIRO Inform. Théor. App.*, 42:395–414, 2008. 770

[56] P. V. Silva and P. Weil. Automorphic orbits in free groups: words versus subgroups. *Internat. J. Algebra Comput.*, 20(4):561–590, 2010. 783

[57] P. V. Silva and P. Weil. On finite-index extensions of subgroups of free groups. *J. Group Theory*, 13(3):365–381, 2010. 771

[58] C. C. Sims. *Computation with finitely presented groups*. Cambridge University Press, 1994. 760, 762

[59] J. R. Stallings. Topology of finite graphs. *Inventiones Math.*, 71(3):551–565, 1983. 762, 772

[60] M. Takahasi. Note on chain conditions in free groups. *Osaka J. Math.*, 3(2):221–225, 1951. 770

[61] W. M. Touikan. A fast algorithm for Stallings' folding process. *Internat. J. Algebra Comput.*, 16(6):1031–1046, 2006. 765

[62] E. Ventura. On fixed subgroups of maximal rank. *Commun. Algebra*, 25(10):3361–3375, 1997. 775

[63] E. Ventura. Fixed subgroups of free groups: a survey. *Contemporary Math.*, 296:231–255, 2002. 775

[64] J. H. C. Whitehead. On equivalent sets of elements in a free group. *Ann. of Math. (2)*, 37(4):782–800, 1936. 774

Groups defined by automata

Laurent Bartholdi[1] *Pedro V. Silva*[2,*]

[1] Mathematisches Institut
Georg-August Universität zu Göttingen
Bunsenstraße 3–5
D-37073 Göttingen, Germany
email: laurent.bartholdi@gmail.com

[2] Centro de Matemática, Faculdade de Ciências
Universidade do Porto
R. Campo Alegre 687
4169-007 Porto, Portugal
email: pvsilva@fc.up.pt

2010 Mathematics Subject Classification: 20F65, 20E08, 20F10, 20F67, 68Q45

Key words: Automatic groups, word-hyperbolic groups, self-similar groups.

Contents

[*] The second author acknowledges support by Project ASA (PTDC/MAT/65481/2006) and C.M.U.P., financed by F.C.T. (Portugal) through the programmes POCTI and POSI, with national and E.U. structural funds.

Finite automata have been used effectively in recent years to define infinite groups. The two main lines of research have as their most representative objects the class of automatic groups (including "word-hyperbolic groups" as a particular case) and automata groups (singled out among the more general "self-similar groups").

The first approach is studied in Section 1 and implements in the language of automata some tight constraints on the geometry of the group's Cayley graph. Automata are used to define a normal form for group elements and to execute the fundamental group operations.

The second approach is developed in Section 2 and focuses on groups acting in a finitely constrained manner on a regular rooted tree. The automata define sequential permutations of the tree, and can even represent the group elements themselves.

The authors are grateful to Martin R. Bridson, François Dahmani, Rostislav I. Grigorchuk, Luc Guyot, and Mark V. Sapir for their remarks on a preliminary version of this text.

1 The geometry of the Cayley graph

Since its inception at the beginning of the 19th century, group theory has been recognized as a powerful language to capture *symmetries* of mathematical objects: crystals in the early 19th century, for Hessel and Frankenheim [56, page 120]; roots of a polynomial, for Galois and Abel; solutions of a differential equation, for Lie, Painlevé, etc. It was only later, mainly through the work of Klein and Poincaré, that the tight connections between group theory and geometry were brought to light.

Topology and group theory are related as follows. Consider a space X, on which a group G acts *freely*: for every $g \neq \mathbb{1} \in G$ and $x \in X$, we have $x \cdot g \neq x$. If the quotient space $Z = X/G$ is compact, then G "looks very much like" X, in the following sense: choose any $x \in X$, and consider the orbit $x \cdot G$. This identifies G with a roughly evenly distributed subset of X.

Conversely, consider a "nice" compact space Z with *fundamental group* G: then $X = \widetilde{Z}$, the *universal cover* of Z, admits a free G-action. In conclusion, properties of the fundamental group of a compact space Z reflect geometric properties of the space's universal cover.

We recall that finitely generated groups were defined in §23.1: they are groups G admitting a surjective map $\pi : F_A \twoheadrightarrow G$, where F_A is the free group on a finite set A.

Definition 1.1. A group G is *finitely presented* if it is finitely generated, say by $\pi : F_A \twoheadrightarrow G$, and if there exists a finite subset $\mathscr{R} \subset F_A$ such the kernel $\ker(\pi)$ is generated by the F_A-conjugates of $\mathscr{R}$, that is, $\ker(\pi) = \langle\langle \mathscr{R} \rangle\rangle$; one then has $G = F_A/\langle\langle \mathscr{R} \rangle\rangle$. These $r \in \mathscr{R}$ are called *relators* of the presentation; and one writes

$$G = \langle A \mid \mathscr{R} \rangle.$$

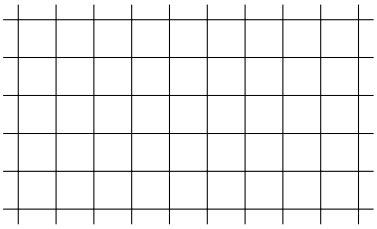
Sometimes it is convenient to write a relator in the form '$a = b$' rather than the more exact form 'ab^{-1}'.

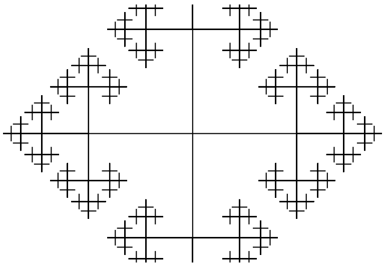
Let G be a finitely generated group, with generating set A. Its *Cayley graph* $\mathscr{C}(G, A)$, introduced by Cayley [47], is the graph with vertex set G and edge set $G \times A$; the edge (g, s) starts at vertex g and ends at vertex gs.

In particular, the group G acts freely on $\mathscr{C}(G, A)$ by left translation; the quotient $\mathscr{C}(G, A)/G$ is a graph with one vertex and $\operatorname{Card} A$ loops.

Assume moreover that G is finitely presented, with relator set $\mathscr{R}$. For each $r = r_1 \cdots r_n \in \mathscr{R}$ and each $g \in G$, the word r traces a closed path in $\mathscr{C}(G, A)$, starting at g and passing successively through $gr_1, gr_1r_2, \dots, gr_1r_2 \cdots r_n = g$. If one "glues" for each such r, g a closed disk to $\mathscr{C}(G, A)$ by identifying the disk's boundary with that path, one obtains a 2-dimensional cell complex in which each loop is contractible — this is a direct translation of the fact that the normal closure of $\mathscr{R}$ is the kernel of the presentation homomorphism $F_A \to G$.

For example, consider $G = \mathbb{Z}^2$, with generating set $A = \{(0, 1), (1, 0)\}$. Its Cayley graph is the standard square grid. The Cayley graph of a free group F_A, generated by A, is a tree.





More generally, consider a right G-set X, for instance the coset space $H\backslash G$. The *Schreier graph* $\mathscr{C}(G, X, A)$ of X is then the graph with vertex set X and edge set $X \times A$; the edge (x, s) starts in x and ends in xs.

1.1 History of geometric group theory

In a remarkable series of papers, Dehn [51, 52, 53], see also [54], initiated the geometric study of infinite groups, by trying to relate algorithmic questions on a group G and geometric questions on its Cayley graph. These problems were described in Definition 23.1.1, to which we refer. For instance, the word problem asks if one can determine whether a path in the Cayley graph of G is closed, knowing only the path's labels.

It is striking that Dehn used, for Cayley graph, the German *Gruppenbild*, literally "group picture". We must solve the word problem in a group G to be able to draw bounded portions of its Cayley graph; and some algebraic properties of G are tightly bound to the algorithmic complexity of the word problem, see §23.3.4. For example, Muller and Schupp prove (see Theorem 23.3.9) that a push-down automaton recognizes precisely the trivial elements of G if and only if G admits a free subgroup of finite index.

We consider now a more complicated example. Let $\mathcal{S}_g$ be an oriented surface of genus $g \geqslant 2$, and let J_g denote its fundamental group. Recall that $[x, y]$ denotes in a group the commutator $x^{-1}y^{-1}xy$. We have a presentation

$$J_g = \langle a_1, b_1, \dots, a_g, b_g \mid [a_1, b_1] \cdots [a_g, b_g] \rangle. \tag{1.1}$$

Let $r = [a_1, b_1] \cdots [a_g, b_g]$ denote the relator, and let $\mathscr{R}^*$ denote the set of cyclic permutations of $r^{\pm 1}$. The word problem in J_g is solvable in polynomial time by the following algorithm: let u be a given word. Freely reduce u by removing all aa^{-1} subwords. Then, if u contains a subword v_1 such that $v_1 v_2 \in \mathscr{R}^*$ and v_1 is longer than v_2, replace v_1 by v_2^{-1} in u and repeat. Eventually, u represents $\mathbb{1} \in G$ if and only if it is the empty word.

The validity of this algorithm relies on a lemma by Dehn, that every nontrivial word representing the identity contains more than half of the relator as a subword.

Incidentally, the Cayley graph of J_g is a tiling of the hyperbolic plane by $4g$-gons, with $4g$ meeting at each vertex.

Tartakovskiĭ [122], Greendlinger [70, 69] and Lyndon [101, 100] then devised "small cancellation" conditions on a group presentation that guarantee that Dehn's algorithm will succeed. Briefly said, they require the relators to have small enough overlaps. These conditions are purely combinatorial, and are described in §24.1.3.

Cannon and Thurston, on the other hand, sought a formalism that would encode the "periodicity of pictures" of a group's Cayley graph. Treating the graph as a metric space with geodesic distance d, already seen in §23.2.4, they make the following definition: the *cone type* of $g \in G$ is

$$C_g = \{h \in G \mid d(\mathbb{1}, gh) = d(\mathbb{1}, g) + d(g, gh)\}; \qquad (1.2)$$

the translate gC_g is the set of vertices that may be connected to $\mathbb{1}$ by a geodesic passing through g. Their intuition is that the cone type of a vertex v remembers, for points near v, whether they are closer or further to the origin than v; for example, $\mathbb{Z}^2$ with its standard generators has 9 cone types: the cone type of the origin (the whole plane), those of vertices on the axes (half-planes), and those of other vertices (quadrants).

Thurston's motivation was to get a good, algorithmic understanding of fundamental groups of threefolds. They should be made of nilpotent (or, more generally, solvable) groups on the one hand, and "automatic" groups on the other hand.

Definition 1.2. Let $G = \langle A \rangle$ be a finitely generated group, and recall that $\tilde{A}$ denotes $A \sqcup A^{-1}$. The *word metric* on G is the geodesic distance in G's Cayley graph $\mathscr{C}(G, A)$. It may be defined directly as

$$d(g, h) = \min\{n \mid g = hs_1 \cdots s_n \text{ with all } s_i \in \tilde{A}\},$$

and is left-invariant: $d(xg, xh) = d(g, h)$. The *ball of radius* n is the set

$$B_{G,A}(n) = \{g \in G \mid d(\mathbb{1}, g) \leqslant n\}.$$

The *growth function* of G is the function

$$\gamma_{G,A}(n) = \operatorname{Card} B_{G,A}(n).$$

The *growth series* of G is the power series

$$\Gamma_{G,A}(z) = \sum_{g \in G} z^{d(\mathbb{1}, g)} = \sum_{n \geqslant 0} \gamma_{G,A}(n) z^n (1 - z).$$

Growth functions are usually compared as follows: $\gamma \precsim \delta$ if there is a constant $C \in \mathbb{N}$ such that $\gamma(n) \leqslant \delta(Cn)$ for all $n \in \mathbb{N}$; and $\gamma \sim \delta$ if $\gamma \precsim \delta \precsim \gamma$. The equivalence class of $\gamma_{G,A}$ is independent of A.

Cannon observed (in an unpublished 1981 manuscript; see also [43]) that, if a group has finitely many cone types, then its growth series satisfies a finite linear system and is therefore a rational function of z. For J_g, for instance, he computes

$$\Gamma_{J_g,A} = \frac{1+2z+\cdots+2z^{2g-1}+z^{2g}}{1+(2-4g)z+\cdots+(2-4g)z^{2g-1}+z^{2g}}.$$

This notion was formalized by Thurston in 1984 using automata, and is largely the topic of the next section. We will return to growth of groups in §24.2.5; see however [29] for a good example of growth series of groups computed thanks to a description of the Cayley graph by automata.

Gromov emphasized the relevance to group theory of the following definition, attributed to Margulis:

Definition 1.3 ([84]). A map $f : X \to Y$ between two metric spaces is a C*-quasi-isometry*, for a constant $C > 0$, if one has

$$C^{-1}d(x,y) - C \leqslant d(f(x), f(y)) \leqslant Cd(x,y) + C$$

for all $y \in Y$ such that $d(f(X), y) \leqslant C$. A *quasi-isometry* is a C-quasi-isometry for some $C > 0$. Two spaces are *quasi-isometric* if there exists a quasi-isometry between them; this is an equivalence relation.

A property of finitely generated groups is *geometric* if it only depends on the quasi-isometry class of its Cayley graph.

Thus for instance the inclusion $\mathbb{Z} \to \mathbb{R}$, and the map $\mathbb{R} \to \mathbb{Z}, x \mapsto \lfloor x \rfloor$ are quasi-isometries.

Being finite, having a finite-index subgroup isomorphic to $\mathbb{Z}$, and being finitely presented are geometric properties. The asymptotics of the growth function is also a geometric invariant; thus for instance having growth function $\precsim n^2$ is a geometric property.

1.2 Automatic groups

Let $G = \langle A \rangle$ be a finitely generated group. We will consider the formal alphabet $\hat{A} = A \sqcup A^{-1} \sqcup \{\mathbb{1}\}$, where $\mathbb{1}$ is treated as a "padding" symbol. Following the main reference [57] by Epstein *et al.*:

Definition 1.4 ([57, 58, 25]). The group G is *automatic* if there are finite-state automata $\mathcal{L}, \mathcal{M}$, the *language* and *multiplication* automata, with the following properties:

(i) $\mathcal{L}$ is an automaton with alphabet $\tilde{A}$;

(ii) $\mathcal{M}$ has alphabet $\hat{A} \times \hat{A}$, and has for each $s \in \hat{A}$ an accepting subset T_s of states; call $\mathcal{M}_s$ the automaton with accepting states T_s;

(iii) the language of $\mathcal{L}$ surjects onto G by the natural map $f : \tilde{A} \to F_A \to G$; words in $L(\mathcal{L})$ are called *normal forms*;

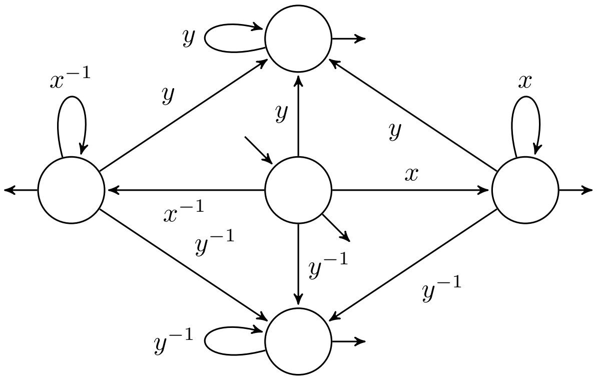
(iv) for any two normal forms $u, v \in L(\mathcal{L})$, consider the word

$$w = (u_1, v_1)(u_2, v_2)\cdots(u_n, v_n) \in (\hat{A} \times \hat{A})^*,$$

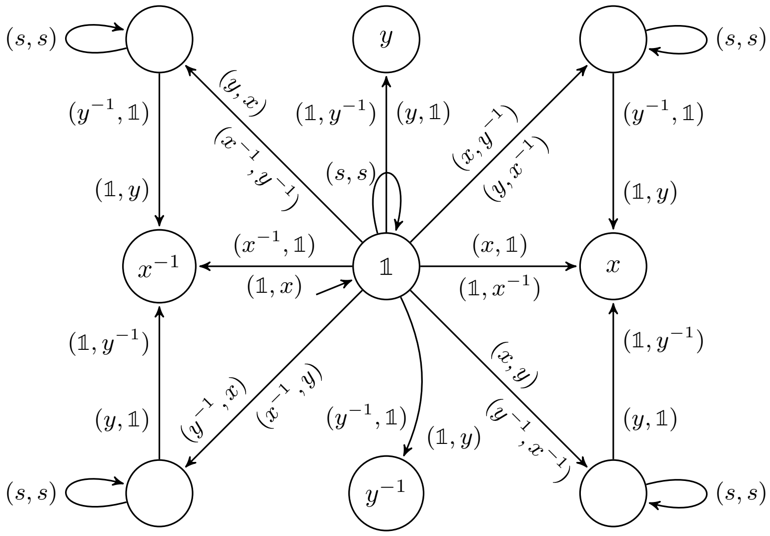
where $n = \max\{|u|, |v|\}$ and $u_i, v_j = \mathbb{1}$ if $i > |u|, j > |v|$. Then $\mathcal{M}_s$ accepts w if and only if $\pi(u) = \pi(vs)$.

In words, G is automatic if the automaton $\mathcal{L}$ singles out sufficiently many words which may be used to represent all group elements; and the automaton $\mathcal{M}_s$ recognizes when two such singled out words represent group elements differing by a generator. The pair $(\mathcal{L}, \mathcal{M})$ is an *automatic structure* for G.

We will give numerous examples of automatic groups in §24.1.3. Here is a simple one that contains the main features: the group $G = \mathbb{Z}^2$, with standard generators x, y. The language accepted by $\mathcal{L}$ is $(x^* \cup (x^{-1})^*)(y^* \cup (y^{-1})^*)$:



The multiplication automaton, in which states in T_s are labeled s, is



The definition we gave is purely automata-theoretic. It does, however, have a more geometric counterpart. A word $w \in \tilde{A}^*$ represents in a natural way a path in the Cayley graph $\mathscr{C}(G, A)$, starting at $\mathbb{1}$ and ending at $\pi(w)$. If $w = w_1 \cdots w_n$, we write $w(j) = w_1 \cdots w_j$ the vertex of $\mathscr{C}(G, A)$ reached after j steps; if $j > n$ then $w(j) = w$. For two paths $u, v \in \tilde{A}^*$, we say they *k-fellow-travel* if $d(u(j), v(j)) \leqslant k$ for all $j \in \{1, \dots, \max\{|u|, |v|\}\}$.

Proposition 1.1. *A group G is automatic if and only if there exists a rational language $L \subseteq \tilde{A}^*$, mapping onto G, and a constant k, such that for any $u, v \in L$ with $d(\pi(u), \pi(v)) \leqslant 1$ the paths u, v k-fellow-travel.*

Sketch of proof. Assume first that G has automatic structure $(\mathcal{L}, \mathcal{M})$, and let c denote the number of states of $\mathcal{M}$. If $u, v \in L(\mathcal{L})$ satisfy $\pi(u) = \pi(vs)$, let s_j denote the state $\mathcal{M}$ is in after having read $(u_1, v_1) \cdots (u_j, v_j)$. There is a path of length $< c$, in $\mathcal{M}$, from s_j to an accepting state (labeled s); let its label be (p, q). Then $\pi(u(j)p) = \pi(v(j)qs)$, so $u(j)$ and $v(j)$ are at distance at most $2c - 1$ in $\mathscr{C}(G, A)$.

Conversely, assume that paths k-fellow-travel and that an automaton $\mathcal{L}$, with state set Q is given, with language surjecting onto G. Recall that $B(k)$ denotes the set of group elements at distance $\leqslant k$ from $\mathbb{1}$ in $\mathscr{C}(G, A)$. Consider the automaton with state set $Q \times Q \times B_k$. Its initial state is $(*, *, \mathbb{1})$, where $*$ is the initial state of $\mathcal{L}$; its alphabet is $\hat{A} \times \hat{A}$, and its transitions are given by $(p, q, g) \cdot (s, t) = (p \cdot s, q \cdot t, s^{-1}gt)$ whenever these are defined. Its accepting set of states, for $s \in \hat{A}$, is $T_s = Q \times Q \times \{s\}$. □

Corollary 1.2. *If the finitely generated group $G = \langle A \rangle$ is automatic, and if B is another finite generating set for G, then there also exists an automatic structure for G using the alphabet B.*

Sketch of proof. Note first that a trivial generator may be added or removed from A or B, using an appropriate finite transducer for the latter.

There exists then $M \in \mathbb{N}$ such that every $a \in \tilde{A}$ can be written as a word $w_a \in \tilde{B}^*$ of length precisely M. Accept as normal forms all $w_{a_1} \cdots w_{a_n}$ such that $a_1 \cdots a_n$ is a normal form in the original automatic structure $\mathcal{L}$. The new normal forms constitute a homomorphic image of $\mathcal{L}$ and therefore define a rational language. If paths in $L(\mathcal{L})$ k-fellow-travel, then their images in the new structure will kM-fellow-travel. □

Note that the language of normal forms is only required to contain "enough" expressions; namely that the evaluation map $L(\mathcal{L}) \to G$ is onto. We may assume that it is bijective, by the following lemma. The language $L(\mathcal{L})$ is then called a "rational cross-section" by Gilman [64]; and $(\mathcal{L}, \mathcal{M})$ is called an *automatic structure with uniqueness*.

Lemma 1.3. *Let G be an automatic group. Then G admits an automatic structure with uniqueness.*

Sketch of proof. Consider $(\mathcal{L}', \mathcal{M})$ an automatic structure. Recall the "short-lex" ordering on words: $u \leqslant v$ if $|u| < |v|$, or if $|u| = |v|$ and u comes lexicographically before v. The language $\{(u, v) \in \hat{A}^* \times \hat{A}^* \mid u \leqslant v\}$ is rational. The language

$$L = L(\mathcal{L}') \cap \{u \in \tilde{A}^* \mid \text{ for all } v \in \hat{A}^*, \text{ if } (u, v) \in L(\mathcal{M}_{\mathbb{1}}) \text{ then } u \leqslant v\}$$

is then also rational, of the form $L(\mathcal{L})$. The automaton $\mathcal{M}$ need not be changed. □

Various notions related to automaticity have emerged, some stronger, some weaker:

- One may require the words accepted by $\mathcal{L}$ to be representatives of minimal length; the automatic structure is then called *geodesic*. It would then follow that the growth series $\Gamma_{G,A}(z)$ of G, which is the growth series of $\mathcal{L}$, is a rational function. Note that there is a constant K such that, for the language produced by Lemma 1.3, all words $u \in L(\mathcal{L})$ satisfy $|u| \leqslant Kd(\mathbb{1}, \pi(u))$.

- The definition is asymmetric; more precisely, we have defined a *right automatic* group, in that the automaton $\mathcal{M}$ recognizes multiplication on the right. One could similarly define *left automatic groups*; then a group is right automatic if and only if it is left automatic.

  Indeed, let $(\mathcal{L}, \mathcal{M})$ be an automatic structure where $\mathcal{L}$ recognizes a rational cross section. Then $L' = \{u^{-1} \mid u \in L(\mathcal{L})\}$ and $M' = \{(u^{-1}, v^{-1}) \mid (u,v) \in L(\mathcal{M})\}$ are again rational languages. Indeed, since rational languages are closed under reversal and morphisms, it follows easily that L' is rational. On the other hand, using the pumping lemma and the fact that group elements admit unique representatives in $L(\mathcal{L})$, the amount of padding at the end of word-pairs in $L(\mathcal{M})$ is bounded, and can be moved from the beginning to the end of the word-pairs in M' by a finite transducer. Therefore, L', M' are the languages of a right automatic structure.

  However, one could require both properties simultaneously, namely, on top of an automatic structure, a third automaton $\mathcal{N}$ accepting (in state $s \in \hat{A}$) all pairs of normal forms (u, v) with $\pi(u) = \pi(sv)$. Such groups are called *biautomatic*. No example is known of a group that is automatic but not biautomatic.
- One might also only keep the geometric notion of "combing": a *combing* on a group is a choice, for every $g \in G$, of a word $w_g \in \tilde{A}^*$ evaluating to g, such that the words w_g and w_{gs} fellow-travel for all $g \in G$, $s \in \tilde{A}$.

  In that sense, a group is automatic if and only if it admits a combing whose words form a rational language; see [32] for details.

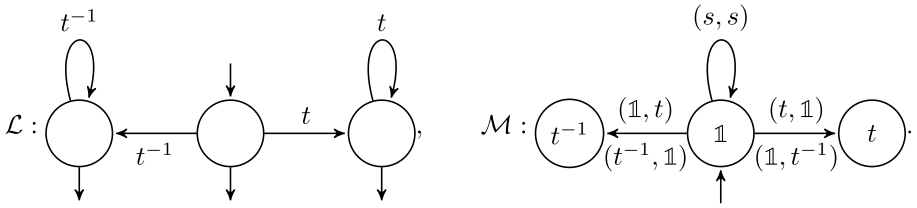
  One may again require the combing lines to be geodesics, i.e., words of minimal length; see Hermiller's work [91, 89, 90].

  One may also put weaker constraints on the words of the combing; for example, require it to be an indexed language. Bridson and Gilman [33] proved that all geometries of threefolds, in particular the Nil (1.3) and Sol geometry, which are not automatic, fall in this framework.
- Another relaxation is to allow the automaton $\mathcal{M}$ to read at will letters from the first or the second word; groups admitting such a structure are called *asynchronously automatic*. Among fundamental groups of threefolds, there is no difference between these definitions [33], but for more general groups there is.
- Finally, Definition 1.4 can be adapted to define automatic semigroups. Properties from automatic groups that can be proved within the automata-theoretic framework can often be generalized to automatic semigroups, or at least monoids [42]. However, establishing an alternative geometric approach has proved to be a tough task and success was reached only in restricted cases [120, 93].

1.3 Main examples of automatic groups

From the very definition, it is clear that finite groups are automatic: one chooses a word representing each group element, and these necessarily form a fellow-travelling rational language.

It is also clear that $\mathbb{Z}$ is automatic: write t for the canonical generator of $\mathbb{Z}$; the language $t^* \cup (t^{-1})^*$ maps bijectively to $\mathbb{Z}$; and the corresponding paths 1-fellow-travel. The automata are



Simple constructions show that the direct and free products of automatic groups are again automatic. Finite extensions and finite-index subgroups of automatic groups are automatic. It is however still an open problem whether a direct factor of an automatic group is automatic.

Recall that we glued disks, one for each $g \in G$ and each $r \in \mathscr{R}$, to the Cayley graph of a finitely presented group $G = \langle A \mid \mathscr{R} \rangle$, so as to obtain a 2-complex $\mathscr{K}$. The *small cancellation conditions* express a combinatorial form of non-positive curvature of $\mathscr{K}$: roughly, $C(p)$ means that every proper edge cycle in $\mathscr{K}$ has length $\geqslant p$, and $T(q)$ means that every proper edge cycle in the dual $\mathscr{K}^\vee$ has length $\geqslant q$; see [101, Chapter V] for details. If G satisfies $C(p)$ and $T(q)$ where $p^{-1} + q^{-1} \leqslant \frac{1}{2}$, then G is automatic.

Consider the configurations defined by n strings in $\mathbb{R}^2 \times [0,1]$, with string $\#i$ starting at $(i,0,0)$ and ending at $(i,0,1)$; these configurations are viewed up to isotopy preserving the endpoints. They can be multiplied (by stacking them above each other) and inverted (by flipping them up-down), yielding a group, the *pure braid group*; if the strings are allowed to end in an arbitrary permutation, one obtains the *braid group*. This group B_n is generated by elementary half-twists of strings $\#i, i+1$ around each other, and admits the presentation

$$B_n = \langle \sigma_1, \dots, \sigma_{n-1} \mid \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}, [\sigma_i, \sigma_j] \text{ whenever } |i-j| \geqslant 2 \rangle.$$

More generally, consider a surface $\mathcal{S}$ of genus g, with n punctures and b boundary components. The *mapping class group* $M_{g,n,b}$ is the group of maps $\mathcal{S} \to \mathcal{S}$ modulo isotopy, and B_n is the special case $M_{0,n,1}$ of mapping classes of the n-punctured disk. All mapping class groups $M_{g,n,b}$ are automatic groups [107].

As another generalization of braid groups, consider *Artin groups*. Let (m_{ij}) be a symmetric $n \times n$-matrix with entries in $\mathbb{N} \cup \{\infty\}$. The *Artin group* of type (m_{ij}) is the group with presentation

$$A(m) = \langle s_1, \dots, s_n \mid (s_i s_j)^{\lfloor m_{ij}/2 \rfloor} = (s_j s_i)^{\lfloor m_{ij}/2 \rfloor} \text{ whenever } m_{ij} < \infty \rangle.$$

The corresponding *Coxeter group* has presentation

$$C(m) = \langle s_1, \dots, s_n \mid s_i^2, (s_i s_j)^{m_{ij}/2} = (s_j s_i)^{m_{ij}/2} \text{ whenever } m_{ij} < \infty \rangle.$$

An Artin group $A(m)$ has *finite type* if $C(m)$ is finite. Artin groups of finite type are biautomatic [48]. Coxeter groups are automatic [35].

Fundamental groups of threefolds, except those with a piece modelled on Nil or Sol geometry [57, chapter 12], are automatic.

1.4 Properties of automatic groups

The definition of automatic groups, by automata, has a variety of interesting consequences. First, automatic groups are finitely presented; more generally, combable groups are finitely presented:

Proposition 1.4 ([4]). *Let G be a combable group. Then G has type F_∞, namely, there exists a contractible cellular complex with free G-action and finitely many G-orbits of cells in each dimension.*

(Finite presentation is equivalent to "finitely many G-orbits of cells in dimension $\leqslant 2$").

Sketch of proof. By assumption, G is finitely generated. Therefore, the Cayley graph contains one G-orbit of 0-cells (vertices), and $\operatorname{Card} A$ orbits of 1-cells (edges). Consider all pairs of paths u, v in the combing that have neigbouring extremities. They k-fellow-travel by hypothesis; so there are for all j paths $w(j)$ of length $\leqslant k$ connecting $u(j)$ to $v(j)$. The closed paths $u(j) - v(j) - v(j+1) - u(j+1) - u(j)$ have length $\leqslant 2k+2$, so they trace finitely many words in F_A. Taking them as relators defines a finite presentation for G. The process may be continued with higher-dimensional cells. □

Proposition 1.5. *Automatic groups satisfy a* quadratic isoperimetric inequality*; that is, for any finite presentation $G = \langle A \mid \mathscr{R} \rangle$ there is a constant k such that, if $w \in F_A$ is a word evaluating to $\mathbb{1}$ in G, then*

$$w = \prod_{i=1}^{\ell} r_i^{w_i} \text{ for some } r_i \in \mathscr{R}^{\pm 1}, w_i \in F_A \text{ and } \ell \leqslant k|w|^2.$$

Sketch of proof. Write $n = |w|$, and draw the combing lines between $\mathbb{1}$ and $w(j)$. There are n combing lines, which have length $\mathcal{O}(n)$; so the gap between neighbouring combing lines can be filled by $\mathcal{O}(n)$ relators. This gives $\mathcal{O}(n^2)$ relators in total. □

Note that being finitely presented is usually of little value as far as algorithmic questions are concerned: there are finitely presented groups whose word problem cannot be solved by a Turing machine [110, 27]. By contrast:

Proposition 1.6. *The word problem in a group given by an automatic structure is solvable in quadratic time. A word may even be put into canonical form in quadratic time.*

Sketch of proof. We may assume, by Lemma 1.3, that every $g \in G$ admits a unique normal form. Now, given a word $u = a_1 \cdots a_n \in \hat{A}^*$, construct the following words: $w_0 \in L(\mathcal{L})$ is the representative of $\mathbb{1}$. Treating $\mathcal{M}_a$ as a non-deterministic automaton in its second variable, find for $i = 1, \dots, n$ a word $w_i \in \hat{A}^*$ such that the padding of (w_{i-1}, w_i) is accepted by $\mathcal{M}_{a_i}$. Then $\pi(u) = \mathbb{1} \in G$ if and only if $w_n = w_0$.

Clearly the w_i have linear length in i, so the total running time is quadratic in n. □

In general, finitely generated subgroups and quotients of automatic groups need not be automatic — they need not even be finitely presented. A subgroup H of a finitely generated group $G = \langle A\rangle$ is *quasi-convex* if there exists a constant δ such that every $h \in H$ is connected to $\mathbb{1} \in G$ by a geodesic in $\mathscr{C}(G, A)$ that remains at distance $\leqslant \delta$ from H. Typical examples are finite-index subgroups, free factors, and direct factors.

On the other hand, a subgroup H of an automatic group G with language $L(\mathcal{L})$ is *$\mathcal{L}$-rational* if the full preimage of H in $L(\mathcal{L})$ is rational. The following is easy but fundamental:

Lemma 1.7 ([61]). *A subgroup H of an automatic group is quasi-convex if and only if it is $\mathcal{L}$-rational.*

It is still unknown whether automatic groups have solvable conjugacy problem; however, there are asynchronously automatic groups with unsolvable conjugacy problem, for instance appropriate amalgamated products of two free groups over finitely generated subgroups. These groups are asynchronously automatic [25, Theorem E], and have unsolvable conjugacy problem [104].

Theorem 1.8 (Gersten-Short). *Biautomatic groups have solvable conjugacy problem.*

Sketch of proof; see [62]. Consider two words $x, y \in \tilde{A}^*$. Using the biautomatic structure, the language

$$C(x, y) = \{(u, v) \in \hat{A}^* \times \hat{A}^* \mid u, v \in \mathcal{L} \text{ and } \pi(u) = \pi(xvy)\}$$

is rational. Now x, y are conjugate if and only if $C(x^{-1}, y) \cap \{(w, w) \mid w \in \mathcal{L}\}$ is nonempty. The problem of deciding whether a rational language is empty is algorithmically solvable. □

In fact, the centralizer of an element of a biautomatic group is a quasi-convex subgroup, and is thus biautomatic [61] (but we remark that it is still unknown whether a quasi-convex subgroup of an automatic group is necessarily automatic). There is therefore a good algorithmic description of *all* elements that conjugate x to y.

1.5 Word-hyperbolic groups

Gromov [85] introduced the fundamental concept of "negative curvature" to group theory. This goes further in the direction of viewing groups as metric spaces, through the geodesic distance on their Cayley graph. The definition is given for *geodesic* metric spaces, i.e., metric spaces in which any two points can be joined by a geodesic segment:

Definition 1.5 ([63, 5, 49]). Let X be a geodesic metric space, and let $\delta > 0$ be given. The space X is *δ-hyperbolic* if, for any three points $A, B, C \in X$ and geodesics arcs a, b, c joining them, every $P \in a$ is at distance at most δ from $b \cup c$.

The space X is *hyperbolic* if it is δ-hyperbolic for some δ. The finitely generated group $G = \langle A\rangle$ is *word-hyperbolic* if it acts by isometries on a hyperbolic metric space X with discrete orbits, finite point stabilizers, and compact quotient X/G.

Equivalently, G is word-hyperbolic if and only if $\mathscr{C}(G, A)$ is hyperbolic.

Gilman [65] gave a purely automata-theoretic definition of word-hyperbolic groups: G is word-hyperbolic if and only if, for some regular combing $\mathcal{M} \subset \hat{A}^*$, the language $\{u\mathbb{1}v\mathbb{1}w \mid u, v, w \in \mathcal{M}, \pi(uvw) = \mathbb{1}\} \subset \hat{A}^*$ is context-free. Using the geometric definition, we note immediately the following examples: first, the hyperbolic plane $\mathbb{H}^2$ is hyperbolic (with $\delta = \log 3$); so is $\mathbb{H}^n$. Any discrete, cocompact group of isometries of $\mathbb{H}^n$ is word-hyperbolic. This applies in particular to the surface group J_g from (1.1), if $g \geqslant 2$. Note however that some word-hyperbolic groups are not small cancellation groups, for instance because for small cancellation groups the complex in Proposition 1.4 has trivial homology in dimension $\geqslant 3$; yet the complex associated with a cocompact group acting on $\mathbb{H}^n$ has infinite cyclic homology in degree n (see [60] for applications of topology to group theory).

It is also possible to define δ-hyperbolicity for spaces X that are not geodesic (such as, e.g., a discrete group):

Definition 1.6. Let X be a metric space, and let $\delta' > 0$ be given. The space X is δ'-*hyperbolic* if, for any four points $A, B, C, D \in X$, the numbers

$$\{d(A,B) + d(C,D), d(A,C) + d(B,D), d(A,D) + d(B,C)\}$$

are such that the largest two differ by at most δ'.

Word-hyperbolic groups arise naturally in geometry, in the following way: let $\mathcal{M}$ be a compact Riemannian manifold with negative (not necessarily constant) sectional curvature. Then $\pi_1(\mathcal{M})$ is a word-hyperbolic group.

Word-hyperbolic groups are also "generic" among finitely-presented groups, in the following sense: fix a number k of generators, and a constant $\epsilon \in [0, 1]$. For large N, there are $\approx (2k-1)^N$ words of length $\leqslant N$ in F_k; choose a subset $\mathscr{R}$ of size $\approx (2k-1)^{\epsilon N}$ of them uniformly at random, and consider the group G with presentation $\langle A \mid \mathscr{R}\rangle$.

Then, with probability $\to 1$ as $N \to \infty$, the group G is word-hyperbolic. Furthermore, if $\epsilon < \frac{1}{2}$, then with probability $\to 1$ the group G is infinite, while if $\epsilon > \frac{1}{2}$, then with probability $\to 1$ the group G is trivial [111].

Word-hyperbolic groups provide us with a large number of examples of automatic groups. Better:

Theorem 1.9 (Gersten-Short, Gromov). *Let G be a word-hyperbolic group. Then G is biautomatic. Moreover, the normal form $\mathcal{L}$ may be chosen to consist of geodesics.*

Sketch of proof. In a δ-word-hyperbolic group G, geodesics $(2\delta + 1)$-fellow-travel. On the other hand, G has a finite number of cone types (1.2), so the language of geodesics is rational, recognized by an automaton with as many states as there are cone types. □

In fact, the automatic structure is, in some precise sense, unique [31]. Calegari and Fujiwara [41] show that, given two finite generating sets A, B for a word-hyperbolic group G, there exists an algebraic number λ and a rational transduction $f : \hat{A}^* \to \hat{B}^*$ that converts a geodesic normal form for A into a geodesic normal form for B, such that $|f(w)| - \lambda|w| = \mathcal{O}(\sqrt{|w|})$.

Hyperbolic spaces X have a natural *hyperbolic boundary* ∂X: fix a point $x_0 \in X$, and consider *quasi-geodesics at* x_0, namely quasi-isometric embeddings $\gamma : \mathbb{N} \to X$ starting at x_0. Declare two such quasi-geodesics γ, δ to be equivalent if $d(\gamma(n), \delta(n))$ is bounded. The set of equivalence classes, with its natural topology, is the boundary ∂X of X. The fundamental tool in studying hyperbolic spaces is the following

Lemma 1.10 (Morse). *Let X be a hyperbolic space and let C be a constant. There is then a constant D such that all C-quasi-geodesics between two points $x, y \in X$ are at distance at most D from one another.*

The hyperbolic boundary ∂X is compact, under appropriate conditions satisfied e.g. by $X = \mathscr{C}(G, A)$, and $X \cup \partial X$ is a compactification of X. Now, in that case, the automaton $\mathcal{L}$ provides a symbolic coding of ∂X as a finitely presented shift space (where the shift action is the "geodesic flow", following one step along infinite paths $\in \hat{A}^\infty$ representing quasi-geodesics).

We note that, for word-hyperbolic groups, the word and conjugacy problem admit extremely efficient solutions: they are both solvable in linear time by a Turing machine. The word problem is actually solvable in real time, namely with a bounded amount of calculation allowed between inputs [94]. The isomorphism problem is decidable for word-hyperbolic groups, say given by a finite presentation [50]. Word-hyperbolic groups also satisfy a linear isoperimetric inequality, in the sense that every $w \in F_A$ that evaluates to $\mathbb{1}$ in G is a product of $\mathcal{O}(|w|)$ conjugates of relators. Better:

Proposition 1.11. *A finitely presented group is word-hyperbolic if and only if it satisfies a linear isoperimetric inequality, if and only if it satisfies a subquadratic isoperimetric inequality.*

Note that the generalized word problem is known to be unsolvable [113], but the order problem is on the other hand solvable in word-hyperbolic groups [28]. It follows that the generalized word problem is unsolvable for automatic groups as well.

There are important weakenings of the definition of word-hyperbolic groups; we mention two. A *bicombing* is a choice, for every pair of vertices $g, h \in \mathscr{C}(G, A)$, of a path $\ell_{g,h}$ from g to h. Since G acts by left-translation on $\mathscr{C}(G, A)$, it also acts on bicombings. A bicombing satisfies the *k-fellow-traveller property* if for any neighbours x' of x and y' of y, the paths $\ell_{x,y}$ and $\ell_{x',y'}$ k-fellow-travel.

A *semi-hyperbolic group* is a group admitting an invariant bicombing by fellow-travelling words. See [34], or the older paper [6]. In particular, biautomatic, and therefore word-hyperbolic, groups are semi-hyperbolic.

Semi-hyperbolic groups are finitely presented and have solvable word and conjugacy problems. In fact, they even have the "monotone conjugation property", namely, if g and h are conjugate, then there exists a word w with $g^{\pi(w)} = h$ and $|g^{\pi(w(i))}| \leqslant \max\{|g|, |h|\}$ for all $i \in \{0, \dots, |w|\}$.

A group G is *relatively hyperbolic* [59] if it acts properly discontinuously on a hyperbolic space X, preserving a family $\mathcal{H}$ of separated horoballs, such that $(X \setminus \mathcal{H})/G$ is compact. All fundamental groups of finite-volume negatively curved manifolds are relatively hyperbolic.

A non-closed manifold has "cusps", going off to infinity, whose interpretation in the fundamental group are conjugacy classes of loops that may be homotoped arbitrarily far into the cusp. Farb [59] captures combinatorially the notion of relative hyperbolicity as follows: let $\mathscr{H}$ be a family of subgroups of a finitely generated group $G = \langle A \rangle$. Modify the Cayley graph of G as follows: for each coset gH of a subgroup $H \in \mathscr{H}$, add a vertex gH, and connect it by an edge to every $gh \in \mathscr{C}(G, A)$, for all $h \in H$. In addition, require that every edge in $\widehat{\mathscr{C}(G, A)}$ belong to only finitely many simple loops of any given length. The group G is *weakly relatively hyperbolic*, relative to the family $\mathscr{H}$, if that modified Cayley graph $\widehat{\mathscr{C}(G, A)}$ is a hyperbolic metric space.

By virtue of its geometric characterization, being word-hyperbolic is a geometric property in the sense of Definition 1.3 (though beware that being hyperbolic is preserved under quasi-isometry only if the metric spaces are geodesic). Being combable and being bicombable are also geometric.

We finally remark that a notion of word-hyperbolicity has been defined for semigroups [92, 55]; the definition uses context-free languages. As for automatic (semi)groups, the theory does not seem uniform enough to warrant a simultaneous treatment of groups and semigroups; again, there is no clear geometric counterpart to the definition of word-hyperbolic semigroups — except in particular cases, such as monoids defined through special confluent rewriting systems [46].

1.6 Non-automatic groups

All known examples of non-automatic groups arise as groups violating some interesting consequence of automaticity.

First, infinitely presented groups cannot be automatic. There are uncountably many finitely generated groups, and only countably many finitely presented groups; therefore automatic groups should be thought of as the rationals among the real numbers.

Groups with unsolvable word problem cannot be automatic.

If a nilpotent group is automatic, then it contains an abelian subgroup of finite index [66]; therefore, for instance, the discrete Heisenberg group

$$G = \begin{pmatrix} 1 & \mathbb{Z} & \mathbb{Z} \\ 0 & 1 & \mathbb{Z} \\ 0 & 0 & 1 \end{pmatrix} = \langle x, y \mid [x, [x, y]], [y, [x, y]] \rangle \tag{1.3}$$

is not automatic. Note also that G satisfies a cubic, but no quadratic, isoperimetric inequality.

Many solvable groups have larger-than-quadratic isoperimetric functions; they therefore cannot be automatic [86]. This applies in particular to the Baumslag-Solitar groups

$$BS_{1,n} = \langle a, t \mid a^n = a^t \rangle. \tag{1.4}$$

Similarly, $\mathrm{SL}_n(\mathbb{Z})$, for $n \geqslant 3$, or $\mathrm{SL}_n(\mathcal{O})$ for $n \geqslant 2$, where $\mathcal{O}$ are the integers in an imaginary number field, are not automatic.

Infinite, finitely generated torsion groups cannot be automatic: they cannot admit a rational normal form, because of the pumping lemma. We will see examples, due to Grigorchuk and Gupta-Sidki, in §24.2.1.

There are combable groups that are not automatic [30], for instance

$$G = \langle a_i, b_i, t_i, s \mid t_1 a_1 = t_2 a_2, [a_i, s] = [a_i, t_i] = [b_i, s] = [b_i, t_i] = \mathbb{1} \quad (i = 1, 2) \rangle,$$

which satisfies only a cubic isoperimetric inequality. Finitely presented subgroups of automatic groups need not be automatic [24].

The following group is asynchronously automatic, but is not automatic: it does not satisfy a quadratic isoperimetric inequality [25, §11]:

$$G = \langle a, b, t, u \mid a^t = ab, b^t = a, a^u = ab, b^u = a \rangle.$$

2 Groups generated by automata

We now turn to another important class of groups related to finite-state automata. These groups act by permutations on a set A^* of words, and these permutations are represented by *Mealy automata*. These are deterministic, initial finite-state transducers $\mathcal{M}$ with input and output alphabet A, that are complete with respect to input; in other words,

$$\text{At every state and for each } a \in A \text{, there is a unique outgoing edge with input } a. \tag{2.1}$$

The automaton $\mathcal{M}$ defines a transformation of A^*, which extends to a transformation of A^ω, as follows. Given $w = a_1 a_2 \cdots \in A^* \cup A^\omega$, there is by (2.1) a unique path in $\mathcal{M}$ starting at the initial state and with input labels w. The image of w under the transformation is the output label along that same path.

Definition 2.1. A map $f : A^* \to A^*$ is *automatic* if f is produced by a finite-state automaton as above.

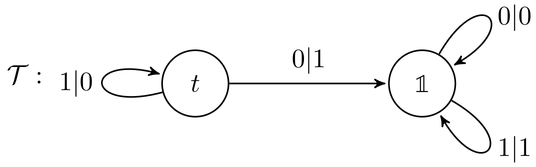
One may forget the initial state of $\mathcal{M}$, and consider the set of all transformations corresponding to all choices of initial state of $\mathcal{M}$; the *semigroup of the automaton* $S(\mathcal{M})$ is the semigroup generated by all these transformations. It is closely connected to Krohn-Rhodes Theory [98]. Its relevance to group theory was seen during Gluškov's seminar on automata [68].

The automaton $\mathcal{M}$ is *invertible* if furthermore it is complete with respect to output; namely,

$$\text{At every state and for each } a \in A \text{, there is a unique outgoing edge with output } a; \tag{2.2}$$

the corresponding transformation of $A^* \cup A^\omega$ is then invertible; and the set of such permutations, for all choices of initial state, generate the *group of the automaton* $G(\mathcal{M})$. Note that $S(\mathcal{M})$ may be a proper subsemigroup of $G(\mathcal{M})$, even if $\mathcal{M}$ is *invertible*. General references on groups generated by automata are [108, 77, 17].

As our first, fundamental example, consider the automaton with alphabet $A = \{0, 1\}$



(2.3)

in which the input i and output o of an edge are represented as '$i|o$'. The transformation associated with state $\mathbb{1}$ is clearly the identity transformation, since any path starting from $\mathbb{1}$ is a loop with same input and output. Consider now the transformation t. One has, for instance, $t \cdot 111001 = 000101$, with the path consisting of three loops at t, the edge to $\mathbb{1}$, and two loops at $\mathbb{1}$. In particular, $G(\mathcal{T}) = \langle t \rangle$. We will see in §24.2.7 that it is infinite cyclic.

Lemma 2.1. *The product of two automatic transformations is automatic. The inverse of an invertible automatic transformation is automatic.*

The proof becomes transparent once we introduce a good notation. If in an automaton $\mathcal{M}$ there is a transition from state q to state r, with input i and output o, we write

$$q \cdot i = o \cdot r. \tag{2.4}$$

In effect, if the state set of $\mathcal{M}$ is Q, we are encoding $\mathcal{M}$ by a function $\tau : Q \times A \to A \times Q$. It then follows from (2.1) that, given $q \in Q$ and $v = a_1 \cdots a_n \in A^*$, there are unique $w = b_1 \cdots b_n \in A^*, r \in Q$ such that $q \cdot a_1 \cdots a_n = b_1 \cdots b_n \cdot r$. The image of v under the transformation q is w. We have in fact extended naturally the function τ to a function $\tau : Q \times A^* \to A^* \times Q$.

Proof of Lemma 2.1. Given $\mathcal{M}, \mathcal{N}$ initial automata with state sets Q, R respectively, consider the automaton $\mathcal{M}\mathcal{N}$ with state set $Q \times R$ and transitions defined by $(q, r) \cdot i = q \cdot (r \cdot i) = o \cdot (q', r')$. If q_0, r_0 are the initial states of $\mathcal{M}, \mathcal{N}$, then the transformation $q_0 \circ r_0$ is the transformation corresponding to state (q_0, r_0) in $\mathcal{M}\mathcal{N}$.

Similarly, if q_0 induces an invertible transformation, consider the automaton $\mathcal{M}^{-1}$ with state set $\{q^{-1} \mid q \in Q\}$ and transitions defined by $q^{-1} \cdot o = i \cdot r^{-1}$ whenever (2.4) holds. The transformation induced by q_0^{-1} is the inverse of q_0. □

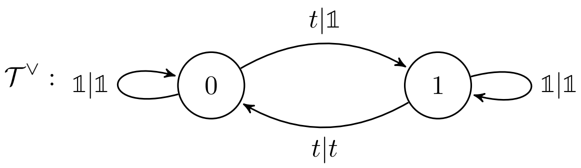
This construction applies naturally to any composition of finitely many automatic transformations. In case they all arise from the same machine $\mathcal{M}$, we *de facto* extend the function τ describing $\mathcal{M}$ to a function $\tau : Q^* \times A^* \to A^* \times Q^*$, and (if $\mathcal{M}$ is invertible) to a function $\tau : F_Q \times A^* \to A^* \times F_Q$. It projects to a function $\tau : S(\mathcal{M}) \times A^* \to A^* \times S(\mathcal{M})$, and, if $\mathcal{M}$ is invertible, to a function $\tau : G(\mathcal{M}) \times A^* \to A^* \times G(\mathcal{M})$.

Note that a function $G(\mathcal{M}) \times A \to A \times G(\mathcal{M})$ naturally gives a function, still written $\tau : G(\mathcal{M}) \to G(\mathcal{M})^A \rtimes \mathrm{Sym}(A)$; this is the semidirect product of functions $A \to G(\mathcal{M})$ by the symmetric group of A (acting by permutation of coordinates), and is commonly called the *wreath product* $G(\mathcal{M}) \wr \mathrm{Sym}(A)$, see also Chapter 16.

This wreath product decomposition also inspires a convenient description of the function τ by a *matrix embedding*; the size and shape of the matrix is determined by the permutation of A, and the nonzero entries by the elements in $G(\mathcal{M})^A$; more precisely, assume $A = \{1, \dots, d\}$, and, for $\tau(q) = ((s_1, \dots, s_d), \pi) \in G(\mathcal{M})^A \rtimes \mathrm{Sym}(A)$, write $\tau'(q) =$ the permutation matrix with s_i at position $(i, i\pi)$. Then these matrices multiply as wreath product elements. More algebraically, we have defined a homomomorphism $\tau' : \Bbbk G \to M_d(\Bbbk G)$, where $\Bbbk G$ is the group ring of G over the field $\Bbbk$. Such an embedding defines an algebra acting on the linear span of A^*; this algebra has important properties, studied in [116] for Gupta-Sidki's example and in [12] for Grigorchuk's example.

The action of $g \in G(\mathcal{M})$ may be described as follows: given a sequence $u = a_1 \cdots a_n$, compute $\tau(g, u) = (w, h)$. Then $g \cdot u = w$; and the image of $g \cdot (uv) = w(h \cdot v)$; that is, the action of g on sequences starting by u is defined by an element $h \in G(\mathcal{M})$ acting on the tail of the sequence. More geometrically, we can picture A^* as an infinite tree. The action of g carries the subtree uA^* to wA^*, and, within uA^* naturally identified with A^*, acts by the element h. For that reason, $G(\mathcal{M})$ is called a *self-similar group*.

The formalism expressing a Mealy machine as a map $\tau : Q \times A \to A \times Q$ is completely symmetric with respect to A and Q; the *dual* of the automaton $\mathcal{M}$ is the automaton $\mathcal{M}^\vee$ with state set A, alphabet Q, and transitions given by $i \cdot q = r \cdot o$ whenever (2.4) holds. For example, the dual of (2.3) is



(2.5)

In case the dual $\mathcal{M}^\vee$ of the automaton $\mathcal{M}$ is itself an invertible automaton, $\mathcal{M}$ is called *reversible*. If $\mathcal{M}$, $\mathcal{M}^\vee$ and $(\mathcal{M}^{-1})^\vee$ are all invertible, then $\mathcal{M}$ is *bireversible*; it then has eight associated automata, obtained through all combinations of $()^{-1}$ and $()^\vee$.

Note that $\mathcal{M}^\vee$ naturally encodes the action of $S(\mathcal{M})$ on A: it is a graph with vertex set A, and an edge, with (input) label q, from a to $q \cdot a$. More generally, $(\mathcal{M}^n)^\vee$ encodes the action of $S(\mathcal{M})$ on the set A^n of words of length n.

More generally, we will consider subgroups of $G(\mathcal{M})$, namely subgroups generated by a subset of the states of an automaton; we call these groups *automata groups*. This is a large class of groups, which contains in particular finitely generated linear groups, see Theorem 2.2 below or [36]. The elements of automata groups are, strictly speaking, automatic permutations of A^*. It is often convenient to identify them with a corresponding automaton, for instance constructed as a power of the original Mealy automaton (keeping in mind the construction for the composition of automatic transformations), with appropriate initial state.

Theorem 2.2 (Brunner-Sidki). *The affine group $\mathbb{Z}^n \rtimes \mathrm{GL}_n(\mathbb{Z})$ is an automata group for each n.*

This will be proven in more generality in §24.2.7.

We mention some closure properties of automata groups. Clearly a direct product of automata groups is an automata group (take the direct product of the alphabets). A more subtle operation, called *tree-wreathing* in [37, 119], gives wreath products with $\mathbb{Z}$.

A more general class of groups has also been considered, and is relevant to §24.2.6: *functionally recursive groups*. Let A denote a finite alphabet, Q a finite set, and $F = F_Q$ the free group on Q. The "automaton" now is given by a set of rules of the form

$$q \cdot a = b \cdot r$$

for all $q \in Q, a \in A$, where $b \in A$ and $r \in F$. In effect, in the dual $\mathcal{M}^\vee$ we are allowing arbitrary words over Q as output symbols.

2.1 Main examples

Automata groups gained significance when simple examples of finitely generated, infinite torsion groups, and groups of intermediate word-growth, were discovered. Alëshin [2] studied the automaton (2.7), and showed that $\langle A, B\rangle$ is an infinite torsion group. Another of his examples is described in §24.2.8.

Grigorchuk [71, 75, 73, 74, 72] simplified Alëshin's example as follows: let $\mathcal{A}$ be obtained from the Alëshin automaton by removing the gray states; the state set of $\mathcal{A}$ is $\{\mathbb{1}, a, b, c, d\}$. He showed that $G(\mathcal{A})$, which is known as the *Grigorchuk group*, is an infinite torsion group; see Theorem 2.9. In fact, $G(\mathcal{A})$ and $\langle A, B\rangle$ have isomorphic finite-index subgroups.

Gupta and Sidki [88, 87] constructed for all prime p an infinite, p-torsion group; their construction, for $p = 3$, is the automata group $G(\mathcal{G})$ generated by the automaton (2.8).

All invertible automata with at most three states and two alphabet letters have been listed in [26]; here are some important examples.

The affine group $BS_{1,3} = \{z \mapsto 3^p z + q/3^r \mid p, q, r \in \mathbb{Z}\}$, see (1.4) is a linear group, and an automata group by Theorem 2.16; see also [22]. It is generated by the automaton (2.9).

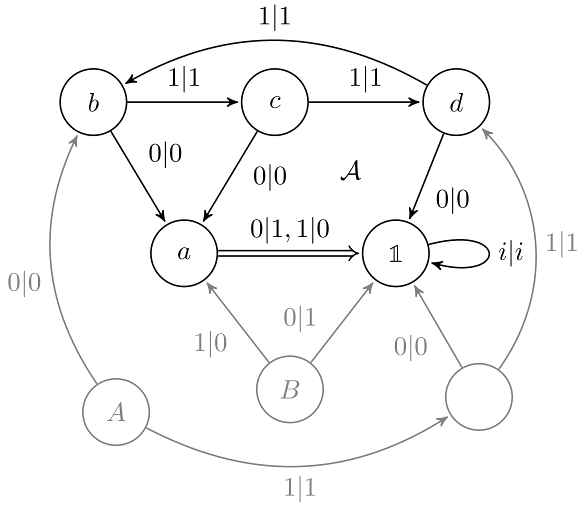
As another important example, consider the lamplighter group

$$G = (\mathbb{Z}/2)^{(\mathbb{Z})} \rtimes \mathbb{Z} = \langle a, t \mid a^2, [a, a^{t^n}] \text{ for all } n \in \mathbb{Z}\rangle. \tag{2.6}$$

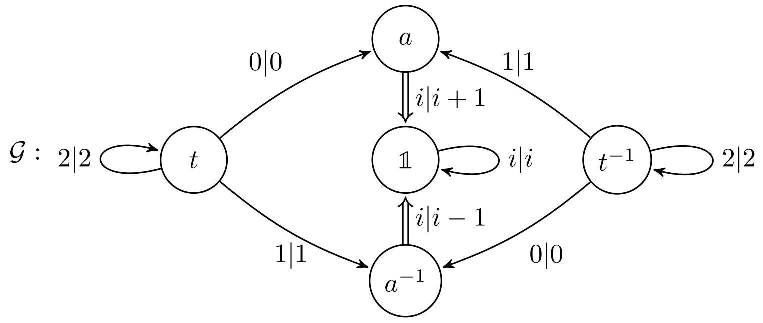
It is an automata group [80], embedded as the set of maps

$$\{z \mapsto (t+1)^p z + q \mid p \in \mathbb{Z}, q \in \mathbb{F}_2[t+1, (t+1)^{-1}]\}$$

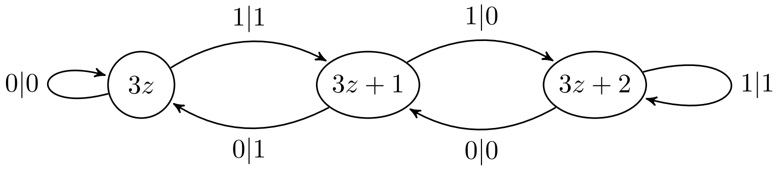
in the affine group of $\mathbb{F}_2[[t]]$. It is generated by the automaton $\mathcal{L}$ in (2.10).



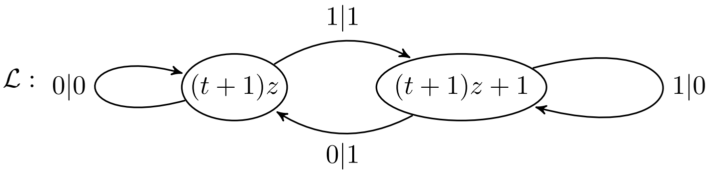
(2.7)



(2.8)

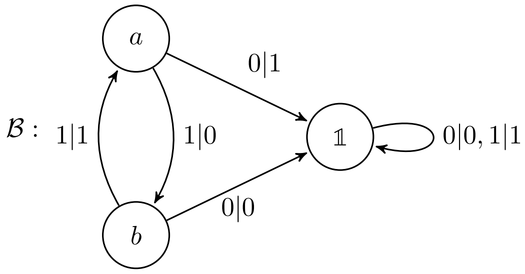


(2.9)



(2.10)

The Basilica group, see [81, 23], will appear again in §24.2.6. It is generated by the automaton (2.11).



(2.11)

There are (unpublished) lists by Sushchansky *et al.* of all (not necessarily invertible) automata with $\leqslant 3$ states, on a binary alphabet; there are more than 2000 such automata; the invertible ones are listed in [26].

How about groups that are *not* automata groups? Groups with unsolvable word problem (or more generally whose word problem cannot be solved in exponential time, see §24.2.2), and groups that are not residually finite (or more generally that are not residually (finite with composition factors of bounded order)) among the simplest examples. In fact, it is difficult to come up with any other ones.

2.2 Decision problems

One virtue of automata groups is that elements may easily be compared, since (Mealy) automata admit a unique minimized form, which furthermore may efficiently be computed in time $\mathcal{O}(\operatorname{Card} A \operatorname{Card} Q \log \operatorname{Card} Q)$, see [95, 97].

Proposition 2.3. *Let G be an automata group. Then the word problem is solvable in G, in at worst exponential time.*

Proof. Let Q be a generating set for G, and for each $q \in Q$ compute the minimal automaton $\mathcal{M}_q$ representing q. Let C be an upper bound for the number of states of any $\mathcal{M}_q$.

Now given a word $w = q_1 \cdots q_n \in (Q \sqcup Q^{-1})^*$, multiply the automata $\mathcal{M}_{q_1}, \dots, \mathcal{M}_{q_n}$. The result is an automaton with $\leqslant C^n$ states. Then w is trivial if and only if all states to which the initial state leads have identical input and output symbols. □

It is unknown if the conjugacy or generalized word problem are solvable in general; though this is known in particular cases, such as the Grigorchuk group $G(\mathcal{A})$, see [114, 99, 78]. The conjugacy problem is solvable as soon as $G(\mathcal{A})$ is *conjugacy separable*, namely, for g, h non-conjugate in $G(\mathcal{A})$ there exists a finite quotient of $G(\mathcal{A})$ in which their images are non-conjugate. Indeed automata groups are recursively presented and residually finite.

It is also unknown whether the order problem is solvable in arbitrary automata groups; but this is known for particular cases, such as bounded automata groups, see §24.2.3.

Nekrashevych's limit space (see Theorem 2.15) may sometimes be used to prove that two contracting, self-similar groups are non-isomorphic: By [79], some groups admit essentially only one weakly branch self-similar action; if the group is also contracting, then its limit space is an isomorphism invariant.

On the other hand, in the more general class of functionally recursive groups, the very solvability of the word problem remains so far an open problem.

2.3 Bounded and contracting automata

As we saw in §24.2.2, it may be useful to note, and use, additional properties of automata groups.

Definition 2.2. An automaton $\mathcal{M}$ is *bounded* if the function which to $n \in \mathbb{N}$ associates the number of paths of length n in $\mathcal{M}$ that do not end at the identity state is a bounded function. A group is *bounded* if its elements are bounded automata; or, equivalently, if it is generated by bounded automata.

More generally, Sidki considered automata for which that function is bounded by a polynomial; see [117]. He showed in [118] that such groups cannot contain non-abelian free subgroups.

Definition 2.3. An automaton $\mathcal{M}$ is *nuclear* if the set of recurrent states of $\mathcal{M}\mathcal{M}$ spans an automaton isomorphic to $\mathcal{M}$; and, for invertible $\mathcal{M}$, if additionally $\mathcal{M} = \mathcal{M}^{-1}$. Recall that a state is *recurrent* if it is the endpoint of arbitrarily long paths.

An invertible automaton $\mathcal{M}$ is *contracting* if $G(\mathcal{M}) = G(\mathcal{N})$ for a (necessarily unique) nuclear automaton $\mathcal{N}$. The *nucleus* of $G(\mathcal{M})$ is then $\mathcal{N}$.

For example, the automata (2.7,2.8) are nuclear; the automata (2.3,2.11) are contracting, with nucleus $\{1, t, t^{-1}\}$ and $\{1, a^{\pm 1}, b^{\pm 1}, b^{-1}a, a^{-1}b\}$; the automaton (2.10) is not contracting.

If $\mathcal{M}$ is contracting, then for every $g \in G(\mathcal{M})$ there is a constant K such that (in the automaton describing g) all paths of length $\geqslant K$ end at a state in $\mathcal{M}$. It also implies that there are constants L, m and $\lambda < 1$ such that, for the word metric $\|\cdot\|$ on $G(\mathcal{M})$, whenever one has $g \cdot a_1 \cdots a_m = b_1 \cdots b_m \cdot h$ with $h, g \in G(\mathcal{M})$, one has $\|h\| \leqslant \lambda\|g\| + L$.

Proposition 2.4 ([108, Theorem 3.9.12]). *Finitely generated bounded groups are contracting.*

Consider the following graph $\mathscr{X}(\mathcal{M})$: its vertex set is A^*. It has two kinds of edges, *vertical* and *horizontal*. There is a vertical edge (u, ua) for all $u \in A^*, a \in A$, and a horizontal edge $(u, q \cdot u)$ for every $u \in A^*, q \in Q$. Note that the horizontal and vertical edges form squares labeled as in (2.4), and that the horizontal edges form the Schreier graphs of the action of $G(\mathcal{M})$ on A^n.

Proposition 2.5 ([108, Theorem 3.8.6]). *If $G(\mathcal{M})$ is contracting then $\mathscr{X}(\mathcal{M})$ is a hyperbolic graph in the sense of Definition 1.5.*

Discrete groups may be broadly separated in two classes: *amenable* and *non-amenable* groups. A group G is *amenable* if it admits a normalized, invariant mean, that is, a map $\mu : \mathcal{P}(G) \to [0, 1]$ with $\mu(A \sqcup B) = \mu(A) + \mu(B)$, $\mu(G) = 1$ and $\mu(gA) = \mu(A)$ for all $g \in G$ and $A, B \subseteq G$. All finite and abelian groups are amenable; so are groups of subexponential word-growth (see §24.2.5). Extensions, quotients, subgroups, and directed unions of amenable groups are amenable. On the other hand, non-abelian free groups are non-amenable.

In understanding the frontier between amenable and non-amenable groups, the Basilica group $G(\mathcal{B})$ stands out as an important example: Bartholdi and Virág proved that it is amenable [23], but its amenability cannot be decided by the criteria of the previous paragraph. We now briefly indicate the core of the argument.

The matrix embedding $\tau' : \Bbbk G \to M_d(\Bbbk G)$ associated with a self-similar group (see page 802) extends to a map $\tau' : \ell^1(G) \to M_d(\ell^1(G))$ on measures on G. A measure μ gives rise to a random walk on G, with one-step transition probability $p_1(x, y) = \mu(xy^{-1})$. On the other hand, $\tau'(\mu)$ naturally defines a random walk on $G \times X$; treating the second variable as an "internal degree of freedom", one may sample the random walk on $G \times X$ each time it hits $G \times \{x_0\}$ for a fixed $x_0 \in X$. In favourable cases, the corresponding random walk on G is *self-similar*: it is a convex combination of $\mathbb{1}$ and μ. One may then deduce that its "asymptotic entropy" vanishes, and therefore that G is amenable. This strategy works in the following cases:

Theorem 2.6 (Bartholdi-Kaimanovich-Nekrashevych [18]). *Bounded groups are amenable.*

Theorem 2.7 (Amir, Angel, Virág[7]). *Automata of linear growth generate amenable groups.*

Nekrashevych conjectures that contracting automata always generate amenable groups, and proves:

Proposition 2.8 (Nekrashevych, [109]). *A contracting self-similar group cannot contain a non-abelian free subgroup.*

We turn to the original claim to fame of automata groups:

Theorem 2.9 (Alëshin-Grigorchuk [2, 71], Gupta-Sidki [88]). *The Grigorchuk group $G(\mathcal{A})$ and the Gupta-Sidki group $G(\mathcal{G})$ are infinite, finitely generated torsion groups.*

Sketch of proof. To see that these groups G are infinite, consider their action on A^*, the stabilizer H of $0 \in A \subset A^*$, and the restriction θ of the action of H to $0A^*$. This defines a homomorphism $\theta : H \to \mathrm{Sym}(0A^*) \cong \mathrm{Sym}(A^*)$, which is in fact onto G. Therefore G possesses a proper subgroup mapping onto G, so is infinite.

To see that these groups are torsion, proceed by induction on the word-length of an element $g \in G$. The initial cases $a^2 = b^2 = c^2 = d^2 = \mathbb{1}$, respectively $a^3 = t^3 = \mathbb{1}$, are easily checked. Now consider again the action of g on $A \subset A^*$. If g fixes A, then its actions on the subsets iA^* are again defined by elements of G, which are shorter by the contraction property; so have finite order. It follows that g itself has finite order.

If, on the other hand, g does not fix A, then $g^{\mathrm{Card}\,A}$ fixes A; the action of $g^{\mathrm{Card}\,A}$ on iA^* is defined by an element of G, of length at most the length of g; and (by an argument that we skip) smaller in the induction order than g; so $g^{\mathrm{Card}\,A}$ is torsion and so is g. □

Contracting groups have recursive presentations (meaning the relators $\mathscr{R}$ of the presentation form a recursive subset of F_Q); in favourable cases, such as branch groups [10], the set of relators is the set of iterates, under an endomorphism of F_Q, of a finite subset of F_Q. For example [102], Grigorchuk's group satisfies

$$G(\mathcal{A}) = \langle a, b, c, d \mid \sigma^n(bcd), \sigma^n(a^2), \sigma^n([d, d^a]), \sigma^n([d, d^{[a,c]a}]) \text{ for all } n \in \mathbb{N}\rangle,$$

where σ is the endomorphism of $F_{\{a,b,c,d\}}$

$$\sigma : a \mapsto aca, b \mapsto d \mapsto c \mapsto b. \tag{2.12}$$

A similar statement holds for the Basilica group (2.11):

$$G(\mathcal{B}) = \langle a, b \mid [a^p, (a^p)^{b^p}], [b^p, (b^p)^{a^{2p}}] \text{ for all } p = 2^n\rangle;$$

here the endomorphism is $\sigma : a \mapsto b \mapsto a^2$.

2.4 Branch groups

Some of the most-studied examples of automata groups are *branch groups*, see [76] or the survey [17]. We will define a strictly smaller class:

Definition 2.4. An automata group $G(\mathcal{M})$ is *regular weakly branch* if it acts transitively on A^n for all n, and if there exists a nontrivial subgroup K of $G(\mathcal{M})$ such that, for all $u \in A^*$ and all $k \in K$, the permutation

$$w \mapsto \begin{cases} u\,k(v) & \text{if } w = uv, \\ w & \text{otherwise} \end{cases}$$

belongs to $G(\mathcal{M})$.

The group $G(\mathcal{M})$ is *regular branch* if furthermore K has finite index in $G(\mathcal{M})$.

If we view A^* as an infinite tree, a regular branch group G contains a rich supply of tree automorphisms in two manners: enough automorphisms to permute any two vertices of the same depth; and, for any disjoint subtrees of A^*, and for (up to finite index) any elements of G acting on these subtrees, an automorphism acting in that manner on A^*.

In particular, if G is a regular branch group, then G and $G \times \cdots \times G$, with $\operatorname{Card} A$ factors, have isomorphic finite-index subgroups (they are *commensurable*, see (2.4)).

Proposition 2.10. *The Grigorchuk group $G(\mathcal{A})$ and the Gupta-Sidki group $G(\mathcal{G})$ are regular branch; the Basilica group $G(\mathcal{B})$ is regular weakly branch.*

Sketch of proof. For $G = G(\mathcal{A})$, note first that G acts transitively on A; since the stabilizer of 0 acts as G on $0A^*$, by induction G acts transitively on A^n for all $n \in \mathbb{N}$.

Define then $x = [a, b]$ and $K = \langle\langle x \rangle\rangle$. Consider the endomorphism (2.12), and note that $\sigma(x) = [aca, d] = [x^{-1}, d] \in K$ using the relation $(ad)^4 = \mathbb{1}$, so σ restricts to an endomorphism $K \to K$, such that $\sigma(k)$ acts as k on $1A^*$ and fixes $0A^*$. Similarly, $\sigma^n(k)$ acts as k on 1^nA^*, so Definition 2.4 is fulfilled for $u = 1^n$. Since G acts transitively on A^n, the definition is also fulfilled for other $u \in A^n$.

Finally, a direct computation shows that K has index 16 in G.

The other groups $G(\mathcal{G})$ and $G(\mathcal{B})$ are handled similarly; for them, one takes $K = [G, G]$. □

Various consequences may be derived from a group being a branch group; in particular,

Theorem 2.11 (Abért, [1]). *A weakly branch group satisfies no identity; that is, if G is a weakly branch group, then for every nontrivial word $w = w(x_1, \ldots, x_k) \in F_k$, there are $a_1, \ldots, a_k \in G$ such that $w(a_1, \ldots, a_k) \neq \mathbb{1}$.*

2.5 Growth of groups

An important geometric invariant of a finitely generated group is the asymptotic behaviour of its growth function $\gamma_{G,A}(n)$. Finite groups, of course, have a bounded growth function. If G has a finite-index nilpotent subgroup, then $\gamma_{G,A}(n)$ is bounded by a polynomial, and one says G has *polynomial growth*; the converse is true [82].

On the other hand, if G contains a free subgroup, for example if G is word-hyperbolic and is not a finite extension of $\mathbb{Z}$, then $\gamma_{G,A}$ is bounded from above and below by exponential functions, and one says that G has *exponential growth*.

By a result of Milnor and Wolf [129, 105], if G has a solvable subgroup of finite index then G has either polynomial or exponential growth. The same conclusion holds, by Tits' alternative [123], if G is linear. Milnor [106] asked whether there exist groups with growth strictly between polynomial and exponential.

Theorem 2.12 (Grigorchuk [72]). *The Grigorchuk group $G(\mathcal{A})$ has intermediate growth.*

In fact, its growth function satisfies the following estimates:

$$e^{n^\alpha} \precsim \gamma_{G,S}(n) \precsim e^{n^\beta},$$

with $\alpha = 0.515$ and $\beta = \log(2)/\log(2/\eta) \approx 0.767$, for $\eta \approx 0.811$ the real root of the polynomial $X^3 + X^2 + X - 2$.

Sketch of proof; see [8, 9]. Recall that G admits an endomorphism σ, see (2.12), such that $\sigma(g)$ acts as g on $1A^*$ and as an element of the finite dihedral group $D_8 = \langle a, d\rangle$ on $0A^*$.

Given $g_0, g_1 \in G$ of length $\leqslant N$, the element $g = a\sigma(g_0)a\sigma(g_1)$ has length $\leqslant 4N$, and acts (up to an element of D_8) as g_i on iA^* for $i = 0, 1$. It follows that g essentially (i.e., up to 8 choices) determines g_0, g_1, and therefore that $\gamma_{G,S}(4N) \geqslant (\gamma_{G,S}(N)/8)^2$. The lower bound follows easily.

On the other hand, the Grigorchuk group G satisfies a stronger property than contraction; namely, for a well-chosen metric (which is equivalent to the word metric), one has that if $g \in G$ acts as $g_i \in G$ on iA^*, then

$$\|g_0\| + \|g_1\| \leqslant \eta(\|g\| + 1), \tag{2.13}$$

with η the constant above.

Then, to every $g \in G$ one associates a description by a finite, labeled binary tree $\iota(g)$. If $\|g\| \leqslant 1/(1-\eta)$, its description is a one-vertex tree with g at its unique leaf. Otherwise, let $i \in \{0, 1\}$ be such that ga^i fixes A, and write g_0, g_1 the elements of G defined by the actions of ga^i on $0A^*, 1A^*$ respectively. Construct recursively the descriptions $\iota(g_0), \iota(g_1)$. Then the description of g is a tree with i at its root, and two descendants $\iota(g_0), \iota(g_1)$.

By (2.13), the tree $\iota(g)$ has at most $\|g\|^\beta$ leaves; and $\iota(g)$ determines g. There are exponentially many trees with a given number of leaves, and the upper bound follows. □

Among groups of exponential growth, Gromov asked the following question [83]: is there a group G of exponential growth, namely such that $\lim \gamma_{G,Q}(n)^{1/n} > 1$ for all (finite) Q, but such that $\inf_{Q \subset G} \lim \gamma_{G,Q}(n)^{1/n} = 1$?

Such examples, called *groups of non-uniform exponential growth*, were first found by Wilson [127]; see [11, 15] for a simplification. Both constructions are heavily based on groups generated by automata.

It is known that essentially any function growing faster than n^2 may be, asymptotically, the growth function of a semigroup. Notably, very small automata generate semigroups of growth $\sim e^{\sqrt{n}}$, and of polynomial growth of irrational degree [20, 21]. However, it is not known whether there exist groups whose growth function is strictly between polynomial and $e^{\sqrt{n}}$. There *is* a gap in the spectrum of growth functions: no group with growth strictly between polynomials and $n^{(\log n)^{1/100}}$ exists [115]. The largest class of growth functions is the following:

Theorem 2.13 (Bartholdi-Erschler [14, 13]). *Let* $\eta \cong 0.811$ *be the positive root of* $X^3 + X^2 + X - 2$ *as above. Let* $f\colon \mathbb{N} \to \mathbb{N}$ *satisfy*

$$f(2n) \leqslant f(n)^2 \leqslant f(\lfloor 2n/\eta \rfloor) \text{ for all } n \gg 0.$$

Then there exists a group with growth function $\sim f$.

2.6 Dynamics and subdivision rules

We show, in this subsection, how automata naturally arise from geometric or topological situations. As a first step, we will obtain a functionally recursive action; in favourable cases it will be encoded by an automaton. We must first adopt a slightly more abstract point of view on functionally recursive groups:

Definition 2.5. A group G is *self-similar* if it is endowed with a *self-similarity biset*, that is, a set $\mathfrak{B}$ with commuting left and right actions, that is free qua right G-set.

The fundamental example is $G = G(\mathcal{M})$ and $\mathfrak{B} = A \times G$, with actions

$$g \cdot (a, h) = (b, kh) \text{ if } \tau(g, a) = (b, k), \qquad (a, g) \cdot h = (a, gh).$$

Conversely, given a self-similar group G, choose a *basis* A of its biset, i.e., express $\mathfrak{B} = A \times G$; then define $\tau(g, a) = (b, k)$ whenever $g \cdot (a, 1) = (b, k)$ in $\mathfrak{B}$. This vindicates the notation (2.4).

Two bisets $\mathfrak{B}, \mathfrak{B}'$ are *isomorphic* if there is a map $\varphi : \mathfrak{B} \to \mathfrak{B}'$ with $g\varphi(b)h = \varphi(gbh)$ for all $g, h \in G, b \in \mathfrak{B}$. They are *equivalent* if there is a map $\varphi : \mathfrak{B} \to \mathfrak{B}'$ and an automorphism $\theta : G \to G$ with $\theta(g)\varphi(b)\theta(h) = \varphi(gbh)$.

Consider now X a topological space, and $f : X \to X$ a *branched covering*; this means that there is an open dense subspace $X_0 \subseteq X$ such that $f : f^{-1}(X_0) \to X_0$ is a covering. The subset $\mathscr{C} = X \setminus f^{-1}(X_0)$ is the *branch locus*, and $\mathscr{P} = \bigcup_{n \geqslant 1} f^n(\mathscr{C})$ is the *post-critical locus*. Write $\Omega = X \setminus \mathscr{P}$, and choose a basepoint $* \in \Omega$.

Two coverings $(f, \mathscr{P}_f)$ and $(g, \mathscr{P}_g)$ are *combinatorially equivalent* if there exists a path g_t through branched coverings, with $g_0 = f, g_1 = g$, such that the post-critical set of g_t varies continuously along the path.

We define a self-similarity biset for $G = \pi_1(\Omega, *)$: set

$$\mathfrak{B}_f = \{\text{homotopy classes of paths } \gamma : [0,1] \to \Omega \mid \gamma(0) = f(\gamma(1)) = *\}.$$

The right action of G prepends a loop at $*$ to γ; the left action appends the unique f-lift of the loop that starts at $\gamma(1)$ to γ.

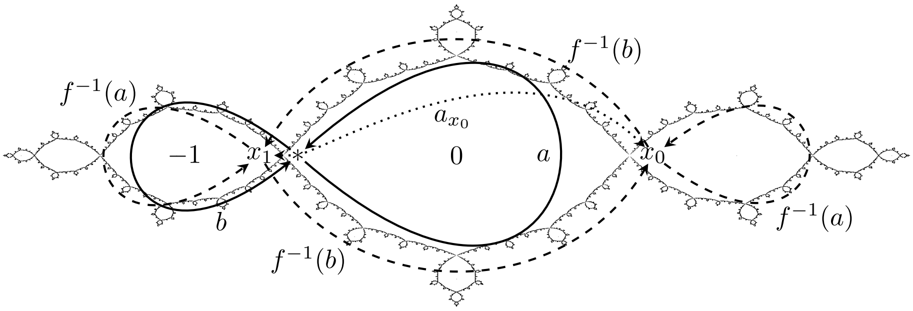
A choice of basis for $\mathfrak{B}$ amounts to choosing, for each $x \in f^{-1}(*)$, a path $a_x \subset \Omega$ from $*$ to x. Set $A = \{a_x \mid x \in f^{-1}(*)\}$. Now, for $g \in G$, and $a_x \in A$, consider a path

γ starting at x such that $f \circ \gamma = g$; such a path is unique up to homotopy, by the covering property of f. The path γ ends at some $y \in f^{-1}(*)$. Set then

$$\tau(g, a_x) = (a_y, a_y^{-1}\gamma a_x),$$

where we write concatenation of paths in reverse order, that is, $\gamma\delta$ is first δ, then γ.

For example, consider the sphere $X = \widehat{\mathbb{C}}$, with branched covering $f(z) = z^2 - 1$. Its post-critical locus is $\mathscr{P} = \{0, -1, \infty\}$. A direct calculation (see e.g. [16]) gives that its biset is the automaton (2.11); the relevant paths are shown here:



Branched self-coverings are encoded by self-similar groups in the following sense:

Theorem 2.14 (Nekrashevych). *Let f, g be branched coverings. Then f, g are combinatorially equivalent if and only if the bisets $\mathfrak{B}_f, \mathfrak{B}_g$ are equivalent.*

This result has been used to answer a long-standing open problem in complex dynamics [19].

If furthermore G is finitely generated and the map f expands a length metric, then the associated biset may be defined by a contracting automaton. This is, in particular, the case for all rational maps acting on the sphere $\widehat{\mathbb{C}}$.

Definition 2.6. Let $f : X \to X$ be a branched self-covering. The *iterated monodromy group* of f is the automata group $G(f) = G(\mathcal{M})$, where $\mathcal{M}$ is an automaton describing the biset $\mathfrak{B}_f$.

If $G = G(\mathcal{M})$ is a contracting self-similar group, consider the hyperbolic boundary $\mathscr{J} = \partial\mathscr{X}(\mathcal{M})$, called the *limit space* of G. It admits an expanding self-covering map $s : \mathscr{J} \to \mathscr{J}$, induced on vertices by the shift map $s(au) = u$.

Theorem 2.15 ([108, Theorems 5.2.6 and 5.4.3]). *The groups $G(s)$ and $G(\mathcal{M})$ are isomorphic.*

Conversely, suppose f is an analytic map, with Julia set *J, the points near which $\{f^{\circ n} \mid n \in \mathbb{N}\}$ does not form a normal family. Then (J, f) and $(\mathscr{J}, s)$ are homeomorphic and topologically conjugate.*

For instance, the Julia set of the Basilica map $f(z) = z^2 - 1$ is depicted above. Appropriately scaled and metrized, the Schreier graphs of the action of $G(\mathcal{M})$ on X^n converge to $\mathscr{J}$.

The first appearance of encodings of branched coverings by automata seems to be the "finite subdivision rules" by Cannon, Floyd and Parry [44]; they wished to know when a branched covering of the sphere may be realized as a conformal map. In their work, a finite subdivision rule is given by a finite subdivision of the sphere, a refinement of it, and a covering map from the refinement to the original subdivision; by iteration, one obtains finer and finer subdivisions of the sphere. The combinatorial information involved is essentially equivalent to a self-similarity biset. Contraction of $G(\mathcal{M})$ and combinatorial versions of expansion have been related in [45].

2.7 Reversible actions

Recall that an automaton $\mathcal{M}$ is *reversible* if its dual $\mathcal{M}^\vee$ is invertible. In other words, if $g \in G(\mathcal{M})$, the action of g is determined by the action on any subset uA^*, for $u \in A^*$.

We have already seen some examples of reversible automata, notably (2.9,2.10). That last example generalizes as follows: consider a finite group G, and set $A = Q = G$. Define an automaton $\mathcal{C}_G$, the "Cayley automaton" of G, by $\tau(q, a) = (qa, qa)$. This automaton seems to have first been considered in [98, page 358]. The automaton $\mathcal{L}$ in (2.10) is the special case $G = \mathbb{Z}/2\mathbb{Z}$. The inverse of the automaton $\mathcal{C}_G$ is a *reset machine*, in that the target of a transition depends only on the input, not on the source state. Silva and Steinberg [121] prove that, if G is abelian, then $G(\mathcal{C}_G) = G \wr \mathbb{Z}$.

A large class of reversible automata is covered by the following construction. Let R be a ring, let M be an R-module, and let N be a submodule of M, with M/N finite. Let $\varphi : N \to M$ be an R-module homomorphism. Define a decreasing sequence of submodules M_i of M by $M_0 = M$ and $M_{n+1} = \varphi^{-1}(M_n)$, and denote by $\operatorname{End}_R(M, \varphi)$ the algebra of R-endomorphisms of M that map M_n into M_n for all n. Assume finally that there is an algebra homomorphism $\widehat{\varphi} : \operatorname{End}_R(M, \varphi) \to \operatorname{End}_R(M, \varphi)$ such that $\varphi(an) = \widehat{\varphi}(a)\varphi(n)$ for all $a \in \operatorname{End}_R(M, \varphi), n \in N$. Consider

$$T_M = \{z \mapsto az + m \mid a \in \operatorname{End}_R(M, \varphi), m \in M\}$$

the affine semigroup of M.

Theorem 2.16. *Let A be a transversal of N in M. Then the semigroup T_M acts self-similarly on A^*, by*

$\tau(az + b, x) = (y, \widehat{\varphi}(a)z + \varphi(ax + b - y))$ *for the unique $y \in A$ with $ax + b - y \in N$.*

This action is

(1) faithful if and only if $\bigcap_n M_n = 0$;

(2) reversible if and only if φ is injective;

(3) defined by a finite-state automaton if $\widehat{\varphi}$ is an automorphism of finite order, and there exists a norm $\|\cdot\| : M \to \mathbb{N}$ such that $\|a + b\| \leqslant \|a\| + \|b\|$, for all $K \in \mathbb{N}$ the ball $\{m \in M \mid K \geqslant \|m\|\}$ is finite, and a constant $\lambda < 1$ satisfies $\|\varphi(n)\| \leqslant \lambda\|n\|$ for all $n \in N$.

We already saw some examples of this construction: the lamplighter automaton $\mathcal{L}$ is obtained by taking $R = M = \mathbb{F}_2[t], N = tM$, $\varphi(tm) = m$, $\widehat{\varphi} = 1$, and $\|f\| = 2^{\deg f}$

with $\lambda = \frac{1}{2}$. The semigroup $S(\mathcal{L})$ is contained in T_M, and the group $G(\mathcal{L})$ is contained in the affine group of $\mathbb{F}_2[[t]]$. More generally, the Cayley automaton of a finite group G is obtained by taking $R = G[[t]]$ with G viewed as a ring with product $xy = 0$ unless $x = 1$ or $y = 1$.

The adding machine (2.3) generates the subgroup of translations in the affine group of M with $R = M = \mathbb{Z}, N = 2M$, $\varphi(2m) = m$, and $\|m\| = |m|$. The same ring-theoretic data produce the Baumslag-Solitar group (2.9); as above, we use $R = \mathbb{Z}$ to obtain a semigroup, and $R = \mathbb{Z}_2$ (or any ring in which 3 is invertible) to obtain a group.

Consider, more generally, $R = \mathbb{Z}, M = \mathbb{Z}^n, N = 2M$, and $\varphi(2m) = m$. These data produce the affine group $\mathbb{Z}^n \rtimes \mathrm{GL}_n(\mathbb{Z})$, proving Theorem 2.2.

A finer construction, giving an action on the binary tree, is to take again $M = \mathbb{Z}^n$ and $N = \varphi^{-1}(M)$ with $\varphi^{-1}(x_1, \dots, x_n) = (2x_n, x_1, \dots, x_{n-1})$; here $\widehat{\varphi}(a) = \varphi \circ a \circ \varphi^{-1}$. This gives a faithful action, on the binary tree, of

$$\mathbb{Z}^n \rtimes \{a \in \mathrm{GL}_n(\mathbb{Z}) \mid a \bmod 2 \text{ is lower triangular}\}.$$

Sketch of proof. (1) The action is faithful if and only if the translation part $\{z \mapsto z + m\}$ acts faithfully; and $z \mapsto z + m$ acts trivially on A^* if and only if $m \in M_n$ for all $n \in \mathbb{N}$.

(2) For any $x \in A$, the map (not a homomorphism!) $T_M \to T_M$ which to $g \in T_M$ associates the permutation of A^* given by $A^* \to xA^* \xrightarrow{g} g(x)A^* \to A^*$ is injective precisely when φ is injective.

(3) Without loss of generality, suppose $\widehat{\varphi} = 1$. Consider $g = z \mapsto az + m \in T_M$. Let K be larger than the norms of $ax + y$ for all $x, y \in A$. Then the states of an automaton describing g are all of the form $z \mapsto az + m'$, with $\|m'\| \leqslant (\|m\| + K)/(1 - \lambda)$; there are finitely many possibilities for such m'. □

Note that the transversal A amounts to a choice of "digits": the analogy is clear in the case of the adding machine (2.3), which has digits $\{0, 1\}$ and "counts" in base 2. For more general radix representations and their association with automata, see e.g. [124].

2.8 Bireversible actions

Recall that an automaton $\mathcal{M}$ is bireversible if $\mathcal{M}, \mathcal{M}^\vee, (\mathcal{M}^{-1})^\vee, ((\mathcal{M}^\vee)^{-1})^\vee$ etc. are all invertible; equivalently, the map $\tau : Q \times A \to A \times Q$ is a bijection for Q the state set of $\mathcal{M} \sqcup \mathcal{M}^{-1}$.

Bireversible automata are interpreted in [103] in terms of *commensurators* of free groups, defined in (2.4) of Chapter 23. Consider a free group F_A on a set A. Its Cayley graph $\mathscr{C}$ is a tree, and F_A acts by isometries on $\mathscr{C}$, so we have $F_A \leqslant \mathrm{Isom}(\mathscr{C})$. Furthermore, $\mathscr{C}$ is oriented: its edges are labeled by $A \sqcup A^{-1}$, and we choose as orientation the edges labeled A. In this way, F_A is contained in the orientation-preserving subgroup of $\mathrm{Isom}(\mathscr{C})$, denoted $\overrightarrow{\mathrm{Isom}(\mathscr{C})}$.

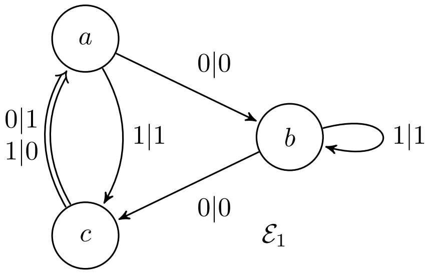
Proposition 2.17. *The stabilizer of* $\mathbb{1}$ *in* $\mathrm{Comm}_{\overrightarrow{\mathrm{Isom}(\mathscr{C})}}(F_A)$ *is the set of bireversible automata with alphabet* A.

Sketch of proof. The proof relies on an interpretation of finite-index subgroups of F_A as complete automata, see §23.2.2.

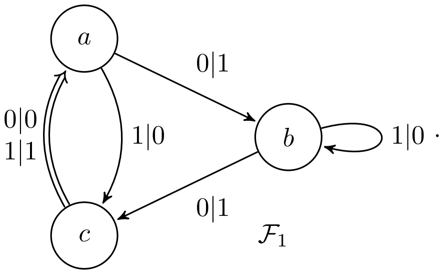
Let $\mathcal{M}$ be a bireversible automaton with alphabet A. Erase first the output labels from $\mathcal{M}$; this defines the Stallings automaton of a finite-index subgroup H_1 (of index $\operatorname{Card} Q$) of F_A. Erase then the input labels from $\mathcal{M}$; this defines an isomorphic subgroup H_2 of F_A. The automaton $\mathcal{M}$ itself defines an isomorphism between these two subgroups, which preserves the Cayley graph.

Conversely, given an isometry g of the Cayley graph of F_A which restricts to an isomorphism $G \to H$ between finite-index subgroups of F_A, the Stallings graphs of G and H and put their labels together, as input and output, to construct a bireversible automaton. □

It is striking that all known bireversible automata generate finitely presented groups. There are, up to isomorphism, precisely two minimized bireversible automata with three states and two alphabet letters:

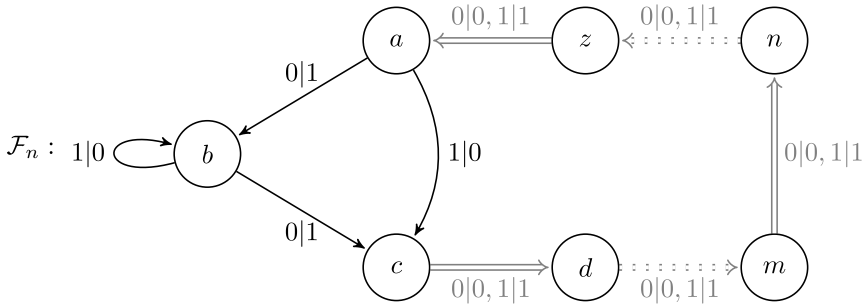


$\mathcal{E}_1$



$\mathcal{F}_1$

These automata are part of families, whose general term $\mathcal{E}_n, \mathcal{F}_n$ has $2n + 1$ states. We describe only $\mathcal{F}_n$:



Alëshin [3] proved that the group generated by the states b_1, b_2 in $\mathcal{F}_1, \mathcal{F}_2$ respectively is a free group on its two generators; but his argument (especially Lemma 8) has been considered incomplete, and a detailed proof appears in [126, Theorem 1.2]. Alëshin's method is to prove by induction that, for any reduced word $w \in \{b_1^{\pm 1}, b_2^{\pm 1}\}^*$, the syntactic monoid of the corresponding automaton acts transitively on its state set.

Sidki conjectured that in fact $G(\mathcal{F}_1)$ is a free group on its three generators; this has been proven in [125]. On the other hand, $G(\mathcal{E}_1)$ is a free product of three cyclic groups of order 2. Both proofs illustrate some techniques used to compute with bireversible automata. They rely on the following

Lemma 2.18. *Let $L \subset Q^*$ be a subset mapping to $G(\mathcal{M})$ through the evaluation map. If*

L is $G(\mathcal{M}^\vee)$-invariant, and every $G(\mathcal{M}^\vee)$-orbit contains a word mapping to a nontrivial element of $G(\mathcal{M})$, then L maps injectively onto $G(\mathcal{M})$.

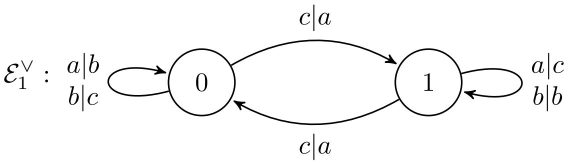
To derive the structure of a bireversible group, we therefore seek a $G(\mathcal{M}^\vee)$-invariant subset $L \subset Q^*$ that maps onto $G(\mathcal{M}) \setminus \{1\}$, and show that every $G(\mathcal{M})$-orbit contains a non-trivial element of $G(\mathcal{M})$.

Theorem 2.19 (Muntyan-Savchuk). $G(\mathcal{E}_1) = \langle a, b, c \mid a^2, b^2, c^2\rangle$.

Note that this result generalizes: $G(\mathcal{E}_n)$ is a free product of $2n + 1$ order-two groups.

Proof. Write $Q = \{a, b, c\}$. We first check the relations $a^2 = b^2 = c^2 = \mathbb{1}$ in $G = G(\mathcal{E}_1)$. Let $L \subset Q^*$ denote those sequences $s_1 \cdots s_n$ with $s_i \neq s_{i+1}$ for all i.

Consider the group $G(\mathcal{E}_1^\vee)$, with generators $0, 1$. It acts on L, and acts transitively on $L \cap Q^n$ for all n; indeed already 0 acts transitively on $Q = L \cap Q^1$, and 1 acts on $\{a, c\}Q^{n-1} \cap L$ as a 2^n-cycle, conjugate to the action (2.3) in the sense that there is an identification of $\{a, c\}Q^{n-1} \cap L$ with $\{0, 1\}^n$ interleaving these actions. It follows that the $3 \cdot 2^{n-1}$ elements of $L \cap A^n$ are in the same orbit.



It remains to note that $L \cap A^n$ contains a word mapping to a nontrivial element of G; for example, $c(ab)^{(n-1)/2}$ or $c(ab)^{n/2-1}a$ depending on the parity of n; and to apply Lemma 2.18. □

Theorem 2.20 (Vorobets). $G(\mathcal{F}_2) = \langle a, b, c \mid \emptyset\rangle \cong F_3$.

Note that this result generalizes: $G(\mathcal{F}_n)$ is a free group of rank $2n + 1$.

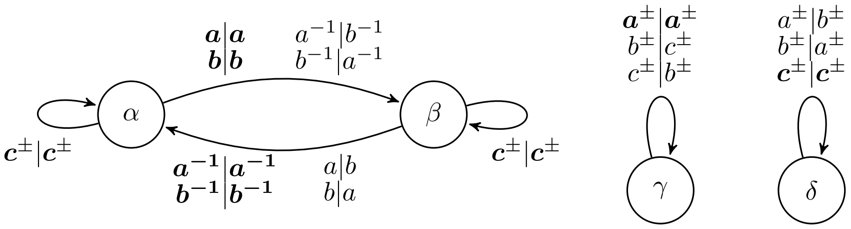
Sketch of proof. Again the key is to control the orbits of $G^\vee = G(\mathcal{F}_2^\vee) = \langle 0, 1\rangle$ on the reduced words over $Q = \{a, b, c\}$ of any given length. Let $s \in (\pm 1)^n$ be a sequence of signs, and consider

$$L_s = \{w = w_1^{s_1} \cdots w_n^{s_n} \in (Q \sqcup Q^{-1})^* \mid w_i^{s_i} \neq w_{i+1}^{-s_{i+1}} \text{ for all } i\}.$$

We show that $G^\vee$ acts transitively on L_s for all s, and that L_s contains a word mapping to a nontrivial element of G. Consider the elements

$$\alpha = 0^2 1^{-2} 0^2 1^{-1}, \quad \beta = 1^2 0^{-2} 1^2 0^{-1}, \quad \gamma = 1^{-1} 0, \quad \delta = 0 1^{-1}$$

of $G^\vee$, where the products are computed left-to-right; they are described by the automata


$\boldsymbol{a}|\boldsymbol{a}$
$\boldsymbol{b}|\boldsymbol{b}$
$a^{-1}|b^{-1}$
$b^{-1}|a^{-1}$
α
β
$\boldsymbol{c}^\pm|\boldsymbol{c}^\pm$
$\boldsymbol{a}^{-1}|\boldsymbol{a}^{-1}$
$\boldsymbol{b}^{-1}|\boldsymbol{b}^{-1}$
$a|b$
$b|a$
$\boldsymbol{c}^\pm|\boldsymbol{c}^\pm$
$\boldsymbol{a}^\pm|\boldsymbol{a}^\pm$
$b^\pm|c^\pm$
$c^\pm|b^\pm$
γ
$a^\pm|b^\pm$
$b^\pm|a^\pm$
$\boldsymbol{c}^\pm|\boldsymbol{c}^\pm$
δ

The elements γ, δ generate a copy of $\mathrm{Sym}(3)$, allowing arbitrary permutations of Q or Q^{-1}. In particular, $G^\vee$ acts transitively on L_s whenever $|s| \leqslant 1$, so we may proceed by induction on $|s|$. The elements α, β, on the other hand, fix a large set of sequences (following the bold edges in the automata).

Consider now $s = s_1 \cdots s_n$, and $s' = s_1 \cdots s_{n-1}$. If $s_{n-1} \neq s_n$, so that $\operatorname{Card} L_s = 2 \operatorname{Card} L_{s'}$, then there exists $w = w_1^{s_1} \cdots w_n^{s_n} \in L_s$, moved by α or β, and such that $w_1^{s_1} \cdots w_{n-1}^{s_{n-1}} \in L_{s'}$ is fixed by α and β; so $G^\vee$ acts transitively on L_s.

If $s_1 \neq s_2$, apply the same argument to $L_{s_n^{-1} \cdots s_1^{-1}}$ and $L_{s_n^{-1} \cdots s_2^{-1}}$.

Finally, if $s_1 = s_2$ and $s_{n-1} = s_n$, consider a typical $w \in L_{s_2 \cdots s_{n-1}}$, and all $w_{qr} = q^{s_1} w r^{s_n}$, for $q, r \in Q$. Using the action of α and β, the words w_{qa} and w_{qb} are in the same $G^\vee$-orbit for all $q \in Q$, and similarly w_{ar} and w_{br} are in the same $G^\vee$-orbit for all $r \in Q$. For all $r \in Q$, finally, $w_{ar}, w_{br'}, w_{cr''}$ are in the same $G^\vee$-orbit for some $r', r'' \in Q$, and similarly $w_{qa,q'b,q''c}$ are in the same $G^\vee$-orbit. It follows that all w_{qr} are in the same $G^\vee$-orbit, so by induction L_s is a single orbit.

It remains to check that every L_s contains a word w mapping to a nontrivial group element. If n is odd, set $w_i = a$ if $s_i = 1$ and $w_i = b$ if $s_i = -1$; then $\overline{w}$ acts nontrivially on A. If n is even, change w_n to c^{s_n}; again $\overline{w}$ acts nontrivially on A. We are done by Lemma 2.18. □

Burger and Mozes [38, 39, 40] have constructed some infinite, finitely presented simple groups, see also [112]. From this chapter's point of view, these groups are obtained as follows: one constructs an "appropriate" bireversible automaton $\mathcal{M}$ with state set Q and alphabet A, defines

$$G_0 = \langle A \cup Q \mid aq = rb \text{ whenever that relation holds in } \mathcal{M} \rangle,$$

and considers G a finite-index subgroup of G_0. We will not explicitly give here the conditions required on $\mathcal{M}$ for their construction to work; but note that automata groups can be understood as a byproduct of their work. Wise constructed finitely presented groups with non-residual finiteness properties that are also related to automata [128].

Burger and Mozes give the following algebraic construction: consider two primes $p, \ell \equiv 1 \pmod 4$. Let A (respectively Q) denote those integral quaternions, up to a unit $\pm 1, \pm i, \pm j, \pm k$, of norm p (respectively ℓ). By a result of Hurwitz, $\operatorname{Card} A = p + 1$ and $\operatorname{Card} Q = \ell + 1$. Furthermore [96], for every $q \in Q, a \in A$ there are unique (again up to units) $b \in A, r \in Q$ with $qa = br$. Use these relations to define an automaton $\mathcal{M}_{p,\ell}$. Clearly $\mathcal{M}_{p,\ell}$ is bireversible, with dual $\mathcal{M}_{p,\ell}^\vee = \mathcal{M}_{\ell,p}$. Again thanks to unique factorization of integral quaternions of odd norm,

Proposition 2.21. $G(\mathcal{M}_{p,\ell}) = F_{(\ell+1)/2}$.

Glasner and Mozes [67] constructed an example of a bireversible automata group with Kazhdan's property (T).

References

[1] Abért, Miklós. Group laws and free subgroups in topological groups. *Bull. London Math. Soc.*, 37(4):525–534, 2005. 809

[2] Alëšin, Stanislav V. Finite automata and the burnside problem for periodic groups. *Mat. Zametki*, 11:319–328, 1972. 804, 808

[3] Alëšin, Stanislav V. A free group of finite automata. *Vestnik Moskov. Univ. Ser. I Mat. Mekh.*, 4:12–14, 1983. 815

[4] Alonso, Juan M. Combings of groups. pages 165–178. 1992. 796

[5] Alonso, Juan M., Brady, Tom, Cooper, Darryl, Ferlini, Vincent, Lustig, Martin, Mihalik, Michael, Shapiro, Michael, and Short, Hamish. Notes on word hyperbolic groups. pages 3–63. 1991. Edited by H. Short. 797

[6] Alonso, Juan M. and Bridson, Martin R. Semihyperbolic groups. *Proc. London Math. Soc. (3)*, 70(1):56–114, 1995. 799

[7] Amir, Gideon, Angel, Omer, and Virág, Bálint. Amenability of linear-activity automaton groups. *J. Eur. Math. Soc. (JEMS)*, 15(3):705–730, 2013. 808

[8] Bartholdi, Laurent. The growth of grigorchuk's torsion group. *Internat. Math. Res. Notices*, 20:1049–1054, 1998. 810

[9] Bartholdi, Laurent. Lower bounds on the growth of a group acting on the binary rooted tree. *Internat. J. Algebra Comput.*, 11(1):73–88, 2001. 810

[10] Bartholdi, Laurent. Endomorphic presentations of branch groups. *J. Algebra*, 268(2):419–443, 2003. 808

[11] Bartholdi, Laurent. A wilson group of non-uniformly exponential growth. *C. R. Math. Acad. Sci. Paris*, 336(7):549–554, 2003. 810

[12] Bartholdi, Laurent. Branch rings, thinned rings, tree enveloping rings. *Israel J. Math.*, 154:93–139, 2006. 802

[13] Bartholdi, Laurent and Erschler, Anna. Groups of given intermediate word growth. 2011. 811

[14] Bartholdi, Laurent and Erschler, Anna. Growth of permutational extensions. *Invent. Math.*, 189(2):431–455, 2012. 811

[15] Bartholdi, Laurent and Erschler, Anna. Ordering the space of finitely generated groups. 2013. 810

[16] Bartholdi, Laurent, Grigorchuk, Rostislav I., and Nekrashevych, Volodymyr V. From fractal groups to fractal sets. pages 25–118. 2003. 812

[17] Bartholdi, Laurent, Grigorchuk, Rostislav I., and Šunić, Zoran. Branch groups. pages 989–1112. 2003. 801, 808

[18] Bartholdi, Laurent, Kaimanovich, Vadim A., and Nekrashevych, Volodymyr V. On amenability of automata groups. *Duke Math. J.*, 154(3):575–598, 2010. 807

[19] Bartholdi, Laurent and Nekrashevych, Volodymyr V. Thurston equivalence of topological polynomials. *Acta Math.*, 197(1):1–51, 2006. 812

[20] Bartholdi, Laurent and Reznykov, Illya I. A mealy machine with polynomial growth of irrational degree. *Internat. J. Algebra Comput.*, 18(1):59–82, 2008. 810

[21] Bartholdi, Laurent, Reznykov, Illya I., and Sushchanskiĭ, Vitaly I. The smallest mealy automaton of intermediate growth. *J. Algebra*, 295(2):387–414, 2006. 810

[22] Bartholdi, Laurent and Šunić, Zoran. Some solvable automaton groups. pages 11–29. 2006. 804

[23] Bartholdi, Laurent and Virág, Bálint. Amenability via random walks. *Duke Math. J.*, 130(1):39–56, 2005. 805, 807

[24] Baumslag, Gilbert, Bridson, Martin R., Miller, Charles F., III, and Short, Hamish. Finitely presented subgroups of automatic groups and their isoperimetric functions. *J. London Math. Soc. (2)*, 56(2):292–304, 1997. 801

[25] Baumslag, Gilbert, Gersten, Stephen M., Shapiro, Michael, and Short, Hamish. Automatic groups and amalgams. *J. Pure Appl. Algebra*, 76(3):229–316, 1991. 791, 797, 801

[26] Bondarenko, Ievgen, Grigorchuk, Rostislav I., Kravchenko, Rostyslav, Muntyan, Yevgen, Nekrashevych, Volodymyr V., Savchuk, Dmytro, and Šunić, Zoran. On classification of groups generated by 3-state automata over a 2-letter alphabet. *Algebra Discrete Math.*, 1:1–163, 2008. 804, 805

[27] Boone, William W. The word problem. *Proc. Nat. Acad. Sci. U.S.A.*, 44:1061–1065, 1958. 796

[28] Brady, Noel. Finite subgroups of hyperbolic groups. *Internat. J. Algebra Comput.*, 10(4):399–405, 2000. 799

[29] Brazil, Marcus. Calculating growth functions for groups using automata. pages 1–18. 1995. 791

[30] Bridson, Martin R. Combings of groups and the grammar of reparameterization. *Comment. Math. Helv.*, 78(4):752–771, 2003. 800

[31] Bridson, Martin R. A note on the grammar of combings. *Internat. J. Algebra Comput.*, 15(3):529–535, 2005. 798

[32] Bridson, Martin R. Non-positive curvature and complexity for finitely presented groups. pages 961–987. 2006. 794

[33] Bridson, Martin R. and Gilman, Robert H. Formal language theory and the geometry of 3-manifolds. *Comment. Math. Helv.*, 71(4):525–555, 1996. 794

[34] Bridson, Martin R. and Haefliger, André. *Metric spaces of non-positive curvature*, volume 319 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999. 799

[35] Brink, Brigitte and Howlett, Robert B. A finiteness property and an automatic structure for coxeter groups. *Math. Ann.*, 296(1):179–190, 1993. 795

[36] Brunner, Andrew M. and Sidki, Said N. The generation of $\mathrm{GL}(n,)$ by finite state automata. *Internat. J. Algebra Comput.*, 8(1):127–139, 1998. 803

[37] Brunner, Andrew M. and Sidki, Said N. Wreath operations in the group of automorphisms of the binary tree. *J. Algebra*, 257(1):51–64, 2002. 803

[38] Burger, Marc and Mozes, Shahar. Finitely presented simple groups and products of trees. *C. R. Acad. Sci. Paris Sér. I Math.*, 324(7):747–752, 1997. 817

[39] Burger, Marc and Mozes, Shahar. Groups acting on trees: from local to global structure. *Inst. Hautes Études Sci. Publ. Math.*, 92:113–150 (2001), 2000. 817

[40] Burger, Marc and Mozes, Shahar. Lattices in product of trees. *Inst. Hautes Études Sci. Publ. Math.*, 92:151–194 (2001), 2000. 817

[41] Calegari, Danny and Fujiwara, Koji. Combable functions, quasimorphisms, and the central limit theorem. *Ergodic Theory Dynam. Systems*, 30(5):1343–1369, 2010. 798

[42] Campbell, Colin M., Robertson, Edmund F., Ruškuc, Nikola, and Thomas, Richard M. Automatic semigroups. *Theoret. Comput. Sci.*, 250(1-2):365–391, 2001. 794

[43] Cannon, James W. The combinatorial structure of cocompact discrete hyperbolic groups. *Geom. Dedicata*, 16(2):123–148, 1984. 791

[44] Cannon, James W., Floyd, William J., and Parry, Walter R. Finite subdivision rules. *Conform. Geom. Dyn.*, 5:153–196 (electronic), 2001. 813

[45] Cannon, James W., Floyd, William J., Parry, Walter R., and Pilgrim, Kevin M. Subdivision rules and virtual endomorphisms. *Geom. Dedicata*, 141:181–195, 2009. 813

[46] Cassaigne, Julien and Silva, Pedro V. Infinite words and confluent rewriting systems: endomorphism extensions. *Internat. J. Algebra Comput.*, 19(4):443–490, 2009. 800

[47] Cayley, Arthur. The theory of groups: Graphical representations. *Amer. J. Math.*, 1(2):174–176, 1878. 789

[48] Charney, Ruth. Artin groups of finite type are biautomatic. *Math. Ann.*, 292(4):671–683, 1992. 795

[49] Coornaert, Michel, Delzant, Thomas, and Papadopoulos, Athanase. *Géométrie et théorie des groupes*, volume 1441 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1990. Les groupes hyperboliques de Gromov. [Gromov hyperbolic groups]; With an English summary. 797

[50] Dahmani, François and Guirardel, Vincent. The isomorphism problem for all hyperbolic groups. *Geom. Funct. Anal.*, 21(2):223–300, 2011. 799

[51] Dehn, Max. über die topologie des dreidimensionalen raumes. *Math. Ann.*, 69(1):137–168, 1910. 789

[52] Dehn, Max. über unendliche diskontinuierliche gruppen. *Math. Ann.*, 71(1):116–144, 1911. 789

[53] Dehn, Max. Transformation der kurven auf zweiseitigen flächen. *Math. Ann.*, 72(3):413–421, 1912. 789

[54] Dehn, Max. *Papers on group theory and topology*. Springer-Verlag, New York, 1987. Translated from the German and with introductions and an appendix by John Stillwell; With an appendix by Otto Schreier. 789

[55] Duncan, Andrew and Gilman, Robert H. Word hyperbolic semigroups. *Math. Proc. Cambridge Philos. Soc.*, 136(3):513–524, 2004. 800

[56] Engel, Peter. *Geometric Crystallography*. Reidel Publishing Company, Dordrecht, 1986. 788

[57] Epstein, David B. A., Cannon, James W., Holt, Derek F., Levy, Silvio V. F., Paterson, Michael S., and Thurston, William P. *Word processing in groups*. Jones and Bartlett Publishers, Boston, MA, 1992. 791, 795

[58] Farb, Benson. Automatic groups: a guided tour. *Enseign. Math. (2)*, 38(3-4):291–313, 1992. 791

[59] Farb, Benson. Relatively hyperbolic groups. *Geom. Funct. Anal.*, 8(5):810–840, 1998. 799, 800

[60] Geoghegan, Ross. *Topological methods in group theory*, volume 243 of *Graduate Texts in Mathematics*. Springer, New York, 2008. 798

[61] Gersten, Stephen M. and Short, Hamish. Rational subgroups of biautomatic groups. *Ann. of Math. (2)*, 134(1):125–158, 1991. 797

[62] Gersten, Stephen M. and Short, Hamish. Small cancellation theory and automatic groups. ii. *Invent. Math.*, 105(3):641–662, 1991. 797

[63] Ghys, Étienne and de la Harpe, Pierre, editors. *Sur les groupes hyperboliques d'après Mikhael Gromov*, volume 83 of *Progress in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 1990. Papers from the Swiss Seminar on Hyperbolic Groups held in Bern, 1988. 797

[64] Gilman, Robert H. Groups with a rational cross-section. pages 175–183. 1987. 793

[65] Gilman, Robert H. On the definition of word hyperbolic groups. *Math. Z.*, 242(3):529–541, 2002. 798

[66] Gilman, Robert H., Holt, Derek F., and Rees, Sarah. Combing nilpotent and polycyclic groups. *Internat. J. Algebra Comput.*, 9(2):135–155, 1999. 800

[67] Glasner, Yair and Mozes, Shahar. Automata and square complexes. *Geom. Dedicata*, 111:43–64, 2005. 818

[68] Gluškov, Victor M. Abstract theory of automata. *Uspehi Mat. Nauk*, 16(5 (101)):3–62, 1961. 801

[69] Greendlinger, Martin. Dehn's algorithm for the word problem. *Comm. Pure Appl. Math.*, 13:67–83, 1960. 790

[70] Greendlinger, Martin. On dehn's algorithms for the conjugacy and word problems, with applications. *Comm. Pure Appl. Math.*, 13:641–677, 1960. 790

[71] Grigorchuk, Rostislav I. On burnside's problem on periodic groups. *Funktsional. Anal. i Prilozhen.*, 14(1):53–54, 1980. 804, 808

[72] Grigorchuk, Rostislav I. On the milnor problem of group growth. *Dokl. Akad. Nauk SSSR*, 271(1):30–33, 1983. 804, 810

[73] Grigorchuk, Rostislav I. Construction of p-groups of intermediate growth that have a continuum of factor-groups. *Algebra i Logika*, 23(4):383–394, 478, 1984. 804

[74] Grigorchuk, Rostislav I. Degrees of growth of finitely generated groups and the theory of invariant means. *Izv. Akad. Nauk SSSR Ser. Mat.*, 48(5):939–985, 1984. 804

[75] Grigorchuk, Rostislav I. Degrees of growth of p-groups and torsion-free groups. *Mat. Sb. (N.S.)*, 126(168)(2):194–214, 286, 1985. 804

[76] Grigorchuk, Rostislav I. Just infinite branch groups. pages 121–179. 2000. 808

[77] Grigorchuk, Rostislav I., Nekrashevych, Volodymyr V., and Sushchanskiĭ, Vitaly I. Automata, dynamical systems, and groups. *Tr. Mat. Inst. Steklova*, 231(Din. Sist., Avtom. i Beskon. Gruppy):134–214, 2000. 801

[78] Grigorchuk, Rostislav I. and Wilson, John S. A structural property concerning abstract commensurability of subgroups. *J. London Math. Soc. (2)*, 68(3):671–682, 2003. 806

[79] Grigorchuk, Rostislav I. and Wilson, John S. The uniqueness of the actions of certain branch groups on rooted trees. *Geom. Dedicata*, 100:103–116, 2003. 806

[80] Grigorchuk, Rostislav I. and Żuk, Andrzej. The lamplighter group as a group generated by a 2-state automaton, and its spectrum. *Geom. Dedicata*, 87(1-3):209–244, 2001. 804

[81] Grigorchuk, Rostislav I. and Żuk, Andrzej. On a torsion-free weakly branch group defined by a three state automaton. *Internat. J. Algebra Comput.*, 12(1-2):223–246, 2002. International Conference on Geometric and Combinatorial Methods in Group Theory and Semigroup Theory (Lincoln, NE, 2000). 805

[82] Gromov, Mikhael L. Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.*, 53:53–73, 1981. 809

[83] Gromov, Mikhael L. *Structures métriques pour les variétés riemanniennes*, volume 1 of *Textes Mathématiques [Mathematical Texts]*. CEDIC, Paris, 1981. Edited by J. Lafontaine and P. Pansu. 810

[84] Gromov, Mikhael L. Infinite groups as geometric objects. pages 385–392. 1984. 791

[85] Gromov, Mikhael L. Hyperbolic groups. pages 75–263. 1987. 797

[86] Groves, John R. J. and Hermiller, Susan M. Isoperimetric inequalities for soluble groups. *Geom. Dedicata*, 88(1-3):239–254, 2001. 800

[87] Gupta, Narain D. and Sidki, Said N. On the burnside problem for periodic groups. *Math. Z.*, 182(3):385–388, 1983. 804

[88] Gupta, Narain D. and Sidki, Said N. Some infinite p-groups. *Algebra i Logika*, 22(5):584–589, 1983. 804, 808

[89] Hermiller, Susan M., Holt, Derek F., and Rees, Sarah. Star-free geodesic languages for groups. *Internat. J. Algebra Comput.*, 17(2):329–345, 2007. 794

[90] Hermiller, Susan M., Holt, Derek F., and Rees, Sarah. Groups whose geodesics are locally testable. *Internat. J. Algebra Comput.*, 18(5):911–923, 2008. 794

[91] Hermiller, Susan M. and Meier, John. Tame combings, almost convexity and rewriting systems for groups. *Math. Z.*, 225(2):263–276, 1997. 794

[92] Hoffmann, Michael, Kuske, Dietrich, Otto, Friedrich, and Thomas, Richard M. Some relatives of automatic and hyperbolic groups. pages 379–406. 2002. 800

[93] Hoffmann, Michael and Thomas, Richard M. A geometric characterization of automatic semigroups. *Theoret. Comput. Sci.*, 369(1-3):300–313, 2006. 794

[94] Holt, Derek F. Word-hyperbolic groups have real-time word problem. *Internat. J. Algebra Comput.*, 10(2):221–227, 2000. 799

[95] Hopcroft, John. An n log n algorithm for minimizing states in a finite automaton. pages 189–196. 1971. 806

[96] Hurwitz, Adolf. *Vorlesungen über die Zahlentheorie der Quaternionen*. J. Springer, Berlin, 1919. 817

[97] Knuutila, Timo. Re-describing an algorithm by hopcroft. *Theoret. Comput. Sci.*, 250(1-2):333–363, 2001. 806

[98] Krohn, Kenneth B. and Rhodes, John L. Algebraic theory of machines. pages 341–384. 1963. 801, 813

[99] Leonov, Yurij G. The conjugacy problem in a class of 2-groups. *Mat. Zametki*, 64(4):573–583, 1998. 806

[100] Lyndon, Roger C. On dehn's algorithm. *Math. Ann.*, 166:208–228, 1966. 790

[101] Lyndon, Roger C. and Schupp, Paul E. *Combinatorial group theory*. Springer-Verlag, Berlin, 1977. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 89. 790, 795

[102] Lysënok, I. G. A set of defining relations for the grigorchuk group. *Mat. Zametki*, 38(4):503–516, 634, 1985. 808

[103] Macedońska, Olga, Nekrashevych, Volodymyr V., and Sushchanskiĭ, Vitaly I. Commensurators of groups and reversible automata. *Dopov. Nats. Akad. Nauk Ukr. Mat. Prirodozn. Tekh. Nauki*, 12:36–39, 2000. 814

[104] Miller, Charles F., III. *On group-theoretic decision problems and their classification*. Princeton University Press, Princeton, N.J., 1971. Annals of Mathematics Studies, No. 68. 797

[105] Milnor, John W. Growth of finitely generated solvable groups. *J. Differential Geometry*, 2:447–449, 1968. 810

[106] Milnor, John W. Problem 5603. *Amer. Math. Monthly*, 75:685–686, 1968. 810

[107] Mosher, Lee. Mapping class groups are automatic. *Ann. of Math. (2)*, 142(2):303–384, 1995. 795

[108] Nekrashevych, Volodymyr V. *Self-similar groups*, volume 117 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2005. 801, 807, 812

[109] Nekrashevych, Volodymyr V. Free subgroups in groups acting on rooted trees. *Groups, Geom. and Dynamics*, 4(4):847–862, 2010. 808

[110] Novikov, Pëtr S. *Ob algoritmičeskoĭ nerazrešimosti problemy toždestva slov v teorii grupp*. Trudy Mat. Inst. im. Steklov. no. 44. Izdat. Akad. Nauk SSSR, Moscow, 1955. 796

[111] Ol′shanskiĭ, Alexander Yu. Almost every group is hyperbolic. *Internat. J. Algebra Comput.*, 2(1):1–17, 1992. 798

[112] Rattaggi, Diego. A finitely presented torsion-free simple group. *J. Group Theory*, 10(3):363–371, 2007. 817

[113] Rips, Eliyahu. Subgroups of small cancellation groups. *Bull. London Math. Soc.*, 14(1):45–47, 1982. 799

[114] Rozhkov, Alexander V. The conjugacy problem in an automorphism group of an infinite tree. *Mat. Zametki*, 64(4):592–597, 1998. 806

[115] Shalom, Yehuda and Tao, Terence. A finitary version of gromov's polynomial growth theorem. *Geom. Funct. Anal.*, 20(6):1502–1547, 2010. 811

[116] Sidki, Said N. A primitive ring associated to a burnside 3-group. *J. London Math. Soc. (2)*, 55(1):55–64, 1997. 802

[117] Sidki, Said N. Automorphisms of one-rooted trees: growth, circuit structure, and acyclicity. *J. Math. Sci. (New York)*, 100(1):1925–1943, 2000. Algebra, 12. 806

[118] Sidki, Said N. Finite automata of polynomial growth do not generate a free group. *Geom. Dedicata*, 108:193–204, 2004. 806

[119] Sidki, Said N. Tree-wreathing applied to generation of groups by finite automata. *Internat. J. Algebra Comput.*, 15(5-6):1205–1212, 2005. 803

[120] Silva, Pedro V. and Steinberg, Benjamin. A geometric characterization of automatic monoids. *Q. J. Math.*, 55(3):333–356, 2004. 794

[121] Silva, Pedro V. and Steinberg, Benjamin. On a class of automata groups generalizing lamplighter groups. *Internat. J. Algebra Comput.*, 15(5-6):1213–1234, 2005. 813

[122] Tartakovskiĭ, Vladimir A. Solution of the word problem for groups with a k-reduced basis for $k > 6$. *Izvestiya Akad. Nauk SSSR. Ser. Mat.*, 13:483–494, 1949. 790

[123] Tits, Jacques. Free subgroups in linear groups. *J. Algebra*, 20:250–270, 1972. 810

[124] Vince, Andrew. Radix representation and rep-tiling. In *Proceedings of the Twenty-fourth Southeastern International Conference on Combinatorics, Graph Theory, and Computing (Boca Raton, FL, 1993)*, volume 98, pages 199–212. 1993. 814

[125] Vorobets, Mariya and Vorobets, Yaroslav. On a free group of transformations defined by an automaton. *Geom. Dedicata*, 124:237–249, 2007. 815

[126] Vorobets, Mariya and Vorobets, Yaroslav. On a series of finite automata defining free transformation groups. *Groups Geom. Dyn.*, 4(2):377–405, 2010. 815

[127] Wilson, John S. On exponential growth and uniformly exponential growth for groups. *Invent. Math.*, 155(2):287–303, 2004. 810

[128] Wise, Daniel T. Complete square complexes. *Comment. Math. Helv.*, 82(4):683–724, 2007. 817

[129] Wolf, Joseph A. Growth of finitely generated solvable groups and curvature of riemanniann manifolds. *J. Differential Geometry*, 2:421–446, 1968. 810

Automata in Number Theory

Boris Adamczewski[1] *and Jason Bell*[2]

[1] CNRS, Aix-Marseille Université,
Institut Mathématiques de Marseille
Centre de Mathématiques et Informatique
Technopôle Château-Gombert
39, rue F. Joliot Curie
13453 Marseille Cedex 13, France
email: Boris.Adamczewski@math.cnrs.fr

[2] Department of Pure Mathematics
University of Waterloo
Waterloo, ON, Canada
N2L 3G1
email: jpbell@uwaterloo.ca

May 18, 2015 22 h 14

2010 Mathematics Subject Classification: 11B85,11J81,11J87

Key words: Finite automata, automatic sequences, automatic sets, prime numbers, algebraic numbers, Diophantine approximation, linear recurrences, Laurent series, generalized power series.

Contents

1 Introduction

Among infinite sequences or infinite sets of integers, some are well-behaved, such as periodic sequences and arithmetic progressions, whereas others, such as random sequences and random sets, are completely chaotic and cannot be described in a simple way. Finite automata are one of the most basic models of computation and thus lie at the bottom of the hierarchy associated with Turing machines. Such machines can naturally be used to generate sequences with values over a finite set, and also as devices to recognize certain subsets of the integers.

One of the main interests of these automatic sequences/sets arises from the fact that they are in many ways very well-behaved without necessarily being trivial. One can thus consider that they lie somewhere between order and chaos, even if, in many respects, they are well-behaved.

In this chapter, we survey some of the connections between automatic sequences/sets and number theory. Several substantial advances have recently been made in this area and we give an overview of some of these new results. This includes discussions about prime numbers, the decimal expansion of algebraic numbers, the search for an analogue of the Skolem-Mahler-Lech theorem in positive characteristic and the description of an algebraic closure of the field $\mathbb{F}_p(t)$.

2 Automatic sequences and automatic sets of integers

In this section, we recall some basic facts about automatic sequences and automatic sets of integers. The main reference on this topic is the book of Allouche and Shallit [4]. An older reference is Eilenberg [13]. In [13] k-automatic sets are called k -recognizable.

2.1 Automatic sequences

Let $k \geqslant 2$ be an integer. An infinite sequence $(a_n)_{n\geqslant 0}$ is said to be k-automatic if a_n is a finite-state function of the base-k representation of n. This means that there exists a deterministic finite automaton with output (DFAO) taking the base-k expansion of n as input and producing the term a_n as output. We say that a sequence is generated by a finite automaton, or for short is automatic, if it is k-automatic for some $k \geqslant 2$.

A more concrete definition of k-automatic sequences can be given as follows. Let A_k denote the alphabet $\{0, 1, \ldots, k-1\}$. By definition, a k-automaton is a 6-tuple

$$\mathcal{A} = (Q, A_k, \delta, q_0, \Delta, \tau)\,,$$

where Q is a finite set of states, $\delta : Q \times A_k \to Q$ is the transition function, q_0 is the initial state, Δ is the output alphabet and $\tau : Q \to \Delta$ is the output function. For a state q in Q and for a finite word $w = w_1 w_2 \cdots w_n$ on the alphabet A_k, we define $\delta(q, w)$ recursively by $\delta(q, w) = \delta(\delta(q, w_1 w_2 \cdots w_{n-1}), w_n)$. Let $n \geqslant 0$ be an integer and let $w_r w_{r-1} \cdots w_1 w_0$ in $(A_k)^{r+1}$ be the base-k expansion of n. Thus $n = \sum_{i=0}^{r} w_i k^i := [w_r w_{r-1} \cdots w_0]_k$. We denote by $w(n)$ the word $w_r w_{r-1} \cdots w_0$. Then a sequence $(a_n)_{n\geqslant 0}$ is said to be k-automatic if there exists a k-automaton $\mathcal{A}$ such that $a_n = \tau(\delta(q_0, w(n)))$ for all $n \geqslant 0$.

Example 2.1. The Thue–Morse sequence $t := (t_n)_{n\geqslant 0}$ is probably the most famous example of automatic sequences. It is defined as follows: $t_n = 0$ if the sum of the binary digits of n is even, and $t_n = 1$ otherwise. We thus have

$$t = 01101001100101\cdots\,.$$

It is easy to check that the Thue–Morse sequence can be generated by the following finite 2-automaton: $\mathcal{A} = (\{A, B\}, \{0, 1\}, \delta, A, \{0, 1\}, \tau)$, where $\delta(A, 0) = \delta(B, 1) = A$, $\delta(A, 1) = \delta(B, 0) = B$, $\tau(A) = 0$ and $\tau(B) = 1$.

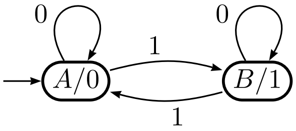


Figure 1. A DFAO generating Thue–Morse word.

Example 2.2. Let $w = (w_n)_{n\geqslant 0}$ be the 3-automatic sequence generated by the DFAO represented in Figure 2. Note that though this 3-automaton has only two states, it seems to be uneasy to give a simple expression of w_n in function of the ternary expansion of n.

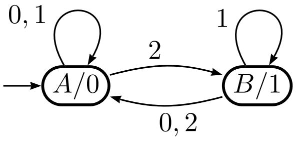


Figure 2. A DFAO generating the sequence w.

2.1.1 Morphisms of free monoids For a finite set A, we let A^* denote the free monoid generated by A. The empty word ε is the identity element of A^*. Let A and B be two finite sets. A map from A to B^* extends uniquely to a homomorphism between the free monoids A^* and B^*. We call such a homomorphism from A^* to B^* a *morphism*. If there is a positive integer k such that each element of A is mapped to a word of length k, then the morphism is called *k-uniform* or simply *uniform*. A *coding* is a 1-uniform morphism.

A morphism σ from A^* to itself is said to be *prolongable* if there exists a letter a such that $\sigma(a) = aw$, where the word w is such that $\sigma^n(w)$ is a nonempty word for every $n \geqslant 0$. In that case, the sequence of finite words $(\sigma^n(a))_{n\geqslant 0}$ converges in $A^\omega = A^* \cup A^{\mathbb{N}}$, endowed with its usual topology, to an infinite word denoted $\sigma^\omega(a)$. This infinite word is clearly a fixed point for σ (extended by continuity to infinite words) and we say that $\sigma^\omega(a)$ is generated by the morphism σ.

For instance, the morphism τ defined over the alphabet $\{0, 1\}$ by $\tau(0) = 01$ and $\tau(1) = 10$ is a 2-uniform morphism which generates the Thue–Morse sequence

$$t = \tau^\omega(0) = 01101001100101\cdots .$$

Uniform morphisms and automatic sequences are strongly connected, as the following classical result of Cobham shows [10]. A notable consequence of Theorem 2.1 is that finite automata are Turing machines that produce sequences in linear time.

Theorem 2.1 (Cobham). *An infinite word is k-automatic if and only if it is the image by a coding of a word that is generated by a k-uniform morphism.*

Example 2.3. Let us consider the 3-uniform morphism ω defined over the monoid $\{0, 1, 2\}^*$ by $\omega(0) = 012$, $\omega(1) = 020$, and $\omega(2) = 021$. This morphism has a unique fixed point

$$x = \omega^\omega(0) = 012020021012021012012\cdots .$$

Letting ϕ denote the coding that maps 0 and 1 to 0, and 2 to 1, we thus obtain that

$$y := y_1 y_2 \cdots := \phi(x) = 001010010001010001001\cdots$$

is a 3-automatic word.

Example 2.4. The word w defined in Example 2.2 is the unique fixed point generated by the binary morphism ψ satisfying $\psi(0) = 001$ and $\psi(1) = 010$.

2.1.2 Kernels An important notion in the study of k-automatic sequences is the notion of k-kernel. The k-kernel of a sequence $a = (a_n)_{n\geqslant 0}$ is defined as the set

$$N_k(a) = \left\{(a_{k^i n+j})_{n\geqslant 0} \mid i \geqslant 0,\, 0 \leqslant j < k^i\right\} .$$

This notion gives rise to another useful characterization of k-automatic sequences which was first proved by Eilenberg in [13].

Theorem 2.2 (Eilenberg). *A sequence is k-automatic if and only if its k-kernel is finite.*

Example 2.5. The 2-kernel of the Thue–Morse sequence t has only two elements t and the sequence $\bar{t}$ obtained by exchanging the symbols 0 and 1 in t.

2.2 Automatic sets of integers

Another important aspect of finite automata is that they can naturally be used as a device to recognize sets of integers.

2.2.1 Automatic subsets of $\mathbb{N}$ A set $\mathcal{N} \subset \mathbb{N}$ is said to be recognizable by a finite k-automaton, or for short *k-automatic*, if the characteristic sequence of $\mathcal{N}$, defined by $a_n = 1$ if $n \in \mathcal{N}$ and $a_n = 0$ otherwise, is a k-automatic sequence. This means that there exists a finite k-automaton that reads as input the base-k expansion of n and accepts this integer (producing as output the symbol 1) if n belongs to $\mathcal{N}$, otherwise this automaton rejects the integer n, producing as output the symbol 0.

Example 2.6. The simplest automatic sets are arithmetic progressions. Moreover, arithmetic progressions have the very special property of being k-automatic sets for every integer $k \geqslant 2$ (see Cobham's theorem in Chapter 26).

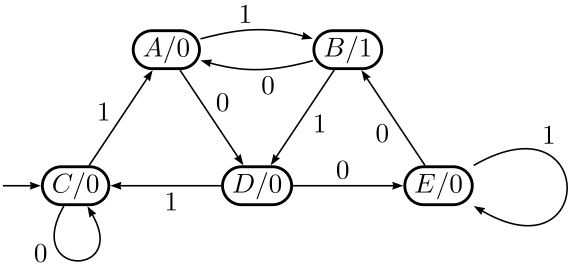


Figure 3. A 2-DFAO recognizing the arithmetic progression $5\mathbb{N} + 3$.

Example 2.7. The set $\{1, 2, 4, 8, 16, \ldots\}$ formed by the powers of 2 is also a typical example of a 2-automatic set.

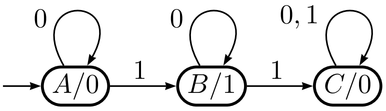


Figure 4. A 2-DFAO recognizing the powers of 2.

Example 2.8. In the same spirit, the set formed by taking all integers that can be expressed as the sum of tat most two powers of 3 is 3-automatic (see Figure 5).

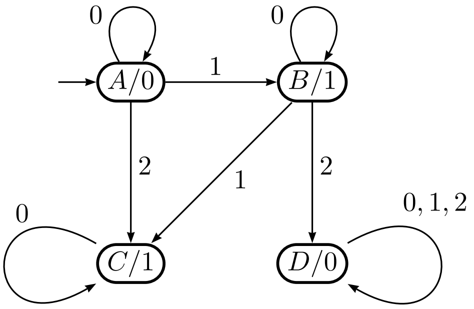


Figure 5. A 3-DFAO recognizing those integers that are the sum of tat most wo powers of 3.

There are also much stranger automatic sets. For instance, the set of integers whose binary expansion has an odd number of digits, does not contain three consecutive 1's, and contains an even number of two consecutive 0's is a 2-automatic set. Furthermore, the class of k-automatic sets is closed under various natural operations such as intersection, union and complement. On the other hand, some classical sets of integers such as the set of prime numbers and the set of perfect squares cannot be recognized by a finite automaton (see Theorem 3.1 and [42, 37]).

2.2.2 Automatic subsets of $\mathbb{N}^d$ and multidimensional automatic sequences Salon [44] extended the notion of automatic sets to include subsets of $\mathbb{N}^d$, where $d \geqslant 1$. To describe Salon's construction, we let A_k denote the alphabet $\{0, 1, \ldots, k-1\}$. We then consider an automaton

$$\mathcal{A} = \left(Q, A_k^d, \delta, q_0, \Delta, \tau\right) ,$$

where Q is a finite set of states, $\delta : Q \times A_k^d \to Q$ is the transition function, q_0 is the initial state, Δ is the output alphabet and $\tau : Q \to \Delta$ is the output function. Just as in the one-dimensional case, for a state q in Q and for a finite word $w = w_1 w_2 \cdots w_n$ on the alphabet A_k^d, we recursively define $\delta(q, w)$ by $\delta(q, w) = \delta(\delta(q, w_1 w_2 \cdots w_{n-1}), w_n)$. We call such an automaton a *d-dimensional k-automaton.*

We identify $\left(A_k^d\right)^*$ with the subset of $(A_k^*)^d$ consisting of all d-tuples $(u_1, \ldots, u_d)$ such that $u_1, \ldots, u_d$ all have the same length. Each nonnegative integer n can be written uniquely as

$$n \;=\; \sum_{j=0}^{\infty} e_j(n) k^j \, ,$$

in which $e_j(n) \in \{0, \ldots, k-1\}$ and $e_j(n) = 0$ for all sufficiently large j. Let $(n_1, \ldots, n_d)$ be a d-tuple of nonnegative integers and let

$$h := \max\left(\lfloor \log n_1 / \log\, k \rfloor, \cdots, \lfloor \log n_d / \log\, k \rfloor\right) \, ,$$

that is, if a_i represents the number of digits in the base-k expansion of n_i, then $h+1$ is the maximum of $a_1, \dots, a_r$. We can then produce an element

$$w_k(n_1, \dots, n_d) := (w_1, \dots, w_d) \in \left(A_k^d\right)^*$$

corresponding to $(n_1, \dots, n_d)$ by defining

$$w_i := e_h(n_i)e_{h-1}(n_i)\cdots e_0(n_i)\,.$$

In other words, we are taking the base-k expansions of $n_1, \dots, n_r$ and then "padding" the expansions of each n_i at the beginning with 0's if necessary to ensure that each expansion has the same length. We say that a map $f : \mathbb{N}^d \to \Delta$ is k-automatic if there is a d-dimensional k-automaton $\mathcal{A} = \left(Q, A_k^d, \delta, q_0, \Delta, \tau\right)$ such that

$$f(n_1, \dots, n_d) = \tau(\delta(q_0, w_d(n_1, \dots, n_d)))\,.$$

Similarly, we define a k-automatic subset of $\mathbb{N}^d$ to be a subset S such that the characteristic function of S, $f : \mathbb{N}^d \to \{0,1\}$, defined by $f(n_1, \dots, n_d) = 1$ if $(n_1, \dots, n_d) \in S$; and $f(n_1, \dots, n_d) = 0$, otherwise, is k-automatic.

Example 2.9. Let $f : \mathbb{N}^2 \to \{0,1\}$ be defined by $f(n,m) = 1$ if the sum of the binary digits of n added to the sum of the binary digits of m is even, and $f(n,m) = 0$ otherwise. Then $f(m,n)$ is a 2-automatic map. One can check that f can be generated by the following 2-dimensional 2-automaton: $\mathcal{A} = \left(\{A,B\}, \{0,1\}^2, \delta, A, \{0,1\}, \tau\right)$, where $\delta(A,(0,0)) = \delta(A,(1,1)) = \delta(B,(1,0)) = \delta(B,(0,1)) = A$, $\delta(A,(1,0)) = \delta(A,(0,1)) = \delta(B,(0,0)) = \delta(B,(1,1)) = B$, $\tau(A) = 1$ and $\tau(B) = 0$.

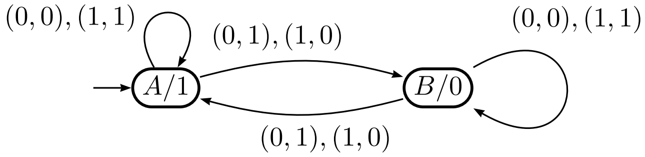


Figure 6. A DFAO generating the map f defined in Example 2.9.

Just as k-automatic sequences can be characterized by the finiteness of the k-kernel, multidimensional k-automatic sequences have a similar characterization.

Definition 2.1. Let d be a positive integer and let Δ be a finite set. We define the *k-kernel* of a map $f : \mathbb{N}^d \to \Delta$ to be the collection of all maps of the form

$$g(n_1, \dots, n_d) := f(k^a n_1 + b_1, \dots, k^a n_d + b_d)$$

where $a \geqslant 0$ and $0 \leqslant b_1, \dots, b_d < k^a$.

For example, if $f : \mathbb{N}^2 \to \{0,1\}$ is the map defined in Example 2.9, then the 2-kernel of f consists of the 2 maps $f_1(m,n) := f(m,n)$, $f_2(m,n) = f(2m+1, 2n)$. Just as Eilenberg [13] showed that being k-automatic is equivalent to having a finite k-kernel for k-automatic sequences, Salon [44, Theorem 1] showed that a similar characterization of multidimensional k-automatic maps holds.

Theorem 2.3 (Salon). *Let d be a positive integer and let Δ be a finite set. A map $f : \mathbb{N}^d \to \Delta$ is k-automatic if and only if its k-kernel is finite.*

3 Prime numbers and finite automata

In this section, we briefly discuss some results concerning primes and finite automata.

3.1 Primes and randomness

An efficient way to produce conjectures about prime numbers comes from the so–called Cramér probabilistic model (see [11, 50, 51]). It is based on the principle that the set $\mathcal{P}$ of prime numbers behaves roughly like a random sequence, in which an integer of size about n has—as inspired by the Prime Number Theorem—a 1 in $\log n$ chance of being prime. Of course, this probabilistic model has some limitations: for instance prime numbers are all odd with only one exception (see [39] for more about such limitations). Thus the set of prime numbers should be thought of as being a hybrid set rather than as a pseudorandom set (see the discussion in [52]). However, the Cramér model allows one to predict precise answers concerning occurrences of large gaps between consecutive prime numbers and concerning small gaps between primes (twin prime conjecture) and of some special patterns in $\mathcal{P}$ such as arithmetic progressions (Hardy–Littlewood conjectures). Some spectacular breakthrough were made recently in the two latter topics. See in particular [19] and [18].

A consequence of this probabilistic way of thinking is that the set $\mathcal{P}$ should be sufficiently random that it cannot be recognized by a finite automaton. This result was in fact proved to be true by Minsky and Papert [37] in 1966.

Theorem 3.1 (Minsky and Papert). *The set of prime numbers cannot be recognized by a finite automaton.*

Schützenberger [46] (see also [22]) even proved the stronger result that an automatic set always contains infinitely many composite numbers.

Theorem 3.2 (Schützenberger). *No infinite subset of the set of prime numbers can be recognized by a finite automaton.*

The intriguing question we are now left with is: *how can we prove that a set of integers is not automatic?* There are actually different approaches: one can use the k-kernel and show that it is infinite or one can use some density properties (the logarithmic frequency of an automatic set exists, also if an automatic set has a positive density then it is rational). Another very efficient tool is the so-called pumping lemma, which is recalled below. For more details about the different ways of proving that a sequence is not automatic, we refer the reader to [4].

Lemma 3.3 (Pumping lemma). *Let $\mathcal{N} \subset \mathbb{N}$ be a k-automatic set. Then for every sufficiently large integer n in $\mathcal{N}$, there exist finite words w_1, w_2 and w_3, with $|w_2| \geqslant 1$, such that $n = [w_1 w_2 w_3]_k$ and $[w_1 w_2^i w_3]_k$ belongs to $\mathcal{N}$ for all $i \geqslant 1$.*

Sketch of proof. Let $n = [a_r a_{r-1} \cdots a_0]_k$ be an element of $\mathcal{N}$ and assume that r is larger than the number of states in the underlying automaton. By the pigeonhole principle, there is a state that is encountered twice when reading the input $a_0 a_1 \cdots a_r$, say just after reading a_i and a_j, $i < j$. Then setting $w_1 = a_r \cdots a_{j+1}$, $w_2 = a_j \cdots a_{i+1}$ and $w_3 = a_i \cdots a_0$, gives the result. □

Proof of Theorem 3.2. Let us assume that $\mathcal{N}$ is an infinite k-automatic set consisting only of prime numbers. Let p be an element of $\mathcal{N}$ that is sufficiently large to apply the pumping lemma. By the pumping lemma, there exist finite words w_1, w_2 and w_3, with $|w_2| \geqslant 1$, such that $p = [w_1 w_2 w_3]_k$ and such that all integers of the form $[w_1 w_2^i w_3]_k$, with $i \geqslant 1$, belong to $\mathcal{N}$. However, it is not difficult to see, by using Fermat's little theorem, that $[w_1 w_2^p w_3]_k \equiv [w_1 w_2 w_3]_k \mod p$ and thus $[w_1 w_2^p w_3]_k \equiv 0 \mod p$. It follows that the integer $[w_1 w_2^p w_3]_k$ belongs to $\mathcal{N}$ but is not a prime number. Hence we obtain a contradiction. □

3.2 Primes in automatic sets

We have just seen that the set of all prime numbers is not automatic. However, it is believed that many automatic sets should contain infinitely many prime numbers. The most basic example of such a result is the famous Dirichlet theorem.

Theorem 3.4 (Dirichlet). *Let a and b be two relatively prime positive integers. Then the arithmetic progression $a\mathbb{N} + b$ contains infinitely many primes.*

Note that the special case of the arithmetic progression $2\mathbb{N} + 1$ was known by Euclid and his famous proof that there are infinitely many prime numbers. A more complete discussion about Dirichlet's theorem can be found in [40]. Beyond Dirichlet's theorem, the more general result concerning automatic sets and prime numbers is Theorem 3.5 from [16]. Recall that an automaton is *irreducible* if for all pairs of states (A, B) there is a path from A to B. Recall also that a positive integer is an *r-almost prime* if it is the product of at most r prime numbers. It is well-known that results about almost-primes are much easier to prove than those concerning primes (compare for instance Chen's theorem [7] with known results about the twin prime conjecture and the Goldbach conjecture).

Theorem 3.5 (Fouvry and Mauduit). *Given an automatic set $\mathcal{N} \subset \mathbb{N}$ associated with an irreducible automaton, there exists a positive integer r such that $\mathcal{N}$ contains infinitely many r-almost primes.*

Theorem 3.5 is not too difficult to prove using results similar to Chen's theorem. In contrast, to prove that there are infinitely many primes in sparse automatic sets such as $\{2^n - 1,\ n \geqslant 1\}$ and $\{2^n + 1,\ n \geqslant 1\}$ appears to be extremely difficult. This would solve two long-standing conjectures about the existence of infinitely many Fermat primes and Mersenne primes.

3.3 A problem of Gelfond: the sum of digits of prime numbers

Given a natural number n and a base b, we let $s_b(n)$ denote the sum of the digits of n in base b. Given two natural numbers a and m with $0 \leqslant a < m$ and $(m, b-1) = 1$, one can then look at the set of positive integers n such that $s_b(n) \equiv a \pmod{m}$. This set is known to be recognizable by a finite b-automaton. In 1968, Gelfond [17] asked about the collection of prime numbers that belong to this set. Theorem 3.5 implies that such a set contains infinitely r-almost primes for some r, but until recently it was still not known whether it contains infinitely many primes. Remarkably, Mauduit and Rivat [35] proved a much stronger result that gives the exact proportion of primes that belong to this automatic set. As usual with analytic number theory, the proof of their result—which relies on strong estimates of exponential sums—is long and difficult. As an example, an immediate corollary of the work of Mauduit and Rivat is that half of prime numbers belong to the Thue–Morse set $\{1, 2, 4, 7, 8, 11, 13, \ldots\}$.

Theorem 3.6 (Mauduit and Rivat). *One has*

$$\lim_{N\to\infty} \frac{\{0 \leqslant n \leqslant N, n \in \mathcal{P} \textit{ and } s_2(n) \equiv 1 \mod 2\}}{\{0 \leqslant n \leqslant N, n \in \mathcal{P}\}} = \frac{1}{2} \cdot$$

4 Expansions of algebraic numbers in integer bases

The decimal expansions of classical constants like $\sqrt{2}$, π and e appear to be very mysterious and have baffled mathematicians for a long time. Numerical observations suggest that a complex underlying structure exists and several famous mathematicians have suggested possible rigorous definitions to try to formalize what "complex structure" actually means (see for instance [6, 38, 23]). These mathematicians were mainly influenced by notions from probability theory, dynamical systems, or theoretical computer science. These pioneering works lead us to a cluster of interesting conjectures concerning expansions of irrational periods in integer bases. However, even some of the simplest questions one can ask about the decimal expansions of classical irrational constants are still far out of reach.

The seminal work of Turing [53] gives rise to a rough classification of real numbers. On one side we find computable real numbers; that is, real numbers whose binary (or more generally base-b) expansion can be produced by a Turing machine, while on the other side lie uncomputable real numbers which, in some sense, "evade computers." Though most real numbers belong to the second class (the first one being countable), classical mathematical constants are usually computable. Following the pioneering ideas of Turing, Hartmanis and Stearns [23] proposed the emphasis of the quantitative aspect of the notion of computability, and to take into account the number $T(n)$ of operations needed by a (multitape) Turing machine to produce the first n digits of the expansion. In this regard, a real number is considered to be simple if its base-b expansion can be produced quickly by a Turing machine. A general problem is then to determine where our mathematical constants take place in such a classification. It is a source of challenging open questions such as the Hartmanis–Stearns problem which asks whether there exists an irrational algebraic number computable in linear time; that is, with $T(n) = O(n)$.

In 1968, Cobham [9] suggested to restrict this problem to a particular class of Turing machines, namely to the case of finite automata. Several attempts at a resolution to this problem are due to Cobham in 1968 [9] and to Loxton and van der Poorten [29, 30] during the Eighties. Both of these works are based on the so-called Mahler transcendence method [31]. The aim of this section is to give a proof, due to Adamczewski and Bugeaud [2], of Cobham's conjecture following a completely different approach based on a deep Diophantine result known as the Schmidt subspace theorem.

Theorem 4.1 (Adamczewski and Bugeaud). *The base-b expansion of an algebraic irrational number cannot be generated by a finite automaton.*

4.1 Rational approximations and transcendence of some automatic numbers

Given an integer $k \geqslant 2$, a real number is said to be *k-automatic* if there exists an integer $b \geqslant 2$ such that its base-b expansion is a k-automatic sequence.

4.1.1 Liouville's inequality In 1844, Liouville [28] proved that transcendental numbers exist. Moreover, he constructed explicit examples of such numbers. His approach relies on the famous Liouville inequality recalled below.

Proposition 4.2 (Liouville's inequality). *Let ξ be an algebraic number of degree $d \geqslant 2$. Then there exists a positive real number c_ξ such that*

$$\left|\xi - \frac{p}{q}\right| \geqslant \frac{c_\xi}{q^d}$$

for every rational number p/q with $q \geqslant 1$.

Proof. Let P denote the minimal polynomial of ξ, P' its derivative, and set

$$c_\xi := 1/(1 + \max_{|\xi - x|<1} |P'(x)|)\,.$$

If $|\xi - p/q| \geqslant 1$, then our choice of c_ξ ensures that $|\xi - p/q| \geqslant c_\xi/q^d$.

Let us now assume that $|\xi - p/q| < 1$. Since P is the minimal polynomial of ξ, it does not vanish at p/q and $q^d P(p/q)$ is a nonzero integer. Consequently,

$$|P(p/q)| \geqslant \frac{1}{q^d}\,\cdot \tag{4.1}$$

Since $|\xi - p/q| < 1$, the mean value theorem implies the existence of a real number t in $(p/q - 1, p/q + 1)$ such that

$$|P(p/q)| = |P(\xi) - P(p/q)| = \left|\xi - \frac{p}{q}\right| \cdot |P'(t)|\,,$$

which ends the proof in view of Inequality (4.1) and the definition of c_ξ. □

Liouville's inequality can be used to easily construct transcendental numbers. Indeed, if ξ is an irrational real number such that for every integer $d \geqslant 2$ there exists a rational number p/q satisfying $|\xi - p/q| < q^{-d}$, then ξ is transcendental. Real numbers enjoying this property are termed *Liouville numbers*. The number $\mathcal{L}$ below is a typical example of Liouville number, often considered as the first example of a transcendental number.

Theorem 4.3 (Liouville). *The real number*

$$\mathcal{L} := \sum_{n=1}^{+\infty} \frac{1}{10^{n!}}$$

is transcendental.

Proof of Theorem 4.3. Let $j \geqslant d \geqslant 2$ be two integers. Then, there exists an integer p_j such that

$$\frac{p_j}{10^{j!}} = \sum_{n=1}^{j} \frac{1}{10^{n!}} \cdot$$

Observe that

$$\left|\mathcal{L} - \frac{p_j}{10^{j!}}\right| = \sum_{n>j} \frac{1}{10^{n!}} < \frac{2}{10^{(j+1)!}} < \frac{1}{(10^{j!})^d} \cdot$$

It then follows from Proposition 4.2 that $\mathcal{L}$ cannot be algebraic of degree less than d. Since d is arbitrary, $\mathcal{L}$ is transcendental. □

Adamczewski and Cassaigne [3] confirmed a conjecture of Shallit by proving that no Liouville number can be generated by a finite automaton. In other words, there is no automatic real number that can be proved to be transcendental by the elementary approach described above. However, we will see in the sequel how some deep improvements of Liouville's inequality can be used in a similar way to prove the transcendence of automatic numbers.

4.1.2 Roth's theorem

The following famous improvement of Liouville's inequality was established by Roth [43] in 1955. This result is the best possible in the sense that the exponent $2 + \varepsilon$ in (4.2) cannot be lowered.

Theorem 4.4 (Roth). *Let ξ be a real algebraic number and let ε be a positive real number. Then there are only a finite number of rational numbers p/q such that $q \geqslant 1$ and*

$$\left|\xi - \frac{p}{q}\right| < \frac{1}{q^{2+\varepsilon}} \cdot \tag{4.2}$$

We give an immediate application of Roth's theorem to the transcendence of automatic real numbers.

Corollary 4.5. *For every integer $k \geqslant 3$, the k-automatic real number*

$$\sum_{n=1}^{+\infty} \frac{1}{10^{k^n}}$$

is transcendental.

Proof. Use the same argument as in the proof of Theorem 4.3. □

However, Roth's theorem gives no information on the arithmetical nature of the 2-automatic real number

$$\sum_{n=1}^{+\infty} \frac{1}{10^{2^n}} \,.$$

Let us now consider the word w defined in Example 2.2. We associate with w the real number

$$\xi_w := \sum_{n \geqslant 0} \frac{w_n}{10^{n+1}} = 0.001\,001\,010\,001\,001\,010\,001 \cdots \,.$$

A characteristic of the number $\mathcal{L}$ and the numbers defined in Corollary 4.5 is that large blocks of zeros appear in their decimal expansion much more frequently than one would expect if the numbers we were dealing with were randomly selected. In contrast, the decimal expansion of ξ_w contains no occurrence of more than three consecutive zeros. However, the combinatorial structure of w can be used to reveal more hidden good rational approximations to ξ_w that imply the following result.

Theorem 4.6. *The* 2*-automatic real number* ξ_w *is transcendental.*

Proof. Let ψ be the binary morphism defined in Example 2.4. For every positive integer j, set $u_j := \psi^j(0)$, $s_j := |u_j|$ and let us consider the rational number ρ_j defined by

$$\rho_j := 0.u_j^{\omega} \,.$$

An easy computation shows that there exists an integer p_j such that

$$\rho_j = \frac{p_j}{10^{s_j} - 1} \,. \tag{4.3}$$

The rational number ρ_j turns out to be a very good approximation to ξ_w. Indeed, by definition of w, the decimal expansion of ξ_w begins with $\psi^j(0)\psi^j(0)\psi^{j-1}(0)$, which is also a prefix of u_j^{ω}. Consequently, the first $(2 + 1/3)s_j = 7 \cdot 3^{j-1}$ digits in the decimal expansion of ξ_w and of ρ_j are the same. We thus obtain that

$$|\xi_w - \rho_j| < 10^{-(2+1/3)s_j} \,. \tag{4.4}$$

Consequently, we infer from (4.4) and (4.3) that

$$\left| \xi_w - \frac{p_j}{10^{s_j} - 1} \right| < \frac{1}{(10^{s_j} - 1)^{2.3}} \,.$$

Furthermore, the rational numbers ρ_j are all different since $\psi^n(0)$ is not a prefix of the infinite word $(\psi^m(0))^{\omega}$ when $n > m$. It thus follows from Roth's theorem that ξ_w is transcendental. □

4.1.3 A p-adic version of Roth's theorem

The following non-Archimedean extension of Roth's theorem was proved in 1957 by Ridout [41]. For every prime number ℓ, the

ℓ-adic absolute value is denoted by $|\cdot|_\ell$ and normalized such that $|\ell|_\ell = \ell^{-1}$. Thus given an integer n, $|n|_\ell = \ell^{-j}$ where j denotes the largest integer for which ℓ^j divides n.

Theorem 4.7. *Let ξ be an algebraic number and ε be a positive real number. Let S be a finite set of distinct prime numbers. Then there are only a finite number of rational numbers p/q such that $q \geqslant 1$ and*

$$\left(\prod_{\ell \in S} |p|_\ell \cdot |q|_\ell\right) \cdot \left|\xi - \frac{p}{q}\right| < \frac{1}{q^{2+\varepsilon}} \cdot$$

We point out a first classical consequence of Ridout's theorem.

Corollary 4.8. *The real number*

$$\mathcal{K} := \sum_{n=1}^{+\infty} \frac{1}{10^{2^n}}$$

is transcendental.

Proof. Let j be a positive integer and let us consider the rational number $\rho_j := \sum_{n=1}^{j} 10^{-2^n}$. There exists an integer p_j such that $\rho_j = p_j/q_j$ with $q_j := 10^{2^j}$. Observe that

$$|\mathcal{K} - \rho_j| = \sum_{n>j} \frac{1}{10^{2^n}} < \frac{2}{10^{2^{j+1}}} = \frac{2}{(q_j)^2} ,$$

and set $S := \{2, 5\}$. An easy computation gives that

$$\left(\prod_{\ell \in S} |q_j|_\ell \cdot |p_j|_\ell\right) \leqslant \prod_{\ell \in S} |q_j|_\ell = \frac{1}{q_j}$$

and thus

$$\left(\prod_{\ell \in S} |q_j|_\ell \cdot |p_j|_\ell\right) \cdot |\mathcal{K} - p_j/q_j| < \frac{2}{(q_j)^3} \cdot$$

Theorem 4.7 then implies that $\mathcal{K}$ is transcendental. □

Let us now consider the 3-automatic word y defined in Example 2.3. Let us associate with y the real number

$$\xi_y := \sum_{n \geqslant 1} \frac{y_n}{10^n} = 0.001\,010\,010\,001\,010\,001\,001 \cdots .$$

Unfortunately, the word y does not have sufficiently large initial repetitive patterns to prove the transcendence of ξ_y by means of Roth's theorem as we did in Theorem 4.6. To overcome this difficulty we use a trick based on Ridout's theorem that was first introduced by Ferenczi and Mauduit [15].

Theorem 4.9. *The 3-automatic real number ξ_y is transcendental.*

Proof. For every integer $j \geqslant 0$, set $u_j := \phi(\omega^j(012020))$, $v_j := \phi(\omega^j(021012))$, $r_j := |u_j|$ and $s_j := |v_j|$. Let us also consider the rational number ρ_j defined by

$$\rho_j := 0.u_j v_j^{\omega} .$$

An easy computation shows that there exists an integer p_j such that

$$\rho_j = \frac{p_j}{10^{r_j}(10^{s_j} - 1)} \cdot \tag{4.5}$$

On the other hand, one can check that y begins with the word

$$\phi(\omega^j(0120200210120210120)) = u_j v_j v_j \phi(\omega^j(0)) .$$

Since $\phi(\omega^j(0))$ is a prefix of v_j, we obtain that the first $r_j + 2s_j + |\phi(\omega^j(0))| = 19 \cdot 3^j$ digits in the decimal expansion of ξ_y and of ρ_j are the same. We thus have

$$|\xi_y - \rho_j| < \frac{1}{10^{19\cdot 3^j}} \cdot \tag{4.6}$$

Note that we obtain very special rational approximations to ξ_y: their denominators can be divided by a very large power of 10. More precisely, letting $S := \{2, 5\}$, we have

$$\prod_{\ell \in S} |10^{r_j}(10^{s_j} - 1)|_\ell = \frac{1}{10^{r_j}} = \frac{1}{10^{6\cdot 3^j}} \cdot \tag{4.7}$$

Set $q_j := 10^{r_j}(10^{s_j} - 1)$. We infer from (4.5), (4.6) and (4.7) that

$$\left(\prod_{\ell \in S} |p_j|_\ell \cdot |q_j|_\ell\right) \cdot \left|\xi_y - \frac{p_j}{q_j}\right| < \frac{1}{10^{25\cdot 3^j}} , \tag{4.8}$$

for every positive integer j. Since $q_j < 10^{r_j + s_j} = 10^{12\cdot 3^j}$, we deduce from (4.8) that

$$\left(\prod_{\ell \in S} |p_j|_\ell \cdot |q_j|_\ell\right) \cdot \left|\xi_y - \frac{p_j}{q_j}\right| < \frac{1}{q_j^{2+1/12}} ,$$

for every integer j large enough. On the other hand, it can be shown that the word y is not eventually periodic, which implies that the set of rational numbers $\{p_j/q_j \mid j \geqslant 1\}$ is infinite. It thus follows from Theorem 4.7 that ξ_y is transcendental, concluding the proof. □

4.2 The Schmidt subspace theorem and a proof of Cobham's conjecture

A wonderful multidimensional generalization of Roth's theorem was obtained by Schmidt in the early 70's (see [45]). It is now referred to as the Schmidt subspace theorem or, for short, as the subspace theorem. We state below a heavily simplified p-adic version of this theorem. However, Theorem 4.10 turns out to be strong enough for our purpose.

Theorem 4.10. *Let $m \geqslant 2$ be an integer and ε be a positive real number. Let S be a finite set of distinct prime numbers. Let $L_1, \ldots, L_m$ be m linearly independent (over the* *field of algebraic numbers) linear forms with real algebraic coefficients. Then the set of solutions* $\mathbf{x} = (x_1, \dots, x_m)$ *in* $\mathbb{Z}^m$ *to the inequality*

$$\left(\prod_{i=1}^{m}\prod_{\ell \in S} |x_i|_\ell\right) \cdot \prod_{i=1}^{m} |L_i(\mathbf{x})| \leqslant (\max\{|x_1|, \dots, |x_m|\})^{-\varepsilon}$$

lies in finitely many proper vector subspaces of $\mathbb{Q}^m$*.*

Let us note that Roth's theorem easily follows from Theorem 4.10. Let $0 < \xi < 1$ be a real algebraic number and let ε be a positive real number. Consider the two independent linear forms $\xi X - Y$ and X. Choosing $S = \{\emptyset\}$, Theorem 4.10 implies that all the integer solutions (p, q) to

$$|q| \cdot |q\xi - p| < |q|^{-\varepsilon} \tag{4.9}$$

are contained in a finite union of proper vector subspaces of $\mathbb{Q}^2$. There thus is a finite set of equations $x_1X + y_1Y = 0, \dots, x_tX + y_tY = 0$ such that, for every solution (p, q) to (4.9), there exists an integer k with $x_kp + y_kq = 0$. This means that there are only finitely many rational solutions to $|\xi - p/q| < |q|^{-2-\varepsilon}$, which immediately gives Roth's theorem.

Proof of Theorem 4.1. Let $0 < \xi < 1$ be an automatic irrational real number. Then there is an integer base $b \geqslant 2$ such that the base-b expansion of ξ is a k-automatic word for some integer $k \geqslant 2$. Let us denote by a the base-b expansion of ξ.

By Theorem 2.1, there exist a coding φ from an alphabet $A = \{1, 2, \dots, r\}$ to the alphabet $\{0, 1, \dots, b-1\}$ and a k-uniform morphism σ from A into itself such that

$$a = \varphi(u)\,,$$

where u is a fixed point of σ. By the pigeonhole principle, the prefix of length $r+1$ of u can be written under the form $w_1cw_2cw_3$, where c is a letter and w_1, w_2, w_3 are (possibly empty) finite words.

For every integer $j \geqslant 1$, set $u_j = \varphi(\sigma^j(w_1))$, $v_j = \varphi(\sigma^j(cw_2))$ and $v'_j = \varphi(\sigma^j(c))$. Since σ is a k-uniform morphism an φ is a coding, we get that

$$|u_j| = s \cdot k^j, \quad |v_j| = t \cdot k^j \quad \text{and} \quad |v'_j| = k^j\,,$$

where $s := |u_1|$ and $t := |v_1|$. Thus the base-b expansion of ξ begins with the word $u_jv_jv'_j$, that is,

$$\xi = 0.u_jv_jv'_j \cdots\,.$$

Let ρ_j be the rational number whose base-b expansion is the infinite word $u_jv_j^\omega$, that is,

$$\rho_j = 0.u_jv_j^\omega\,.$$

A simple computation shows that there exists an integer p_j such that

$$\rho_j = \frac{p_j}{b^{s\cdot k^j}(b^{t\cdot k^j} - 1)}\,\cdot$$

Since ρ_j and ξ have the same first $(s+t+1)\cdot k^j$ digits, we have

$$|\xi - \rho_j| < \frac{1}{b^{(s+t+1)\cdot k^j}}\,\cdot$$

Henceforth, we assume that ξ is an algebraic number, and we will reach a contradiction. Consider the three linearly independent linear forms with real algebraic coefficients:

$$\begin{aligned} L_1(X_1,X_2,X_3) = {} & \xi X_1 - \xi X_2 - X_3\,, \\ L_2(X_1,X_2,X_3) = {} & X_1\,, \\ L_3(X_1,X_2,X_3) = {} & X_2\,. \end{aligned}$$

For $j \geqslant 1$, evaluating them on the integer triple

$$\mathbf{x}_j := (x_1^{(j)}, x_2^{(j)}, x_3^{(j)}) := (b^{(s+t)\cdot k^j}, b^{s\cdot k^j}, p_j)\,,$$

we obtain that

$$\prod_{i=1}^{3} |L_i(\mathbf{x}_j)| \leqslant b^{(2s+t-1)\cdot k^j}\,. \tag{4.10}$$

On the other hand, letting S be the set of prime divisors of b, we get that

$$\prod_{i=1}^{3}\prod_{\ell\in S} |x_i^{(j)}|_\ell \leqslant \prod_{\ell\in S} |b^{(s+t)\cdot k^j}|_\ell \cdot \prod_{\ell\in S} |b^{s\cdot k^j}|_\ell = b^{-(2s+t)\cdot k^j}\,. \tag{4.11}$$

Combining (4.10) and (4.11), we get that

$$\left(\prod_{i=1}^{3}\prod_{\ell\in S} |x_i^{(j)}|_\ell\right) \cdot \prod_{i=1}^{3} |L_i(\mathbf{x}_j)| \leqslant b^{-k^j}\,.$$

Set $\varepsilon = 1/(s+t)$. We thus obtain

$$\left(\prod_{i=1}^{3}\prod_{\ell\in S} |x_i^{(j)}|_\ell\right) \cdot \prod_{i=1}^{3} |L_i(\mathbf{x}_j)| \leqslant \left(\max\{b^{(s+t)\cdot k^j}, b^{s\cdot k^j}, p_j\}\right)^{-\varepsilon}\,,$$

for every positive integer j.

We then infer from Theorem 4.10 that all integer points $\mathbf{x}_j$ lie in a finite number of proper vector subspaces of $\mathbb{Q}^3$. Thus there exist a nonzero integer triple (z_1, z_2, z_3) and an infinite set of distinct positive integers $\mathcal{J}$ such that

$$z_1 b^{(s+t)\cdot k^j} + z_2 b^{s\cdot k^j} + z_3 p_j = 0\,, \tag{4.12}$$

for every j in $\mathcal{J}$. Recall that $p_j/b^{(s+t)\cdot k^j}$ tends to ξ when j tends to infinity. Dividing (4.12) by $b^{(s+t)\cdot k^j}$ and letting j tend to infinity along $\mathcal{J}$, we get that ξ is a rational number since (z_1, z_2, z_3) is a nonzero triple. This provides a contradiction. □

5 The Skolem-Mahler-Lech theorem in positive characteristic

5.1 Zeros of linear recurrences over fields of characteristic zero

The Skolem-Mahler-Lech theorem is a celebrated result which describes the set of solutions in n to the equation $a(n) = 0$, where $a(n)$ is a sequence satisfying a linear recurrence over a field of characteristic 0. We recall that if $\mathbb{K}$ is a field and $a(n)$ is a $\mathbb{K}$-valued sequence, then $a(n)$ *satisfies a linear recurrence* over $\mathbb{K}$ if there exists a natural number d and values $c_1, \ldots, c_d \in \mathbb{K}$ such that

$$a(n) = \sum_{i=1}^{d} c_i a(n-i)$$

for all sufficiently large values of n. The zero set of the linear recurrence a is defined by

$$\mathcal{Z}(a) := \{n \in \mathbb{N} \mid a(n) = 0\} \ .$$

Theorem 5.1 (Skolem-Mahler-Lech). *Let a be a linear recurrence over a field of characteristic 0. Then $\mathcal{Z}(a)$ is a union of a finite set and a finite number of arithmetic progressions.*

This theorem was first proved for linear recurrences over the rational numbers by Skolem [49]. It was next proved for linear recurrences over the algebraic numbers by Mahler [32]. The version above was proven first by Lech [27] and later by Mahler [33], [34]. This history of this theorem can be found in the book by Everest et al. [14]. The techniques used by Lech to prove the Skolem-Mahler-Lech theorem are a modification of a method first used by Skolem [49]. The idea of the proof is to first note that it is no loss of generality to assume that $\mathbb{K}$ is a finitely generated extension of $\mathbb{Q}$. We can then embed $\mathbb{K}$ in a p-adic field $\mathbb{Q}_p$ for some prime p. One can then show that there exists a natural number a such that for each $i = 0, \ldots, a-1$, there is a p-adic analytic map θ_i on $\mathbb{Z}_p$ such that $\theta_i(n) = f(an+i)$ for all sufficiently large positive integers $n \in \mathbb{N}$. If $f(an+i)$ is zero for infinitely many natural numbers n, then the map θ_i has infinitely many zeros in $\mathbb{Z}_p$. Since an analytic function cannot have infinitely many zeros in a compact subset of its domain of convergence unless that function is identically zero, this implies that either $f(an+i) = 0$ for all n sufficiently large, or there are only finitely many n for which $f(an+i) = 0$, which gives the result.

There are many different proofs and extensions of the Skolem-Mahler-Lech theorem in the literature [5, 21, 54, 14]. These proofs all use p-adic methods in some way, although the result is valid in any field of characteristic 0. A well-known aspect of Theorem SML is that it is an ineffective result. Indeed, it is still an open problem whether the set $\mathcal{Z}(a)$ can always be determined for a given linear recurrence $a(n)$ defined over a field of characteristic 0 (see the discussions in [14] and [52]). In particular, it is still unknown whether the fact that $\mathcal{Z}(a)$ is empty or not is a decidable question.

5.2 Zeros of linear recurrences over fields of positive characteristic

5.2.1 Pathological examples over fields of positive characteristic It is interesting to note that the Skolem-Mahler-Lech theorem does not hold for fields $\mathbb{K}$ of positive characteristic. The simplest counter-example was given by Lech [27]. Let p be a prime and let $\mathbb{K} = \mathbb{F}_p(t)$ be the field of rational functions in one variable over $\mathbb{F}_p$. Let

$$a(n) := (1+t)^n - t^n - 1 .$$

It is easy to check that $a(n)$ satisfies the recurrence

$$a(n) - (2+2t)a(n-1) + (1+3t+t^2)a(n-2) - (t+t^2)a(n-3) \;=\; 0$$

for $n > 3$. On the other hand, we have

$$a(p^j) = (1+t)^{p^j} - t^{p^j} - 1 = 0$$

and $a(n) \neq 0$ if n is not a power of p, and so we have

$$\mathcal{Z}(a) = \{1, p, p^2, p^3, \ldots\} .$$

In fact, there are even more pathological examples, which show that the correct analogue of the Skolem-Mahler-Lech theorem in positive characteristic is much more subtle. For example, consider the sequence $a(n)$ in $\mathbb{F}_2(x, y, z)$ defined by

$$a(n) := (x+y+z)^n - (x+y)^n - (x+z)^n - (y+z)^n + x^n + y^n + z^n .$$

We note that if V denotes the $\mathbb{K}$-vector space consisting of all $\mathbb{K}$-valued sequences and $S : V \to V$ is the "shift" linear operator that sends a sequence $a(1), a(2), \ldots$ to the sequence $0, a(1), a(2), \ldots$, then $a(n)$ satisfies a linear recurrence if and only if there is a nonzero polynomial $P(t)$ with coefficients in $\mathbb{K}$ such that when $P(S)$ is applied to the sequence $a(n)$ we obtain a sequence whose terms are eventually zero. Then one can see that the operator

$$(1-(x+y+z)S)(1-(x+y)S)(1-(y+z)S)(1-xS)(1-yS)(1-zS)$$

sends the sequence $a(n)$ to a sequence whose terms are eventually zero.

We claim that the zero set of $a(n)$ is precisely all natural numbers n of the form $2^i + 2^j$ or of the form 2^i. To see this, observe that $a(2^i) = 0$ follows simply from the fact that $(b+c)^{2^i} = b^{2^i} + c^{2^i}$ for elements b and c in a field of characteristic 2. To check that $a(2^i + 2^j) = 0$ we note that

$$\begin{aligned} G(x_1, y_1, z_1; x_2, y_2, z_2) \;&:=\; (x_1+y_1+z_1)(x_2+y_2+z_2) \\ &-\; (x_1+y_1)(x_2+y_2) - (x_1+z_1)(x_2+z_2) \\ &-\; (y_1+z_1)(y_2+z_2) + x_1x_2 + y_1y_2 + z_1z_2 \end{aligned}$$

is identically zero in every field. Notice that if $c_1, c_2, c_3 \in \mathbb{F}_2$ then

$$(c_1x + c_2y + c_3z)^{2^i+2^j} = (c_1x^{2^i} + c_2y^{2^i} + c_3z^{2^i})(c_1x^{2^j} + c_2y^{2^j} + c_3z^{2^j}) .$$

Hence

$$a(2^i + 2^j) = G(x^{2^i}, y^{2^i}, z^{2^i}; x^{2^j}, y^{2^j}, z^{2^j}) = 0 .$$

On the other hand, if n is not a power of 2 or of the form $2^i + 2^j$, then we can write $n = 2^i + 2^j + 2^k m$ where $i > j$, $2^j > 2^k m$ and m is an odd positive integer. Note that

$$\begin{aligned} (x+y+z)^n &= (x+y+z)^{2^i}(x+y+z)^{2^j}\left((x+y+z)^{2^k}\right)^m \\ &= (x^{2^i}+y^{2^i}+z^{2^i})(x^{2^j}+y^{2^j}+z^{2^j})(x^{2^k}+y^{2^k}+z^{2^k})^m . \end{aligned}$$

Consider the coefficient of $x^{2^i}y^{2^j}z^{2^k m}$ in $(x+y+z)^n$. The only way to get this term is to take x^{2^i} from the first term in the product, y^{2^j} from the second term, and $z^{2^k m}$ from the third term. Hence the coefficient is 1. Since $(x+y+z)^n$ is the only term in $a(n)$ that has monomials of the form $x^b y^c z^d$ with $b, c, d > 0$ appearing, we see that $a(n)$ is nonzero if the binary expansion of n has more than two 1's.

5.2.2 Derksen's theorem We now give a remarkable result due to Derksen [12]. We have seen that the zero set of a linear recurrence in a field of characteristic $p > 0$ is often more pathological than in characteristic zero. At the same time, in our pathological examples, the base-p expansion of a number n gives insight into whether the nth term of our linearly recurrent sequence vanishes. In fact, Derksen [12] shows that the zero set of a linearly recurrent sequence can always be described in terms of automata.

Theorem 5.2 (Derksen). *Let a be a linear recurrence over a field of characteristic p. Then the set $\mathcal{Z}(a)$ is a p-automatic set.*

Derksen gave a further refinement of this result, however the main ingredient of his proof is the fact that the zero set is p-automatic. Furthermore, each step in Derksen's proof can be made effective!

We prove an extension of Derksen's result for algebraic power series in several variables in the next section. To explain the connection between Derksen's result and power series, we recall the following classical result.

Proposition 5.3. *Let $\mathbb{K}$ be a field and let $a(n)$ be a $\mathbb{K}$-valued sequence. The following conditions are equivalent.*

(i) *The sequence $a(n)$ satisfies a linear recurrence over $\mathbb{K}$.*

(ii) *There is a natural number d, a matrix $A \in M_d(\mathbb{K})$, and vectors v and w in $\mathbb{K}^d$ such that $a(n) = w^T A^n v$.*

(iii) *$\sum_{n\geqslant 0} a(n)t^n$ is the power series expansion of a rational function in $\mathbb{K}(t)$.*

Proof. $(i) \implies (ii)$. Suppose that $a(n)$ satisfies a linear recurrence

$$a(n) := \sum_{j=1}^{d} c_j a(n-j)$$

for all $n \geqslant d$. We let

$$v(i) := [a(i)\ a(i+1)\ \cdots\ a(i+d-1)]^T$$

and

$$w := [1\ 0\ 0\ \cdots\ 0]^T .$$

Finally, we let

$$A := \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ c_d & c_{d-1} & c_{d-2} & c_{d-3} & \cdots & c_1 \end{pmatrix}.$$

Then one easily sees that $v(i+1) = Av(i)$ and so $w^T A^n v = a(n)$, where $v = v(0)$. Thus (i) implies (ii).

$(ii) \implies (iii)$. Set

$$f(t) := \sum_{n=0}^{\infty} (w^T A^n v) t^n .$$

By the Cayley-Hamilton theorem, A satisfies a polynomial $A^d + \sum_{j=0}^{d-1} c_j A^j \; = \; 0$ and hence

$$w^T A^{n+d} v + \sum_{j=0}^{d-1} c_j w^T A^{j+n} v \; = \; 0$$

for all n. It follows that $f(t)(1 + \sum_{j=0}^{d-1} c_j t^{d-j})$ is a polynomial in t and so $f(t)$ is the power series expansion of a rational function.

$(iii) \implies (i)$. Suppose that $f(t) = \sum_{n=0}^{\infty} a(n) t^n$ is the power series expansion of a rational function $P(t)/Q(t)$ with $P(t)$ and $Q(t)$ polynomials and $Q(t)$ nonzero. We may assume that $Q(0) = 1$. We write $Q(t) = 1 + \sum_{j=1}^{d} c_j t^j$. Then $P(t) = f(t)Q(t)$ and so $a(n) + \sum_{j=1}^{d} c_j a(n-j) = 0$ for all n larger than the degree of $P(t)$. It follows that $a(n)$ satisfies a linear recurrence. □

5.3 Vanishing coefficients of algebraic power series

In light of Proposition 5.3, we may interpret Derksen's result as a statement about the zero coefficients of the power series expansion of a rational power series over a field of characteristic $p > 0$. In this section, we show that this interpretation gives rise to a far-reaching generalization of Derksen's result.

We first note that rational power series are a subset of *algebraic power series* (choosing $m = 1$ in the definition below).

Definition 5.1. Let $\mathbb{K}$ be a field. We say that a power series

$$f(t) = \sum_{n=0}^{\infty} a(n) t^n \in \mathbb{K}[[t]]$$

is *algebraic* if it is algebraic over the field of rational functions $\mathbb{K}(t)$, that is, if there exists a natural number m and polynomials $A_0(t), \ldots, A_m(t) \in \mathbb{K}[t]$, with $A_m(t)$ nonzero, such that

$$\sum_{j=0}^{m} A_j(t) f(t)^j \; = \; 0 .$$

More generally, we say that $f(t_1,\ldots,t_d) \in \mathbb{K}[[t_1,\ldots,t_d]]$ is *algebraic* if there exist polynomials $A_0,\ldots,A_m \in \mathbb{K}[t_1,\ldots,t_d]$, not all zero, such that

$$\sum_{j=0}^{m} A_j(t_1,\ldots,t_d) f(t_1,\ldots,t_d)^j \ = \ 0\,.$$

Given a multivariate power series $f(t_1,\ldots,t_d) = \sum_{n_1,\ldots,n_d} a_{n_1,\ldots,n_d} t_1^{n_1}\cdots t_d^{n_d} \in \mathbb{K}[[t_1,\ldots,t_d]]$, we denote by $\mathcal{Z}(f)$ the set of vanishing coefficients, that is,

$$\mathcal{Z}(f) = \{(n_1,\ldots,n_d) \in \mathbb{N}^d \mid a_{n_1,\ldots,n_d} = 0\}\,.$$

It is interesting to note that the Skolem-Mahler-Lech theorem in characteristic 0 has no analogue for multivariate rational functions. For instance,

$$f(t_1,t_2) = \sum_{m,n} (2^m - n^2) t_1^m t_2^n$$

is a bivariate rational power series in $\mathbb{Q}[[t_1,t_2]]$ with

$$\mathcal{Z}(f) = \{(m,n) \ : \ m \equiv 0 \ (\mathrm{mod}\ 2), n = 2^{m/2}\}\,.$$

Thus we cannot expect the zero set to be given in terms of arithmetic progressions or even in terms of finite automata.

To see some of the complexities that can occur in the multivariate case, consider the power series

$$f(t_1,t_2) = \sum_{m,n\geqslant 0} (3^m - 2^n - 1) t_1^m t_2^n\,.$$

We see that

$$f(t_1,t_2) = (1-3t_1)^{-1}(1-t_2)^{-1} - (1-t_1)^{-1}(1-2t_2)^{-1} - (1-t_1)^{-1}(1-t_2)^{-1}$$

and so it is a rational power series. On the other hand the coefficient of $t_1^m t_2^n$ is zero if and only if $3^m = 2^n + 1$. It is now known that this occurs only when (m,n) is $(2,3)$ or $(1,1)$, due to Mihăilescu's solution to Catalan's conjecture [36]. In general finding the zero set often involves difficult diophantine problems.

Remarkably, in positive characteristic an analogue of Derksen's result holds for multivariate rational power series as shown in [1]—in fact it even holds for multivariate algebraic power series! In the sequel of this chapter, we will use $\mathbf{n}$ and $\mathbf{j}$ to represent respectively the d-tuple of natural numbers $(n_1,\ldots,n_d)$ and $(j_1,\ldots,j_d)$. We will also let $\mathbf{t}^{\mathbf{n}}$ denote the monomial $t_1^{n_1}\cdots t_d^{n_d}$.

Theorem 5.4 (Adamczewski and Bell). *Let $\mathbb{K}$ be a field of characteristic $p > 0$ and let $f(\mathbf{t}) \in \mathbb{K}[[\mathbf{t}]]$ be the power series expansion of an algebraic function over $\mathbb{K}(\mathbf{t})$. Then $\mathcal{Z}(f)$ is a p-automatic subset of $\mathbb{N}^d$.*

We note that this immediately implies Theorem 5.2 by taking $d = 1$ and taking $f(t)$ to be a rational function. On the other hand, by taking $\mathbb{K}$ to be a finite field, Theorem 5.4 reduces to the difficult part of the multivariate version of Christol's theorem (see Theorem 6.2). As with Derksen's proof, it seems that Theorem 5.4 can be made effective.

Furthermore, given any p-automatic set $\mathcal{N}$ in $\mathbb{N}^d$, $\mathcal{N}$ is the zero set of the power series $\sum_{\mathbf{n}\notin\mathcal{N}} \mathbf{t}^{\mathbf{n}} \in \mathbb{F}_p((\mathbf{t}))$ which is known to be algebraic over $\mathbb{F}_p(\mathbf{t})$ by Theorem 6.2. At this level of generality, we thus have a nice correspondence between p-automatic sets and the zero set of algebraic multivariate functions over fields of characteristic p.

5.3.1 Proof of Theorem 5.4 In order to prove this result we need to introduce some notation.

Let p be a prime number and let d be a natural number. For each $\mathbf{j} = (j_1, \ldots, j_d) \in \{0, 1, \ldots, p-1\}^d$, we define $e_{\mathbf{j}} : \mathbb{N}^d \to \mathbb{N}^d$ by

$$e_{\mathbf{j}}(n_1, \ldots, n_d) := (pn_1 + j_1, \ldots, pn_d + j_d)\,. \tag{5.1}$$

We let Σ denote the semigroup generated by the collection of all $e_{\mathbf{j}}$ under composition.

Remark 5.5. Note that if Δ is a finite set, then $f : \mathbb{N}^d \to \Delta$ is p-automatic if and only the set of functions $\{f \circ e \ :\ e \in \Sigma\}$ is a finite set.

We also recall that a field $\mathbb{K}$ of characteristic $p > 0$ is *perfect* if the map $x \mapsto x^p$ is surjective on $\mathbb{K}$. Let p be a prime number and let $\mathbb{K}$ be a perfect field of characteristic p. For a power series $f(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} a(\mathbf{n})\mathbf{t}^{\mathbf{n}} \in \mathbb{K}[[\mathbf{t}]]$, we define

$$E_{\mathbf{j}}(f(\mathbf{t})) \ :=\ \sum_{\mathbf{n}\in\mathbb{N}^d} (a \circ e_{\mathbf{j}}(\mathbf{n}))^{1/p}\mathbf{t}^{\mathbf{n}} \tag{5.2}$$

for $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$. We let Ω denote the semigroup generated by the collection of $E_{\mathbf{j}}$ under composition. We let $\Omega(f)$ denote the $\mathbb{K}$-vector space spanned by all power series of the form $E \circ f$ with $E \in \Omega$. We note that if $g \in \Omega(F)$ then $E \circ g \in \Omega(f)$ for all $E \in \Omega$.

A theorem of Sharif and Woodcock [48] gives a concrete characterization of the algebraic power series over a perfect field of positive characteristic.

Theorem 5.6 (Sharif and Woodcock). *Let p be a prime number and let $\mathbb{K}$ be a perfect field of characteristic p. A power series $f(\mathbf{t}) \in \mathbb{K}[[\mathbf{t}]]$ is algebraic if and only if $\Omega(f)$ is a finite-dimensional $\mathbb{K}$-vector space.*

One can rephrase the theorem of Sharif and Woodcock in terms of the coefficients of an algebraic power series.

Lemma 5.7. *Let p be a prime number, let $\mathbb{K}$ be a perfect field of characteristic p, and let $a : \mathbb{N}^d \to \mathbb{K}$ be a sequence with the property that*

$$f(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} a(\mathbf{n})\mathbf{t}^{\mathbf{n}} \in \mathbb{K}[[\mathbf{t}]]$$

is a nonzero algebraic function over $\mathbb{K}(\mathbf{t})$. Then there exists a natural number m and there exist maps $a_1, \ldots, a_m : \mathbb{N}^d \to \mathbb{K}$ with the following properties.

(i) *The formal power series $f_i(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} a_i(\mathbf{n})\mathbf{t}^{\mathbf{n}}$, $1 \leqslant i \leqslant m$, form a basis of $\Omega(f)$ as a $\mathbb{K}$-vector space.*

(ii) $f_1 = f$.
(iii) *If* $b : \mathbb{N}^d \to \mathbb{K}$ *has the property that* $g(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} b(\mathbf{n})\mathbf{t}^{\mathbf{n}}$ *belongs to* $\Omega(f)$*, then* $b \circ e_{\mathbf{j}} \in \mathbb{K}\, a_1^p + \cdots + \mathbb{K}\, a_m^p$ *for every* $\mathbf{j} \in \{0, \ldots, p-1\}^d$.

Proof. Since $f(\mathbf{t})$ is algebraic, $\dim_{\mathbb{K}}(\Omega(f))$ is finite by Theorem 5.6. We can thus pick maps $a_1, \ldots, a_m : \mathbb{N}^d \to \mathbb{K}$ such that the m power series $f_i(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} a_i(\mathbf{n})\mathbf{t}^{\mathbf{n}}$ form a basis of $\Omega(f)$, and with $f_1 = f$. Let $b : \mathbb{N}^d \to \mathbb{K}$ be such that $g(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} b(\mathbf{n})\mathbf{t}^{\mathbf{n}}$ belongs to $\Omega(f)$. Observe that the power series g can be decomposed as

$$g(\mathbf{t}) = \sum_{\mathbf{j}\in\{0,\ldots,p-1\}^d} \mathbf{t}^{\mathbf{j}} E_{\mathbf{j}}(g(\mathbf{t}))^p\,. \tag{5.3}$$

By assumption, $E_{\mathbf{j}}(g(\mathbf{t})) \in \mathbb{K}\, f_1(\mathbf{t}) + \cdots + \mathbb{K}\, f_m(\mathbf{t})$ and hence $E_{\mathbf{j}}(g(\mathbf{t}))^p \in \mathbb{K}\, f_1(\mathbf{t})^p + \cdots + \mathbb{K}\, f_m(\mathbf{t})^p$. Let $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$. Considering the coefficient of $\mathbf{t}^{p\mathbf{n}+\mathbf{j}}$ in Equation (5.3), we see that $b \circ e_{\mathbf{j}}(\mathbf{n})$ is equal to the coefficient of $\mathbf{t}^{p\mathbf{n}}$ in $E_{\mathbf{j}}(g(\mathbf{t}))^p$, which belongs to $\mathbb{K}\, a_1(\mathbf{n})^p + \cdots + \mathbb{K}\, a_m(\mathbf{n})^p$. □

Before proving Theorem 5.4, we first fix a few notions. Given a finitely generated field extension $\mathbb{K}_0$ of $\mathbb{F}_p$, we let $\mathbb{K}_0^{\langle p\rangle}$ denote the subfield consisting of all elements of the form x^p with $x \in \mathbb{K}_0$. Given $\mathbb{F}_p$-vector subspaces V and W of $\mathbb{K}_0$ we let VW denote the $\mathbb{F}_p$-subspace of $\mathbb{K}_0$ spanned by all products of the form vw with $v \in V, w \in W$. We let $V^{\langle p\rangle}$ denote the $\mathbb{F}_p$-vector subspace consisting of all elements of the form v^p with $v \in V$. We note that since $\mathbb{K}_0$ is a finitely generated field extension of $\mathbb{F}_p$, $\mathbb{K}_0$ is a finite-dimensional $\mathbb{K}_0^{\langle p\rangle}$-vector space. If we fix a basis

$$\mathbb{K}_0 = \bigoplus_{i=1}^{r} \mathbb{K}_0^{\langle p\rangle} h_i$$

then we have *projections* $\pi_1, \ldots, \pi_r : \mathbb{K}_0 \to \mathbb{K}_0$ defined by

$$x = \sum_{i=1}^{r} \pi_i(x)^p h_i\,. \tag{5.4}$$

Remark 5.8. For $1 \leqslant i \leqslant r$ and $a, b, c \in \mathbb{K}_0$ we have

$$\pi_i(c^p a + b) = c\pi_i(a) + \pi_i(b)\,.$$

The last ingredient of the proof is a technical (but very useful) result due to Derksen, which we state here without proof.

Proposition 5.9 (Derksen). *Let* $\mathbb{K}_0$ *be a finitely generated field extension of* $\mathbb{F}_p$ *and let* $\pi_1, \ldots, \pi_r : \mathbb{K}_0 \to \mathbb{K}_0$ *be as in Equation (5.4). Let* V *be a finite-dimensional* $\mathbb{F}_p$*-vector subspace of* $\mathbb{K}_0$*. Then there exists a finite-dimensional* $\mathbb{F}_p$*-vector subspace* W *of* $\mathbb{K}_0$ *containing* V *such that* $\pi_i(WV) \subseteq W$ *for* $1 \leqslant i \leqslant r$*.*

Proof of Theorem 5.4. By enlarging $\mathbb{K}$ if necessary, we may assume that $\mathbb{K}$ is perfect. By Lemma 5.7 we can find maps $a_1, \ldots, a_m : \mathbb{N}^d \to \mathbb{K}$ with the following properties.

(1) the power series $f_i(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} a_i(\mathbf{n})\mathbf{t}^{\mathbf{n}}$, $1 \leqslant i \leqslant m$, form a basis for $\Omega(f)$.

(2) $f_1 = f$.

(3) If $b : \mathbb{N}^d \to \mathbb{K}$ has the property that $g(\mathbf{t}) := \sum_{\mathbf{n}\in\mathbb{N}^d} b(\mathbf{n})\mathbf{t}^{\mathbf{n}}$ belongs to $\Omega(f)$, then $b \circ e_{\mathbf{j}} \in \mathbb{K}\, a_1^p + \cdots + \mathbb{K}\, a_m^p$ for every $\mathbf{j} \in \{0, \ldots, p-1\}^d$.

In particular, given $1 \leqslant i \leqslant m$ and $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$, there are elements $\lambda(i, \mathbf{j}, k)$, $1 \leqslant k \leqslant m$, such that

$$a_i \circ e_{\mathbf{j}} = \sum_{k=1}^{m} \lambda(i, \mathbf{j}, k) a_k^p \,. \tag{5.5}$$

Since $f_1, \ldots, f_m$ are algebraic power series, there exists a finitely generated field extension of $\mathbb{F}_p$ such that all coefficients of $f_1, \ldots, f_m$ are contained in this field extension. It follows that the subfield $\mathbb{K}_0$ of $\mathbb{K}$ generated by the coefficients of $f_1(\mathbf{t}), \ldots, f_m(\mathbf{t})$ and all the elements $\lambda(i, \mathbf{j}, k)$ is a finitely generated field extension of $\mathbb{F}_p$.

Since $\mathbb{K}_0$ is a finite-dimensional $\mathbb{K}_0^{\langle p\rangle}$-vector space, we can fix a basis $\{h_1, \ldots, h_r\}$ of $\mathbb{K}_0$, that is,

$$\mathbb{K}_0 = \bigoplus_{i=1}^{r} \mathbb{K}_0^{\langle p\rangle} h_i \,.$$

Then we have *projections* $\pi_1, \ldots, \pi_r : \mathbb{K}_0 \to \mathbb{K}_0$ defined by

$$c = \sum_{i=1}^{r} \pi_i(c)^p h_i \,. \tag{5.6}$$

We let V denote the finite-dimensional $\mathbb{F}_p$-vector subspace of $\mathbb{K}_0$ spanned by the elements $\lambda(i, \mathbf{j}, k)$, $1 \leqslant i, k \leqslant m$ and $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$, and by 1. By Equation (5.5), we have

$$a_i \circ e_{\mathbf{j}} \in \sum_{k=1}^{m} V a_k^p \,, \tag{5.7}$$

for $1 \leqslant i \leqslant m$ and $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$. By Proposition 5.9 there exists a finite-dimensional $\mathbb{F}_p$-vector subspace W of $\mathbb{K}_0$ containing V such that $\pi_i(WV) \subseteq W$ for $1 \leqslant i \leqslant r$. Set

$$U := Wa_1 + \cdots + Wa_m \subseteq \{b \ : b : \mathbb{N}^d \to \mathbb{K}_0\} \,.$$

We note that $\operatorname{Card} U \leqslant (\operatorname{Card} W)^m < \infty$. Note also that if $\ell \in \{1, \ldots, r\}$, $i \in \{1, \ldots, m\}$, and $j \in \{0, 1, \ldots, p-1\}^d$ then by Equation (5.7) and Remark 5.8 we have

$$\begin{aligned}\pi_\ell(Wa_i \circ e_{\mathbf{j}}) \subseteq \pi_\ell(WVa_1^p + \cdots + WVa_m^p) \subseteq \sum_{k=1}^{m} \pi_\ell(WV) a_k \\ \subseteq \sum_{k=1}^{m} W a_k = U \,.\end{aligned}$$

Thus by Remark 5.8, if $b \in U$ and $\mathbf{j} \in \{0, 1, \ldots, p-1\}^d$, then $b_\ell := \pi_\ell(b \circ e_{\mathbf{j}}) \in U$ for $1 \leqslant \ell \leqslant r$. In particular, $b(p\mathbf{n} + \mathbf{j}) = 0$ if and only if $b_1(\mathbf{n}) = b_2(\mathbf{n}) = \cdots = b_r(\mathbf{n}) = 0$.

Given $b : \mathbb{N}^d \to \mathbb{K}_0$, we let $\chi_b : \mathbb{N}^d \to \{0,1\}$ be defined by

$$\chi_b(\mathbf{n}) \;=\; \begin{cases} 0 \text{ if } b(\mathbf{n}) \neq 0 \\ 1 \text{ if } \; b(\mathbf{n}) = 0\,. \end{cases}$$

Set

$$\mathcal{S} := \{\chi_{b_1} \cdots \chi_{b_t} \;:\; t \geqslant 0, b_1, \ldots, b_t \in U\}.$$

We note that since $\chi_b^2 = \chi_b$ for all $b \in U$ and U is finite, $\mathcal{S}$ is finite. Note that if $b \in U$ and $\mathbf{j} \in \{0,1,\ldots,p-1\}^d$, then $b_\ell := \pi_\ell(b \circ e_{\mathbf{j}}) \in U$ for $1 \leqslant \ell \leqslant r$. By the above remarks,

$$(\chi_b \circ e_{\mathbf{j}})(\mathbf{n}) = \prod_{\ell=1}^{r} \chi_{b_\ell}(\mathbf{n})\,,$$

and so we see that if $\chi \in \mathcal{S}$ then $\chi \circ e \in \mathcal{S}$ for all $e \in \Sigma$. Since $\mathcal{S}$ is finite, this proves that $\chi : \mathbb{N}^d \to \{0,1\}$ is p-automatic. In particular, since $a(\mathbf{n}) = a_1(\mathbf{n}) \in U$, we obtain that χ_a is p-automatic. In other words, the set of $\mathbf{n} \in \mathbb{N}^d$ such that $a(\mathbf{n}) = 0$ is a p-automatic set. This ends the proof. □

6 The algebraic closure of $\mathbb{F}_p(t)$

6.1 Christol's theorem

One of the most beautiful results in the theory of automatic sequences is Christol's theorem, which characterizes Laurent series with coefficients in a finite field which are algebraic over the field of rational functions.

Theorem 6.1 (Christol). *Let $\mathbb{K}$ be a finite field of characteristic $p > 0$. Then $f(t) = \sum_{n\geqslant 0}^{\infty} a(n)t^n \in \mathbb{K}((t))$ is algebraic over $\mathbb{K}(t)$ if and only if the sequence $a(n)$ is p-automatic.*

Christol's theorem consists of two parts: the "easy" direction in which one shows that if the sequence of coefficients of a Laurent series is p-automatic, then the Laurent series is algebraic, and the "hard" direction in which one must show that the coefficients of an algebraic Laurent series form a p-automatic sequence. The hard direction is generally proved using Ore's lemma, which is the observation that if $f(t)$ is algebraic over a field $\mathbb{K}(t)$, then the set $\{f, f^p, f^{p^2}, \ldots\}$ is linearly dependent over $\mathbb{K}(t)$. Christol's theorem was generalized to multivariate Laurent series by Salon [44].

Theorem 6.2 (Salon). *Let $\mathbb{K}$ be a finite field of characteristic $p > 0$. Then $f(\mathbf{t}) = \sum_{\mathbf{n}\in\mathbb{N}^d} a(\mathbf{n})\mathbf{t}^{\mathbf{n}} \in \mathbb{K}((\mathbf{t}))$ is algebraic if and only if the sequence $a(\mathbf{n})$ is p-automatic.*

Salon's theorem turns out to be a special case of Theorems 5.6 and 5.4.

Proof of Theorem 6.2. We suppose first that $a : \mathbb{N}^d \to \mathbb{K}$ is p-automatic and we consider the power series

$$f(\mathbf{t}) := \sum_{\mathbf{n} \in \mathbb{N}^d} a(\mathbf{n}) \mathbf{t}^{\mathbf{n}} .$$

Using the notation of Equations 5.1 and 5.2, we infer from Remark 5.5 that there are only finitely many distinct functions of the form $a \circ e$ where e runs over Σ. Consequently, there are only finitely many functions of the form $E \circ f$ where E runs over Ω. Thus $\Omega(f)$ is finite-dimensional and Theorem 5.6 implies that $f(\mathbf{t})$ is algebraic.

We next suppose that $f(\mathbf{t})$ is algebraic and let $c \in \mathbb{K}$. Since $f(\mathbf{t})$ is algebraic, then so is $f(\mathbf{t}) - c$ and by Theorem 5.4 the set S_c of d-tuples of natural numbers $\mathbf{n}$ such that $a(\mathbf{n}) - c = 0$ is p-automatic. It follows that the sequence $a_c : \mathbb{N}^d \to \mathbb{K}$ given by $a_c(\mathbf{n}) = 1$ if $\mathbf{n} \in S_c$ and $a_c(\mathbf{n}) = 0$ otherwise is p-automatic. Thus $a(\mathbf{n}) = \sum_{c \in \mathbb{K}} c a_c(\mathbf{n})$ is also p-automatic as p-automatic sequences taking values in a field are closed under the taking of finite sums and scalar multiplication. □

While Christol's theorem gives a concrete description of the elements of $\mathbb{F}_q((t))$ that are algebraic over $\mathbb{F}_q(t)$, it does not give the whole picture. As Kedlaya [25] points out, the field $\mathbb{F}_q((t))$ is far from being algebraically closed. Indeed, for an algebraically closed field $\mathbb{K}$ of characteristic 0, a classical result of Puiseux is that the field

$$\bigcup_{i=1}^{\infty} \mathbb{K}((t^{1/i}))$$

is itself algebraically closed and contains, in particular, the algebraic closure of $\mathbb{K}(t)$. However, over field of positive characteristic, the situation is more subtle. In particular, the algebraic closure of $\mathbb{F}_q((t))$ is much more complicated to describe, due to the existence of wildly ramified field extensions. For instance, Chevalley remarked [8] that the Artin-Schreier polynomial $x^p - x - 1/t$ does not split in the Puiseux field $\bigcup_{n=1}^{+\infty} \mathbb{F}_q((t^{1/n}))$.

6.2 Generalized power series

It turns out that the appropriate framework to describe the algebraic closure of $\mathbb{F}_p(t)$ is provided by the fields of generalized power series $\mathbb{F}_q((t^{\mathbb{Q}}))$ introduced by Hahn [20]. We briefly describe this construction.

We recall that a subset S of a totally ordered group is said to be *well-ordered* if every nonempty subset of S has a minimal element or, equivalently, if there is no infinite decreasing sequence within S. Given a commutative ring R and a totally ordered Abelian group G we construct a commutative ring, denoted by $R((t^G))$, which is defined to be the collection of all elements of the form

$$f(t) \ := \ \sum_{\alpha \in G} r_\alpha t^\alpha$$

which satisfy the following conditions.

(i) $r_\alpha \in R$ for all $\alpha \in G$.

(ii) The support of $f(t)$ is well ordered, that is, the subset $\{\alpha \mid r_\alpha \neq 0_R\}$ is a well-ordered set.

Addition and multiplication are defined via the rules

$$\sum_{\alpha\in G} r_\alpha t^\alpha + \sum_{\alpha\in G} s_\alpha t^\alpha = \sum_{\alpha\in G} (r_\alpha + s_\alpha) t^\alpha$$

and

$$\left(\sum_{\alpha\in G} r_\alpha t^\alpha\right)\left(\sum_{\alpha\in G} s_\alpha t^\alpha\right) = \sum_{\alpha\in G}\sum_{\beta\in G} (r_\beta s_{\alpha-\beta}) t^\alpha .$$

We note that the fact that the supports of valid series expansions are well-ordered means that no problems with possible infinite sums appearing in the expression for the coefficients in a product of two generalized power series will occur. We call the ring $R((t^G))$ the *ring of generalized power series over R with exponents in G.*

We recall that a group is *divisible* if for every $g \in G$ and $n \geqslant 1$, there exists some $h \in G$ such that $h^n = g$. For an algebraically closed field $\mathbb{K}$ and a totally ordered divisible Abelian group G, the field $\mathbb{K}((t^G))$ is known to be algebraically closed [24] (see also [25, 47]). In what follows, we will only consider the particular case of the divisible group $\mathbb{Q}$ and of a finite field $\mathbb{F}_q$ (q being a power of a prime p).

We then have the series of containments

$$\mathbb{F}_q(t) \subset \mathbb{F}_q((t)) \subset \mathbb{F}_q((t^{\mathbb{Q}})) .$$

Though $\mathbb{F}_q((t^{\mathbb{Q}}))$ is not algebraically closed, it is sufficient for our purpose to consider such fields. Indeed, taking $\bigcup_{n\geqslant 1} \mathbb{F}_{p^n}$ as an algebraic closure of $\mathbb{F}_p$, it follows from the remark above that the field $\left(\bigcup_{n\geqslant 1} \mathbb{F}_{p^n}\right)((t^{\mathbb{Q}}))$ is algebraically closed. For example, the Artin-Schreier polynomial $x^p - x - 1/t$ does split in $\mathbb{F}_p((t^{\mathbb{Q}}))$. Indeed, we can check that the generalized power series

$$c + \sum_{i=1}^{\infty} t^{-1/p^i}, \; c \in \mathbb{F}_p ,$$

are the roots of this polynomial.

6.3 Kedlaya's theorem

Kedlaya [25] considered whether one can, as in Christol's theorem, give an automaton-theoretic characterization of the elements of $\mathbb{F}_q((t^{\mathbb{Q}}))$ that are algebraic over $\mathbb{F}_q(t)$. The work of Kedlaya [26] is thus precisely devoted to a description of the algebraic closure of $\mathbb{F}_p(t)$ as a subfield of generalized power series. For this purpose, Kedlaya introduces the notion of a p-quasi-automatic function over the rational numbers.

Kedlaya uses automata to produce power series whose exponents take values in the rational numbers. Hence it is necessary to create automata which accept rational numbers as opposed to just accepting integers. We now explain how Kedlaya does this.

Let $k > 1$ be a positive integer. We set

$$\Sigma'_k = \{0, 1, \ldots, k-1, \bullet\}$$

and we denote by $\mathcal{L}(k)$ the language on the alphabet Σ'_k consisting of all words on Σ'_k with exactly one occurrence of the letter '$\bullet$' (the radix point) and whose first and last letters are not equal to 0. This is a regular language [26, Lemma 2.3.3]. We let S_k denote the set of nonnegative k-adic rationals, that is,

$$S_k = \{a/k^b \mid a, b \in \mathbb{Z},\ a \geqslant 0\}.$$

We note that there is a bijection $[\,\cdot\,]_k : \mathcal{L}(k) \to S_k$ given by

$$s_1 \cdots s_{i-1} \bullet s_{i+1} \cdots s_n \in \mathcal{L}(k) \mapsto \sum_{j=1}^{i-1} s_j k^{i-1-j} + \sum_{j=i+1}^{n} s_j k^{i-j},$$

where $s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n \in \{0, 1, \ldots, k-1\}$. So, for example, we have $[110\bullet32]_4 = [20\bullet875]_{10} = 167/8$. We also note that the fact that we exclude strings whose initial and terminal letters are 0 means that we have the awkward looking expression $[\,\bullet\,]_k = 0$.

Definition 6.1. We say that a map $h : S_k \to \Delta$ is *k-automatic* if there is a finite state machine which takes words on Σ'_k as input such that for each $W \in \mathcal{L}_k$, $h([W]_k)$ is generated by the machine using the word W as input.

Since the support of a generalized power series is well-ordered, we need a more general notion of automatic functions defined over the set of rationals. For this purpose, we always implicitly consider sets Δ containing a special element called zero and denoted by 0 (of course, when Δ is a subset of $\mathbb{R}$ or $\mathbb{N}$, or if it denotes a finite field, zero will preserve its usual meaning). Then we will talk about functions $h : \mathbb{Q} \to \Delta$ as being k-automatic if their support is contained in S_k and the restriction of h to S_k is k-automatic (the support of such a function being defined as the set $S = \{\alpha \in \mathbb{Q} \mid h(\alpha) \neq 0\}$).

Example 6.1. For $w \in \mathcal{L}(2)$, define

$$h([w]_2) = \begin{cases} 0 \text{ if there are an even number of } 1\text{'s in } w \\ 1 \text{ otherwise.} \end{cases}$$

Then $h : S_2 \to \{0, 1\}$ is K$_2$-automatic.

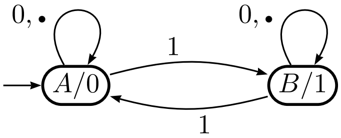


Figure 7. The DFAO associated with the function h of Example 6.1.

Definition 6.2. Let k be a positive integer, let Δ be a finite set containing a special element 0, and let $h : \mathbb{Q} \to \Delta$. We say that h is k-quasi-automatic if it satisfies the following conditions.

(i) The support S of h is well-ordered.
(ii) There exist a positive integer a and an integer b such that the set $aS + b$ consists of nonnegative k-adic rationals and the map $h((x - b)/a)$ is a k-automatic function from S_k to Δ.

We are now ready to state Kedlaya's theorem.

Theorem 6.3 (Kedlaya). *Let p be a prime, let q be a power of p, and let $a : \mathbb{Q} \to \mathbb{F}_q$. Then $\sum_{\alpha\in\mathbb{Q}} a(\alpha)t^\alpha$ is algebraic over $\mathbb{F}_q(t)$ if and only if the function $a : \mathbb{Q} \to \mathbb{F}_q$ is p-quasi-automatic.*

In light of Salon's result [44], Kedlaya asked whether his theorem has an extension to multivariate generalized power series $\mathbb{F}_q((t_1^{\mathbb{Q}}, \ldots, t_m^{\mathbb{Q}}))$. As far as we know, this problem has not yet been solved.

References

[1] B. Adamczewski and J. Bell. On vanishing coefficients of algebraic power series over fields of positive characteristic. *Invent. Math.*, 187:343–393, 2012. 846

[2] B. Adamczewski and Y. Bugeaud. On the complexity of algebraic numbers. I. Expansions in integer bases. *Ann. of Math. (2)*, 165:547–565, 2007. 835

[3] B. Adamczewski and J. Cassaigne. Diophantine properties of real numbers generated by finite automata. *Compos. Math.*, 142(6):1351–1372, 2006. 836

[4] J.-P. Allouche and J. O. Shallit. *Automatic Sequences, Theory, Applications, Generalizations*. Cambridge University Press, 2003. 827, 832

[5] J.-P. Bézivin. Une généralisation du théorème de Skolem-Mahler-Lech. *Quart. J. Math. Oxford Ser. (2)*, 40(158):133–138, 1989. 842

[6] E. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti Circ. Mat. Palermo*, 27:247–271, 1909. 834

[7] J. R. Chen. On the representation of a large even integer as the sum of a prime and the product of at most two primes. *Kexue Tongbao (Foreign Lang. Ed.)*, 17:385–386, 1966. 833

[8] C. Chevalley. *Introduction to the Theory of Algebraic Functions of One Variable*. Mathematical Surveys, No. VI. American Mathematical Society, New York, N. Y., 1951. 851

[9] A. Cobham. On the Hartmanis-Stearns problem for a class of tag machines. In *IEEE Conference Record of 1968 Ninth Annual Symposium on Switching and Automata Theory*, pages 51–60, 1968. Also appeared as IBM Research Technical Report RC-2178, August 23 1968. 835

[10] A. Cobham. Uniform tag sequences. *Math. Systems Theory*, 6:164–192, 1972. 828

[11] H. Cramér. On the order of magnitude of the difference between consecutive prime numbers. *Acta Arith.*, 2:23–46, 1936. 832

[12] H. Derksen. A Skolem-Mahler-Lech theorem in positive characteristic and finite automata. *Invent. Math.*, 168(1):175–224, 2007. 844

[13] 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. 827, 829, 831

[14] 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. 842

[15] S. Ferenczi and C. Mauduit. Transcendence of numbers with a low complexity expansion. *J. Number Theory*, 67:146–161, 1997. 838

[16] E. Fouvry and C. Mauduit. Sommes des chiffres et nombres presque premiers. *Math. Ann.*, 305:571–599, 1996. 833

[17] A. O. Gelfond. Sur les nombres qui ont des propriétés additives et multiplicatives données. *Acta Arith.*, 13:259–265, 1967/1968. 834

[18] D. A. Goldston, J. Pintz, and C. Y. Yıldırım. Primes in tuples. I. *Ann. of Math. (2)*, 170:819–862, 2009. 832

[19] B. Green and T. Tao. The primes contain arbitrarily long arithmetic progressions. *Ann. of Math. (2)*, 167:481–547, 2008. 832

[20] H. Hahn. *Gesammelte Abhandlungen/Collected works. Band 1/Vol. 1*. Springer-Verlag, Vienna, 1995. With biographical sketches by Karl Popper and by L. Schmetterer and K. Sigmund, and commentaries on Hahn's work by H. Heuser, H. Sagan and L. Fuchs, Edited by Schmetterer and Sigmund and with a foreword by Popper. 851

[21] G. Hansel. Une démonstration simple du théorème de Skolem-Mahler-Lech. *Theoret. Comput. Sci.*, 43(1):91–98, 1986. 842

[22] J. Hartmanis and H. Shank. On the recognition of primes by automata. *J. Assoc. Comput. Mach.*, 15:382–389, 1968. 832

[23] J. Hartmanis and R. E. Stearns. On the computational complexity of algorithms. *Trans. Amer. Math. Soc.*, 117:285–306, 1965. 834

[24] I. Kaplansky. Maximal fields with valuations. *Duke Math. J.*, 9:303–321, 1942. 852

[25] K. S. Kedlaya. The algebraic closure of the power series field in positive characteristic. *Proc. Amer. Math. Soc.*, 129(12):3461–3470 (electronic), 2001. 851, 852

[26] K. S. Kedlaya. Finite automata and algebraic extensions of function fields. *J. Théor. Nombres Bordeaux*, 18(2):379–420, 2006. 852, 853

[27] C. Lech. A note on recurring series. *Ark. Mat.*, 2:417–421, 1953. 842, 843

[28] J. Liouville. Sur des classes très étendues de quantités dont la valeur n'est ni algébrique, ni même reductible à des irrationelles algébriques. *C. R. Acad. Sci. Paris*, 18:883–885, 910–911, 1844. 835

[29] J. H. Loxton and A. J. van der Poorten. Arithmetic properties of the solutions of a class of functional equations. *J. Reine Angew. Math.*, 330:159–172, 1982. 835

[30] J. H. Loxton and A. J. van der Poorten. Arithmetic properties of automata: regular sequences. *J. Reine Angew. Math.*, 392:57–69, 1988. 835

[31] K. Mahler. Arithmetische Eigenschaften der Lösungen einer Klasse von Funktionalgleichungen. *Math. Annalen*, 101:342–366, 1929. Corrigendum, **103** (1930), 532. 835

[32] K. Mahler. Eine arithmetische eigenshaft der taylor-koeffizienten rationaler funktionen. In *Proc. Kon. Nederlandsche Akad. v. Wetenschappen*, volume 38, pages 50–60. 1935. 842

[33] K. Mahler. On the Taylor coefficients of rational functions. *Proc. Cambridge Philos. Soc.*, 52:39–48, 1956. 842

[34] K. Mahler. Addendum to the paper "On the Taylor coefficients of rational functions". *Proc. Cambridge Philos. Soc.*, 53:544, 1957. 842

[35] C. Mauduit and J. Rivat. Sur un problème de gelfond : la somme des chiffres des nombres premiers. *Ann. of Math. (2)*, 171:1591–1646, 2010. 834

[36] P. Mihăilescu. Primary cyclotomic units and a proof of Catalan's conjecture. *J. Reine Angew. Math.*, 572:167–195, 2004. 846

[37] M. Minsky and S. Papert. Unrecognizable sets of numbers. *J. Assoc. Comput. Mach.*, 13:281–286, 1966. 830, 832

[38] M. Morse and G. A. Hedlund. Symbolic Dynamics. *Amer. J. Math.*, 60:815–866, 1938. 834

[39] J. Pintz. Cramér vs. Cramér. On Cramér's probabilistic model for primes. *Funct. Approx. Comment. Math.*, 37(, part 2):361–376, 2007. 832

[40] P. Ribenboim. *The new book of prime number records*. Springer-Verlag, New York, 1996. 833

[41] D. Ridout. Rational approximations to algebraic numbers. *Mathematika*, 4:125–131, 1957. 837

[42] R. W. Ritchie. Finite automata and the set of squares. *J. Assoc. Comput. Mach.*, 10:528–531, 1963. 830

[43] K. F. Roth. Rational approximations to algebraic numbers. *Mathematika*, 2:1–20, 1955. Corrigendum, p. 168. 836

[44] O. Salon. Suites automatiques à multi-indices et algébricité. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(12):501–504, 1987. 830, 831, 850, 854

[45] W. M. Schmidt. *Diophantine Approximation*, volume 785 of *Lecture Notes in Mathematics*. Springer-Verlag, 1980. 839

[46] M.-P. Schützenberger. A remark on acceptable sets of numbers. *J. Assoc. Comput. Mach.*, 15:300–303, 1968. 832

[47] J.-P. Serre. *Local fields*, volume 67 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1979. Translated from the French by Marvin Jay Greenberg. 852

[48] H. Sharif and C. F. Woodcock. Algebraic functions over a field of positive characteristic and Hadamard products. *J. London Math. Soc. (2)*, 37(3):395–403, 1988. 847

[49] T. Skolem. Ein verfahren zur behandlung gewisser exponentialer gleichungen und diophantischer gleichungen. In *C. r. 8 congr. scand. à Stockholm*, pages 163–188. 1934. 842

[50] K. Soundararajan. The distribution of prime numbers. In *Equidistribution in number theory, an introduction*, volume 237 of *NATO Sci. Ser. II Math. Phys. Chem.*, pages 59–83. Springer, Dordrecht, 2007. 832

[51] K. Soundararajan. Small gaps between prime numbers: the work of Goldston-Pintz-Yıldırım. *Bull. Amer. Math. Soc. (N.S.)*, 44:1–18, 2007. 832

[52] T. Tao. *Structure and randomness*. American Mathematical Society, Providence, RI, 2008. Pages from year one of a mathematical blog. 832, 842

[53] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.*, 42:230–265, 1936. 834

[54] A. J. van der Poorten. Some facts that should be better known, especially about rational functions. In *Number theory and applications (Banff, AB, 1988)*, volume 265 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 497–528. Kluwer Acad. Publ., Dordrecht, 1989. 842

On Cobham's theorem

Fabien Durand[1] *and Michel Rigo*[2]

[1]LAMFA - Université de Picardie Jules Verne - CNRS UMR 6140
33 rue Saint Leu, F-80039 Amiens cedex 1, France
email: fabien.durand@u-picardie.fr

[2]Université de Liège, Institut de Mathématiques
12 Grande Traverse (B37), B-4000 Liège, Belgique
email: M.Rigo@ulg.ac.be

2010 Mathematics Subject Classification: 68Q45, 68R15, 11B85, 11A67, 11U05, 37B20

Key words: Numeration systems, recognizable sets of integers, Cobham's theorem, ultimate periodicity, substitutions, dynamical systems.

Contents

1 Introduction

In this chapter we essentially focus on the representation of non-negative integers in a given numeration system. The main role of such a system — like the usual integer base k numeration system — is to replace numbers or more generally sets of numbers by their corresponding representations, *i.e.*, by words or by languages. First we consider integer base numeration systems to present the main concepts but rapidly we will introduce non-standard systems and their relationships with substitutions.

Let $k \in \mathbb{N}_{\geqslant 2}$ be an integer where $\mathbb{N}_{\geqslant 2}$ denotes the set of non-negative integers larger or equal to 2. The set $\{0, \ldots, k\}$ is denoted by $[\![0, k]\!]$. If we do not allow leading zeroes when representing numbers, the function mapping a non-negative integer n onto its k-ary representation $\mathrm{rep}_k(n) \in [\![0, k-1]\!]^*$ is a one-to-one correspondence. In the literature, one also finds notation like $\langle n \rangle_k$, $(n)_k$ or $\rho_k(n)$ instead of $\mathrm{rep}_k(n)$. In particular, 0 is assumed to be represented by the empty word ε. Hence any set $X \subseteq \mathbb{N}$ is associated with the language $\mathrm{rep}_k(X)$ consisting of the k-ary representations of the elements of X.

It is natural to study the relation existing between the arithmetic or number-theoretic properties of integers and the syntactical properties of the corresponding representations in a given numeration system. We focus on those sets $X \subseteq \mathbb{N}$ for which a finite automaton can be used to decide for any given word w over $[\![0, k-1]\!]$ whether or not w belongs to $\mathrm{rep}_k(X)$. Sets having the property that $\mathrm{rep}_k(X)$ is regular[1] are called *k-recognizable* sets. Such a set can be considered as a particularly simple set because using the k-ary numeration system it has a somehow elementary algorithmic description. In the framework of infinite-state systems verification, one also finds the terminology of *Number Decision Diagram* or *NDD* [122].

The essence of Cobham's theorem is to express that the property for a set to be recognizable by a finite automaton *strongly depends* on the choice of the base and more generally on the considered numeration system. Naturally this fact leads to and motivates the introduction and the study of recognizable sets in non-standard numeration systems. Considering alternative numeration systems may provide new recognizable sets and these non-standard systems also have applications in computer arithmetic [59]. Last but not least, the proof of Cobham's theorem is non-trivial and relies on quite elaborate argu-

[1] We use the terminology of regular language, instead of rational language.

ments.

Now let us state this celebrated result from 1969 and give all the needed details and definitions. Several surveys have been written on the same subject, see [25, 26, 28, 98].

Theorem 1.1 (Cobham's theorem [35]). *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. A set $X \subseteq \mathbb{N}$ is both k-recognizable and ℓ-recognizable if and only if it is ultimately periodic.*

In the various contexts that we will describe, showing that an ultimately periodic set is recognizable is always the easy direction to prove. See Remark 1.3. So we focus on the other direction.

Definition 1.1. A subset of $\mathbb{N}$ is *ultimately periodic* if it is the union of a finite set and a finite number of infinite arithmetic progressions. In particular, X is ultimately periodic if and only if there exist $N \geqslant 0$ and $p \geqslant 1$ such that for all $n \geqslant N$, $n \in X \Leftrightarrow n + p \in X$. Recall that an *arithmetic progression* is a set of the kind $a\mathbb{N} + b := \{an + b \mid n \geqslant 0\}$.

Definition 1.2. Let $\alpha, \beta > 1$ be two real numbers. If the equation $\alpha^m = \beta^n$ with $m, n \in \mathbb{N}$ has only the trivial integer solution $m = n = 0$, then α and β are said to be *multiplicatively independent*. Otherwise, α and β are said to be *multiplicatively dependent*.

Let $k, \ell \geqslant 2$ be two integers. Notice that k and ℓ are multiplicatively independent if and only if $\log k / \log \ell$ is irrational. Note that for k and ℓ to be multiplicatively dependent it is not enough that k and ℓ share exactly the same prime factors occurring in their decomposition. For instance, 6 and 18 are multiplicatively independent. But coprime integers are multiplicatively independent.

The irrationality of $\log k / \log \ell$ is a crucial point in the proof of Cobham's theorem (see Subsection 5.3). Recall that if $\theta > 0$ is irrational, then the set $\{\{n\theta\} \mid n > 0\}$ of fractional parts of the multiples of θ is dense in $[0, 1]$. For a proof of the so-called *Kronecker's theorem*, see [66].

Remark 1.2. The fact for two integers to be multiplicatively dependent is an equivalence relation $\mathfrak{M}$ over $\mathbb{N}_{\geqslant 2}$. If k and ℓ are multiplicatively dependent, then there exist a minimal $q \geqslant 2$ and two positive integers m, n such that $k = q^m$ and $\ell = q^n$. Let us give the first (with respect to their minimal element) few equivalence classes for $\mathfrak{M}$ partitioning $\mathbb{N}_{\geqslant 2}$: $[2]_{\mathfrak{M}}$, $[3]_{\mathfrak{M}}$, $[5]_{\mathfrak{M}}$, $[6]_{\mathfrak{M}}$, $[7]_{\mathfrak{M}}$, $[10]_{\mathfrak{M}}$, $[11]_{\mathfrak{M}}$, $[12]_{\mathfrak{M}}, \ldots$.

Remark 1.3. We show that if a set $X \subseteq \mathbb{N}$ is ultimately periodic then, for all $k \geqslant 2$, X is k-recognizable. In the literature, one also finds the terminology of a *recognizable set* X (without any mention to a base), meaning that X is k-recognizable for all $k \geqslant 2$. Note that a finite union of regular languages is again a regular language. Hence it is enough to check that $\operatorname{rep}_k(a\mathbb{N} + b)$ is regular with $0 \leqslant b < a$. We can indeed assume that $b < a$ because if we add or remove a finite number of words to a regular language, we still have a regular language. Consider a DFA having $Q = [\![0, a-1]\!]$ as its set of states. For all $i \in Q$, $d \in [\![0, k-1]\!]$, the transitions are given by

$$i \xrightarrow{d} ki + d \mod a.$$

The initial state is 0 and the unique final state is b. As an example, a DFA accepting exactly binary representations of the integers congruent to 3 mod 4 is given in Figure 1. A study of the minimal automaton recognizing such divisibility criteria expressed in an integer base is given in [3]. See also the discussion in [109, Prologue]. The fact that a divisibility criterion exists in every base for any fixed divisor was already observed by Pascal in [97, pp. 84–89].

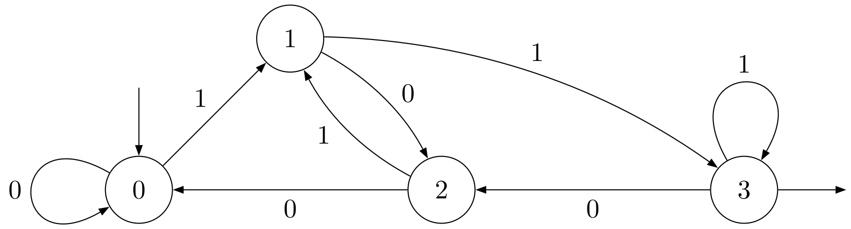


Figure 1. A finite automaton accepting $\text{rep}_2(4\mathbb{N}+3)$.

2 Numeration basis

It is remarkable that the recognizability of ultimately periodic sets extends to wider contexts (see Proposition 2.6 and Theorem 5.1). Let us introduce a first generalization of the integer base numeration system.

Definition 2.1. A *numeration basis* is a sequence $U = (U_n)_{n \geqslant 0}$ of integers such that U is increasing, $U_0 = 1$ and that the set $\{U_{i+1}/U_i \mid i \geqslant 0\}$ is bounded. This latter condition ensures the finiteness of the alphabet of digits used to represent integers. If $w = w_\ell \cdots w_0$ is a word over a finite alphabet $A \subset \mathbb{Z}$ then the numerical value of w is

$$\pi_{A,U}(w) = \sum_{i=0}^{\ell} w_i \, U_i.$$

Using the greedy algorithm [57], any integer n has a unique *(normal) U-representation* $\text{rep}_U(n) = w_\ell \cdots w_0$ which is a finite word over a minimal finite alphabet called the *canonical alphabet* of U and denoted by A_U. The normal U-representation satisfies

$$\pi_{A_U,U}(\text{rep}_U(n)) = n \text{ and for all } i \in [\![0, \ell-1]\!],\ \pi_{A_U,U}(w_i \cdots w_0) < U_{i+1}.$$

Again, $\text{rep}_U(0) = \varepsilon$. See [85, Chapter 7] or Ch. Frougny and J. Sakarovitch's chapter in [12, Chapter 2]. A subset $X \subseteq \mathbb{N}$ is *U-recognizable* if $\text{rep}_U(X)$ is accepted by a finite automaton. Let $B \subset \mathbb{Z}$ be a finite alphabet. If $w \in B^*$ is such that $\pi_{B,U}(w) \geqslant 0$, then the function mapping w onto $\text{rep}_U(\pi_{B,U}(w))$ is called *normalization*.

Definition 2.2. A numeration basis U is said to be *linear* if there exist $k \in \mathbb{N} \setminus \{0\}$, $d_1, \ldots, d_k \in \mathbb{Z}$, $d_k \neq 0$, such that, for all $n \geqslant k$, $U_n = d_1 U_{n-1} + \cdots + d_k U_{n-k}$. The polynomial $P_U(X) = X^k - d_1 X^{k-1} - \cdots - d_{k-1}X - d_k$ is called the *characteristic polynomial* of U.

Definition 2.3. Recall that a *Pisot-Vijayaraghavan number* is an algebraic integer $\beta > 1$ whose Galois conjugates have modulus strictly less than one. We say that $U = (U_n)_{n\geqslant 0}$ is a *Pisot numeration system* if the numeration basis U is linear and $P_U(X)$ is the minimal polynomial of a Pisot number β. Integer base numeration systems are particular cases of Pisot systems. For instance, see [27] where it is shown that most properties related to k-recognizable sets, $k \in \mathbb{N}_{\geqslant 2}$, can be extended to Pisot systems. In such a case, there exists some $c > 0$ such that $|U_n - c\,\beta^n| \to 0$, as n tends to infinity.

Example 2.1. Consider the Fibonacci sequence defined by $U_0 = 1$, $U_1 = 2$ and $U_{n+2} = U_{n+1} + U_n$ for all $n \geqslant 0$. A word over $\{0, 1\}$ is a U-representation if and only if it belongs to the language $L = 1\{0, 01\}^* \cup \{\varepsilon\}$. For instance 10110 is not a U-representation. Since $\pi_{A_U,U}(10110) = 13$, the normalization maps 10110 to $\mathrm{rep}_U(13) = 100000$. The characteristic polynomial of this linear numeration basis is the minimal polynomial of the Pisot number $(1 + \sqrt{5})/2$. This Pisot numeration system is presented in [123].

The following result is an easy exercise but also can be carried on in a wider context.

Theorem 2.1. [115] *Let U be a numeration basis. If $\mathbb{N}$ is U-recognizable, then U is linear.*

Definition 2.4. [13] A *Bertrand numeration basis* U is a numeration basis satisfying the following property: $w \in \mathrm{rep}_U(\mathbb{N})$ if and only if, for all $n \in \mathbb{N}$, $w0^n \in \mathrm{rep}_U(\mathbb{N})$. It is a natural condition satisfied by all integer base $k \geqslant 2$ systems. For instance, the sequence defined by $U_0 = 1$, $U_1 = 3$ and, for all $n \geqslant 0$, $U_{n+2} = U_{n+1} + U_n$ is not a Bertrand numeration basis because $\mathrm{rep}_U(2) = 2$, but $\pi_{A_U,U}(20) = 6$ and $\mathrm{rep}_U(6) = 102$.

Let $\alpha > 1$ be a real number. The notion of α-expansion was introduced by Parry in [96], (also see Rényi's paper [104]). See again [85, Chapter 7]. All $x \in [0, 1]$ can be uniquely written in the following way:

$$x = \sum_{n\geqslant 1} a_n \alpha^{-n}, \tag{2.1}$$

with $x_1 = x$ and for all $n \geqslant 1$, $a_n = \lfloor \alpha\, x_n \rfloor$ and $x_{n+1} = \{\alpha x_n\}$, where $\lfloor\cdot\rfloor$ stands for the integer part. The sequence $d_\alpha(x) = (a_n)_{n\geqslant 1}$ is the *α-expansion* of x and $L(\alpha)$ denotes the set of finite words having an occurrence in some sequences $d_\alpha(x)$, $x \in [0, 1]$. Let $d_\alpha(1) = (t_n)_{n\geqslant 1}$. If there exist $N \geqslant 0$, $p > 0$ such that, for all $n \geqslant N$, $t_{n+p} = t_n$ then α is said to be a *Parry number*, sometimes called a *β-number* (for more details or information about these numbers, see [96] or [58]). Observe that integers greater or equal to 2 are Parry numbers.

The following result relates Bertrand numeration systems to languages defined by some real number.

Theorem 2.2 (A. Bertrand-Mathis [14]). *Let U be a numeration basis. It is a Bertrand numeration basis if and only if there exists a real number $\alpha > 1$ such that $\mathrm{rep}_U(\mathbb{N}) = L(\alpha)$. In this case, if U is linear then α is a root of the characteristic polynomial of U.*

Theorem 2.3 (A. Bertrand-Mathis [13]). *Let $\alpha > 1$ be a real number. The language $L(\alpha)$ is regular if and only if α is a Parry number.*

Associated with a Parry number β, one can define the notion of beta-polynomial. For details see [68] or [12, Chapter 2]. First we define the *canonical beta-polynomial*. If $d_\beta(1)$ is eventually constant and equal to 0: $d_\beta(1) = t_1 \cdots t_m 0^\omega$, with $t_m \neq 0$, then we set $G_\beta(X) = X^m - \sum_{i=1}^{m} t_i X^{m-i}$ and $r = m$. Otherwise, $d_\beta(1)$ is eventually periodic: $d_\beta(1) = t_1 \cdots t_m (t_{m+1} \cdots t_{m+p})^\omega$, with m and p being minimal. Then we set $G_\beta(X) = X^{m+p} - \sum_{i=1}^{m+p} t_i X^{m+p-i} - X^m + \sum_{i=1}^{m} t_i X^{m-i}$ and $r = p$. Let β be a Parry number. An *extended beta-polynomial* is a polynomial of the form $H_\beta(X) = G_\beta(X)(1 + X^r + \cdots + X^{rk})X^n$ for $k, n \in \mathbb{N}$.

Proposition 2.4. [68] *Let U be a linear numeration basis with dominant root β, i.e., $\lim_{n\to\infty} U_{n+1}/U_n = \beta$ for some $\beta > 1$. If $\mathrm{rep}_U(\mathbb{N})$ is regular, then β is a Parry number.*

Theorem 2.5 (M. Hollander [68]). *Let U be a linear numeration basis whose dominant root β is a Parry number.*

- *If $d_\beta(1)$ is infinite and eventually periodic, then $\mathrm{rep}_U(\mathbb{N})$ is regular if and only if U satisfies an extended beta-polynomial for β.*
- *If $d_\beta(1)$ is finite of length m, then: if U satisfies an extended beta-polynomial for β then $\mathrm{rep}_U(\mathbb{N})$ is regular; and conversely if $\mathrm{rep}_U(\mathbb{N})$ is regular, then U satisfies either an extended beta-polynomial for β, $H_\beta(X)$, or a polynomial of the form $(X^m - 1)H_\beta(X)$.*

Ultimately periodic sets are recognizable for any linear numeration basis.

Proposition 2.6 (Folklore [12, 85]). *Let $a, b \geqslant 0$. If $U = (U_n)_{n\geqslant 0}$ is a linear numeration basis, then*

$$\pi^{-1}_{A_U,U}(a\,\mathbb{N} + b) = \left\{ c_\ell \cdots c_0 \in A_U^* \mid \sum_{k=0}^{\ell} c_k\, U_k \in a\,\mathbb{N} + b \right\}$$

is accepted by a DFA that can be effectively constructed. In particular, if $\mathbb{N}$ is U-recognizable, then any ultimately periodic set is U-recognizable.

To conclude this section, consider again the integer base numeration systems.

Example 2.2. The set $P_2 = \{2^n \mid n \geqslant 0\}$ of powers of two is trivially 2-recognizable because $\mathrm{rep}_2(P_2) = 10^*$. Since the difference between any two consecutive elements in P_2 is of the kind $2^{n+1} - 2^n = 2^n$, P_2 is not ultimately periodic. As a consequence of Cobham's theorem, P_2 is for instance neither 3-recognizable nor 5-recognizable.

One could also consider the case when the two bases k and ℓ are multiplicatively dependent. This case is much easier and can be considered as an exercise.

Proposition 2.7. *Let $k, \ell \geqslant 2$ be two multiplicatively dependent integers. A set $X \subseteq \mathbb{N}$ is k-recognizable if and only if it is ℓ-recognizable.*

The theorem of Cobham implies that ultimately periodic sets are the only infinite sets that are k-recognizable for every $k \geqslant 2$. We have seen so far that there exist sets (like the set P_2 of powers of two) that are only recognizable for some specific bases: exactly all bases belonging to a unique equivalence class for the equivalence relation $\mathfrak{M}$ over $\mathbb{N}_{\geqslant 2}$. To see that a given infinite ordered set $X = \{x_0 < x_1 < x_2 < \cdots\}$ is k-recognizable for *no* base $k \geqslant 2$ at all, we can use results like the following one, where the behavior of the ratio (*resp.* difference) of any two consecutive elements in X is studied through the quantities

$$\mathbf{R}_X = \limsup_{i\to\infty} \frac{x_{i+1}}{x_i} \text{ and } \mathbf{D}_X = \limsup_{i\to\infty} (x_{i+1} - x_i) .$$

Theorem 2.8 (Gap theorem [36]). *Let $k \geqslant 2$. If $X \subseteq \mathbb{N}$ is a k-recognizable infinite subset of $\mathbb{N}$, then either $\mathbf{R}_X > 1$ or $\mathbf{D}_X < +\infty$.*

Corollary 2.9. *Let $a \in \mathbb{N}_{\geqslant 2}$. The set of primes and the set $\{n^a \mid n \geqslant 0\}$ are never k-recognizable for any integer base $k \geqslant 2$.*

Proofs of the Gap theorem and its corollary can also be found in [51]. For more results on primes, see also the chapter "Automata in number theory" of this handbook.

Definition 2.5. An infinite ordered set $X = \{x_0 < x_1 < x_2 < \cdots\}$ such that $\mathbf{D}_X < +\infty$ is said to be *syndetic* or with *bounded gaps*: there exists $C > 0$ such that for all $n \geqslant 0$, $x_{n+1} - x_n < C$. In particular, any ultimately periodic set is syndetic. The converse does not hold, see for instance Example 3.1.

Remark 2.10. Note that syndeticity occurs in various contexts like in ergodic theory. As an example, a subset of an Abelian group G is said to be syndetic if finitely many translates of it cover G. The term "syndetic" was first quoted in [62]. Note that in [64] the following result is proved. Let $\alpha, \beta > 1$ be multiplicatively independent real numbers. If a set $X \subseteq \mathbb{N}$ is α-recognizable and β-recognizable, for the Bertrand numeration systems based respectively on the real numbers α and β in the sense of [14] and Theorem 2.2, then X is syndetic.

Cobham's original proof of Theorem 1.1 appeared in [35] and we quote [51] "*The proof is correct, long and hard. It is a challenge to find a more reasonable proof of this fine theorem*". Then G. Hansel proposed a simpler presentation in [63], also one can see [98] or the dedicated chapter in [9] for an expository presentation. Prior to these last two references, one should read [108]. Usually a first step to prove Cobham's theorem is to show the syndeticity of the considered set. See Section 5.3.

3 Automatic sequences

As explained in Corollary 3.3 presented in this section, the formalism of k-recognizable sets is equivalent to the one of k-automatic sequences[2]. Let us recall briefly what they are.

[2] We indifferently use the terms sequence and infinite word.

An infinite word $x = (x_n)_{n\geqslant 0} \in B^{\mathbb{N}}$ over an alphabet B is said to be k*-automatic* if there exists a DFAO (deterministic finite automaton with output) over the alphabet $[\![0, k-1]\!]$, $(Q, [\![0, k-1]\!], \cdot, q_0, B, \tau)$ such that, for all $n \geqslant 0$,

$$x_n = \tau(q_0 \cdot \mathrm{rep}_k(n)) \ .$$

The transition function is $\cdot : Q \times [\![0, k-1]\!] \to Q$ and can easily be extended to $Q \times [\![0, k-1]\!]^*$ by $q \cdot \varepsilon = q$ and $q \cdot wa = (q \cdot w) \cdot a$. The output function is $\tau : Q \to B$. Roughly speaking, the nth term of the sequence is obtained by feeding a DFAO with the k-ary representation of n. For a complete and comprehensive exposition on k-automatic sequences and their applications see the book [9]. We equally use the terms of sequences or (right-) infinite words. For more information about combinatorics on words, see [84, 85] or also J. Cassaigne and F. Nicolas' chapter in [12, Chapter 4].

Definition 3.1. Let $\sigma : A^* \to A^*$ be a morphism, *i.e.*, $\sigma(uv) = \sigma(u)\sigma(v)$ for all $u, v \in A^*$. Naturally such a map can be defined on A^ω. A finite or infinite word x such that $\sigma(x) = x$ is said to be a *fixed point* of σ. A morphism $\sigma : A^* \to A^*$ is completely determined by the images of the letters in A. In particular, if there exists $k \geqslant 0$ such that for all $a \in A$, $|\sigma(a)| = k$, then σ is said to be of k*-uniform* or simply *uniform*. A 1-uniform morphism is called a *coding*. If there exist a letter $a \in A$ and a word $u \in A^+$ such that $\sigma(a) = au$ and moreover, if $\lim_{n\to+\infty} |\sigma^n(a)| = +\infty$, then σ is said to be *prolongable* on a or to be a *substitution*. Let $\sigma : A^* \to A^*$ be a morphism prolongable on a. We have

$$\sigma(a) = a\,u,\ \sigma^2(a) = a\,u\,\sigma(u),\ \sigma^3(a) = a\,u\,\sigma(u)\,\sigma^2(u), \ldots\ .$$

Since for all $n \in \mathbb{N}$, $\sigma^n(a)$ is a prefix of $\sigma^{n+1}(a)$ and because $|\sigma^n(a)|$ tends to infinity when $n \to +\infty$, the sequence $(\sigma^n(a))_{n\geqslant 0}$ converges (for the usual product topology on words, see for instance (6.2)) to an infinite word denoted by $\sigma^\infty(a)$ and given by

$$\sigma^\infty(a) := \lim_{n\to+\infty} \sigma^n(a) = a\,u\,\sigma(u)\,\sigma^2(u)\,\sigma^3(u)\cdots\ .$$

This infinite word is a fixed point of σ. An infinite word obtained in this way by iterating a prolongable morphism is said to be *purely substitutive* (or *pure morphic*). If $\sigma : A^* \to B^*$ is a non-erasing morphism, it can be extended to a map from $A^{\mathbb{N}}$ to $B^{\mathbb{N}}$ as follows. If $x = x_0x_1\cdots$ is an infinite word over A, then the sequence of words $(\sigma(x_0\cdots x_{n-1}))_{n\geqslant 0}$ is easily seen to be convergent towards an infinite word over B. Its limit is denoted by $\sigma(x) = \sigma(x_0)\sigma(x_1)\sigma(x_2)\cdots$. If $x \in A^{\mathbb{N}}$ is purely substitutive and if $\tau : A \to B$ is a coding, then the word $y = \tau(x)$ is said to be *substitutive*.

Another result due to A. Cobham is the following one, see [36]. The idea is to canonically associated with any k-uniform morphism a DFA over $[\![0, k-1]\!]$.

Theorem 3.1. *Let $k \geqslant 2$. A sequence $x = (x_n)_{n\geqslant 0} \in B^{\mathbb{N}}$ is k-automatic if and only if there exists a k-uniform morphism $\sigma : A^* \to A^*$ prolongable on a letter $a \in A$ and a coding $\tau : A \to B$ such that $x = \tau(\sigma^\infty(a))$.*

Theorem 3.2 (Eilenberg [51]). *A sequence $x = (x_n)_{n\geqslant 0}$ is k-automatic if and only if its* k-kernel $N_k(x) = \{(x_{k^e n+d})_{n\geqslant 0} \mid e \geqslant 0,\ 0 \leqslant d < k^e\}$ *is finite.*

Definition 3.2. The *characteristic sequence* $\mathbb{1}_X \in \{0,1\}^{\mathbb{N}}$ of a set $X \subseteq \mathbb{N}$ is defined by $\mathbb{1}_X(n) = 1$ if and only if $n \in X$.

An infinite word $x \in A^\omega$ is *ultimately periodic* if there exist two finite words $u \in A^*$ and $v \in A^+$ such that $x = uv^\omega$. If $u = \varepsilon$, x is *periodic*. Obviously, a set $X \subseteq \mathbb{N}$ is ultimately periodic if and only if $\mathbb{1}_X$ is an ultimately periodic word over $\{0,1\}$. In that case, there exist two finite words $u \in \{0,1\}^*$ and $v \in \{0,1\}^+$ such that $\mathbb{1}_X = uv^\omega$. In particular, $|v|$ is a period of X. If u and v are chosen of minimal length, then $|u|$ (resp. $|v|$) is said to be the *preperiod* or *index* of X (resp. the *period* of X). If $u = \varepsilon$, X is *(purely) periodic*. Periodic sets are in particular ultimately periodic.

Corollary 3.3. *Let $k \geqslant 2$. If $x = (x_n)_{n\geqslant 0} \in B^{\mathbb{N}}$ is a k-automatic sequence then, for all $b \in B$, the set $\{n \geqslant 0 \mid x_n = b\}$ is k-recognizable. Conversely, if a set $X \subseteq \mathbb{N}$ is k-recognizable, then its characteristic sequence is k-automatic.*

Theorem 3.4 (Cobham's theorem, version 2). *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. An infinite word $x = (x_n)_{n\geqslant 0} \in B^{\mathbb{N}}$ is both k-automatic and ℓ-automatic if and only if it is ultimately periodic.*

Remark 3.5. Using the framework of k-automatic sequences instead of the formalism of k-recognizable sets turns out to be useful. For instance, consider the *complexity function* of an infinite word x which maps $n \in \mathbb{N}$ onto the number $p_x(n)$ of distinct factors of length n occurring in x. Morse–Hedlund's theorem states that x is ultimately periodic if and only if p_x is bounded by some constant. This result appeared first in [90]. Proofs can be found in classical textbooks like [9, 84].

It is also well known that for a k-automatic sequence x, $p_x \in \mathcal{O}(n)$, again see the seminal paper [36]. This latter result can be used to show that particular sets are not k-recognizable for any $k \geqslant 2$: for instance, those sets whose characteristic sequence $\mathbb{1}_X$ has a complexity function such that $\lim_{n\to+\infty} p_{\mathbb{1}_X}(n)/n = +\infty$. For the behavior of p_x in the substitutive case, see the survey [4] or [12, Chapter 4].

Example 3.1. Iterating the morphism $\sigma : 0 \mapsto 01, 1 \mapsto 10$, we get the *Thue–Morse word* $(t_n)_{n\geqslant 0} = \sigma^\infty(0) = 0110100110010110100101100110\cdots$. For an account on this celebrated word, see [8] and [56, Chapter 2]. It is a 2-automatic word, the nth letter in the word is 0 if and only if $\mathrm{rep}_2(n)$ contains an even number of 1's. This word is generated by the DFAO represented in Figure 2. In particular, the set

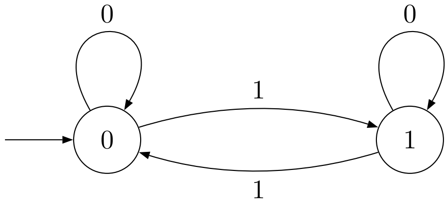


Figure 2. A DFAO generating the Thue–Morse word.

$$X_2 = \left\{ n \in \mathbb{N} \mid \mathrm{rep}_2(n) = c_t \cdots c_0 \text{ and } \sum_{i=0}^{t} c_i \equiv 0 \pmod 2 \right\}$$

is 2-recognizable. The Thue–Morse word is not ultimately periodic (see for instance [23] or [39] where the complexity function of this word is studied carefully) and therefore X_2 is k-recognizable only for those k of the form 2^m, $m \in \mathbb{N}_{\geqslant 1}$. Nevertheless, one can notice that X_2 is syndetic.

4 Multidimensional extension and first order logic

4.1 Subsets of $\mathbb{N}^d$

To extend the concept of k-recognizability to subsets of $\mathbb{N}^d$, $d \geqslant 2$, it is natural to consider d-tuples of k-ary representations. To get d words of the same length that have to be read simultaneously by an automaton, the shortest ones are padded with leading zeroes. We extend the definition of rep_k to a map of domain $\mathbb{N}^d$ as follows. If $n_1, \ldots, n_d$ are non-negative integers, we consider the word

$$\operatorname{rep}_k(n_1, \ldots, n_d) := \begin{pmatrix} 0^{m-|\operatorname{rep}_k(n_1)|}\operatorname{rep}_k(n_1) \\ \vdots \\ 0^{m-|\operatorname{rep}_k(n_d)|}\operatorname{rep}_k(n_d) \end{pmatrix} \in \left(\llbracket 0, k-1 \rrbracket^d\right)^*$$

where $m = \max\{|\operatorname{rep}_k(n_1)|, \ldots, |\operatorname{rep}_k(n_d)|\}$. A subset X of $\mathbb{N}^d$ is *k-recognizable* if the corresponding language $\operatorname{rep}_k(X)$ is accepted by a finite automaton over the alphabet $\llbracket 0, k-1 \rrbracket^d$ which is the Cartesian product of d copies of $\llbracket 0, k-1 \rrbracket$. This automaton is reading d digits at a time (one for each component): this is why we need d words of the same length.

Example 4.1. Consider the automaton depicted in Figure 3 (the sink is not represented). It accepts $(\varepsilon, \varepsilon)$ and all pairs of words of the kind $(u0, 0u)$ where $u \in 1\{0,1\}^*$. This means that the set $\{(2n, n) \mid n \geqslant 0\}$ is 2-recognizable.

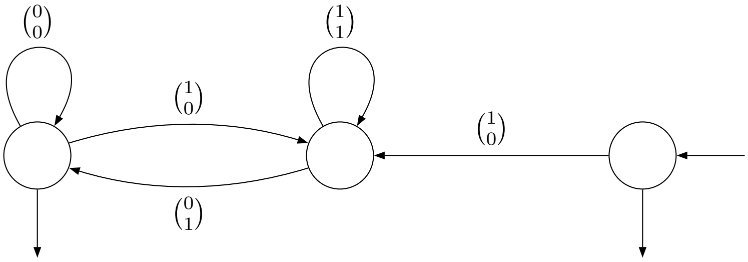


Figure 3. A DFA recognizing $\{(2n, n) \mid n \geqslant 0\}$.

Note that the notion of k-automatic sequence and Theorem 3.1 have been extended accordingly in [111, 112] where the images by a morphism of letters are d-dimensional cubes of size k.

Extending the concept of ultimately periodic sets to subsets of $\mathbb{N}^d$, with $d \geqslant 2$, is at first glance not so easy. We use bold face letters to represent elements in $\mathbb{N}^d$. For instance, one could take the following definition of a (purely) periodic subset $X \subseteq \mathbb{N}^d$. There exists a non-zero element $\mathbf{p} \in \mathbb{N}^d$ such that $\mathbf{x} \in X$ if and only if $\mathbf{x} + \mathbf{p} \in X$. As we will see (Remark 4.2, Proposition 6.9 and Theorem 6.11), it turns out that this definition does not fit to the extension of Cobham's theorem in d dimensions. Therefore we will consider sets definable in $\langle \mathbb{N}, + \rangle$. Let us mention *Nivat's conjecture* connecting such a notion of periodicity in higher dimensions with the notion of block complexity as introduced in Remark 3.5: let $X \subset \mathbb{Z}^2$, if there exist positive integers n_1, n_2 such that $p_X(n_1, n_2) \leqslant n_1 n_2$ then X is periodic, where $p_X(n_1, n_2)$ counts the number of distinct blocks of size $n_1 \times n_2$ occurring in X. See [92] and in particular [102] for details and pointers to the existing bibliography.

4.2 Logic and k-definable sets

The formalism of first order logic is probably the best suited to present a natural extension (in the sense of Cobham's theorem) of the definition of ultimately periodic sets in d dimensions. See [100, 101] or the survey [16]. In the *Presburger arithmetic* $\langle \mathbb{N}, + \rangle$, the variables range over $\mathbb{N}$ and we have at our disposal the connectors $\wedge, \vee, \neg, \rightarrow, \leftrightarrow$, the equality symbol $=$ and the quantifiers $\forall$ and $\exists$ that can only be applied to variables. This is the reason we speak of first order logic; in second order logic, quantifiers can be applied to relations, and in monadic second order logic, only variables and unary relations, *i.e.,* sets, may be quantified. If a variable is not within the scope of any quantifier, this variable is said to be *free*. Formulas are build inductively from terms and atomic formulas. Here details have been omitted, see for instance [28, Section 3.1]. For instance, order relations $<, \leqslant, \geqslant$ and $>$ can be added to the language by noticing that $x \leqslant y$ is equivalent to

$$(\exists z)(y = x + z). \tag{4.1}$$

In the same way, constants can also be added. For instance, $x = 0$ is equivalent to $(\forall y)(x \leqslant y)$ and $x = 1$ is equivalent to $\neg(x = 0) \wedge (\forall y)(\neg(y = 0) \rightarrow (x \leqslant y))$. In general, the *successor* function $\mathbf{S}(x) = y$ of x is defined by

$$(x < y) \wedge (\forall z)((x < z) \rightarrow (y \leqslant z)) \, .$$

For a complete account on the interactions between first order logic and k-recognizable sets, see the excellent survey [28].

Remark 4.1. We mainly discuss the case $\langle \mathbb{N}, + \rangle$ but all developments can be made for $\langle \mathbb{Z}, +, \leqslant \rangle$. Note that if the variables belong to $\mathbb{Z}$ then it is no longer possible to define $\leqslant$ as in (4.1). So this order relation has to be added to the structure. The constant 0 can be defined by $x + x = x$.

Let $\varphi(x_1, \dots, x_d)$ be a formula with d free variables $x_1, \dots, x_d$. Interpreting φ in $\langle \mathbb{N}, + \rangle$ permits one to define the set of d-tuples of non-negative integers for which the formula holds true:

$$\{(r_1, \dots, r_d) \mid \langle \mathbb{N}, + \rangle \models \varphi[r_1, \dots, r_d]\}.$$

We write $\langle \mathbb{N}, + \rangle \models \varphi[r_1, \dots, r_d]$ if $\varphi(x_1, \dots, x_d)$ is satisfied in $\langle \mathbb{N}, + \rangle$ when interpreting

x_i by r_i for all $i \in \{1, \ldots, d\}$. For the reader having no background in logic and model theory, the first chapters of [50] are worth reading.

Remark 4.2. The ultimately periodic sets of $\mathbb{N}$ are exactly the sets that are definable in the Presburger arithmetic. It is obvious that ultimately periodic sets of $\mathbb{N}$ are definable. For instance, the set of even integers can be defined by $\varphi(x) \equiv (\exists y)(x = y + y)$. Since constants can easily be defined, it is easy to write a formula for any arithmetic progression. As an example, the formula $\varphi(x) \equiv (\exists y)(x = \mathbf{S}(\mathbf{S}(y + y + y)))$ defines the progression $3\mathbb{N} + 2$. In particular, multiplication by a fixed constant is definable in $\langle \mathbb{N}, + \rangle$. Note that it is a classical result that the theory of $\langle \mathbb{N}, +, \times \rangle$ is undecidable, see for instance [15].

Adding congruences modulo any integer m permits quantifier elimination, which means that any formula expressed in the Presburger arithmetic is equivalent to a formula using only $\wedge$, $\vee$, $=$, $<$ and congruences, see [100, 101]. Presentations can also be found in [52, 80].

Theorem 4.3 (Presburger). *The structure $\langle \mathbb{N}, +, <, (\equiv_m)_{m>0} \rangle$ admits elimination of quantifiers.*

This result can be used to prove that the theory of $\langle \mathbb{N}, + \rangle$ is decidable. This can be done using the formalism of automata, see for instance [28].

Corollary 4.4. *Any formula $\varphi(x)$ in the Presburger arithmetic $\langle \mathbb{N}, + \rangle$ defines an ultimately periodic set of $\mathbb{N}$.*

Let $k \geqslant 2$. We add to the structure $\langle \mathbb{N}, + \rangle$ a function V_k defined by $V_k(0) = 1$ and for all $x > 0$, $V_k(x)$ is the greatest power of k dividing x. As an example, we have $V_2(6) = 2$, $V_2(20) = 4$ and $V_2(2^n) = 2^n$ for all $n \geqslant 0$. Again the theory of $\langle \mathbb{N}, +, V_k \rangle$ can be shown to be decidable [28]. The next result shows that, as for the k-automatic sequences, the logical framework within the richer structure $\langle \mathbb{N}, +, V_k \rangle$ gives an equivalent presentation of the k-recognizable sets in any dimension. Proofs of the next three theorems can again be found in [28] where a full account of the different approaches used to prove Theorem 4.5 is presented. For Büchi's original paper, see [29].

Theorem 4.5 (Büchi's theorem). *Let $k \geqslant 2$ and $d \geqslant 1$. A set $X \subseteq \mathbb{N}^d$ is k-recognizable if and only if it can be defined by a first order formula $\varphi(x_1, \ldots, x_d)$ of $\langle \mathbb{N}, +, V_k \rangle$.*

For instance the set P_2 introduced in Example 2.2 can be defined by the formula $\varphi(x) \equiv V_2(x) = x$. Note that Theorem 4.5 holds for Pisot numeration systems given in Definition 2.3, see [27] where the function V_k is modified accordingly. This is partially based on the fact that in a Pisot numeration system the normalization function is realized by a finite automaton, see [58], which allows one to consider addition of integers: first perform addition digit-wise without any carry, then normalize the result.

Theorem 4.6 (Cobham's theorem, version 3). *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. A set $X \subseteq \mathbb{N}$ can be defined by a first order formula in $\langle \mathbb{N}, +, V_k \rangle$ and by a first order formula in $\langle \mathbb{N}, +, V_\ell \rangle$ if and only if it can be defined by a first order formula in $\langle \mathbb{N}, + \rangle$.*

This theorem still holds in higher dimensions and is called the Cobham–Semenov theorem. In this respect, the notion of subset of $\mathbb{N}^d$ definable in the Presburger arithmetic $\langle \mathbb{N}, + \rangle$ is the right extension of periodicity in a multidimensional setting. For Semenov's original paper, see [113].

Theorem 4.7 (Cobham–Semenov theorem). *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. A set $X \subseteq \mathbb{N}^d$ can be defined by a first order formula in $\langle \mathbb{N}, +, V_k \rangle$ and by a first order formula in $\langle \mathbb{N}, +, V_\ell \rangle$ if and only if it can be defined by a first order formula in $\langle \mathbb{N}, + \rangle$.*

Subsets of $\mathbb{N}^d$ defined by a first order formula in $\langle \mathbb{N}, + \rangle$ are characterized in [61]. The nice criterion of Muchnik appeared first in 1991 and is given in [91]. See Proposition 6.9 for its precise statement. Using this latter characterization, a proof of Theorem 4.7 is presented in [28]. The logical framework has given rise to several works. Let us mention chronologically [118, 119] and [88, 89]. In [89, Section 5] the authors interestingly show how to reduce Semenov's theorem to Cobham's theorem: "Nothing new in higher dimensions". Also extensions to non-standard numeration systems are considered in [99] and [15]. In this latter paper, the Cobham–Semenov theorem is proved for two Pisot numeration systems.

5 Numeration systems and substitutions

5.1 Substitutive sets and abstract numeration systems

In Sections 4.1 and 4.2, we have mainly extended the notion of recognizability to subsets of $\mathbb{N}^d$. Now we consider another extension of recognizability. In Corollary 3.3, we have seen that a k-recognizable set has a characteristic sequence generated by a uniform substitution and the application of an extra coding. It is pretty easy to define sets of integers encoded by a characteristic sequence generated by an arbitrary substitution and an extra coding, that is whose characteristic sequence is morphic. This generalization permits one to reach a larger class of infinite words, hence a larger class of sets of integers.

Example 5.1. Consider the morphism $\sigma : \{a, b, c\}^* \to \{a, b, c\}^*$ given by $\sigma(a) = abcc$, $\sigma(b) = bcc$, $\sigma(c) = c$ and the coding $\tau : a, b \mapsto 1, c \mapsto 0$. We get

$$\sigma^\infty(a) = abccbccccbccccccbccccccccbccccccccccbcc \cdots$$

and $\tau(\sigma^\infty(a)) = 0100100001000000100000000100000000001 00 \cdots$. Using the special form of the images by σ of b and c, it is not difficult to see that the difference between the position of the nth b and the $(n+1)$st b in $\sigma^\infty(a)$ is $2n+1$. Hence $\tau(\sigma^\infty(a))$ is the characteristic sequence of the set of squares and it is substitutive. From Corollary 2.9 the set of squares is never k-recognizable for any integer base k.

Definition 5.1. As a natural extension of the concept of recognizability, we may consider sets $X \subseteq \mathbb{N}$ having a characteristic sequence $\mathbb{1}_X$ which is (purely) substitutive. Such a set is said to be a *(purely) substitutive set*. In particular k-recognizable sets are substitutive.

With Theorem 5.2 it will turn out that the formalism of substitutive sets is equivalent to the one of abstract numeration systems.

Definition 5.2. [81] An *abstract numeration system* or *ANS* is a triple $\mathcal{S} = (L, A, <)$ where L is an infinite regular language over a totally ordered alphabet $(A, <)$. The map $\mathrm{rep}_{\mathcal{S}} : \mathbb{N} \to L$ is the one-to-one correspondence mapping $n \in \mathbb{N}$ onto the $(n+1)$th word in the genealogically ordered language L, which is called the *$\mathcal{S}$-representation* of n. In particular, a set $X \subseteq \mathbb{N}$ is *$\mathcal{S}$-recognizable*, if $\mathrm{rep}_{\mathcal{S}}(X)$ is regular, and, $\mathbb{N}$ is trivially $\mathcal{S}$-recognizable because $\mathrm{rep}_{\mathcal{S}}(\mathbb{N}) = L$. Recall that in the *genealogical order* (also called *radix* or *military* order), words are first ordered by increasing length and for words of the same length, one uses the lexicographic ordering induced by the order $<$ on A.

Example 5.2. Consider the language $L = a^*b^* \cup a^*c^*$ with $a < b < c$. The first words in L are $\varepsilon, a, b, c, aa, ab, ac, bb, cc, aaa, aab, aac, abb, \ldots$. This means that for the ANS $\mathcal{S}$ built on L, 0 is represented by ε, 1 by a, 2 by b, 3 by c, 4 by aa, *etc*. Since L contains exactly $2n+1$ words of length n for all $n \geqslant 0$, we have that n^2 is represented by a^n for all $n \geqslant 0$. In particular, the set $\{n^2 \mid n \geqslant 0\}$ is $\mathcal{S}$-recognizable because a^* is regular. It is well known that in a regular language L, the set of the first words of each length in the genealogically ordered language L is regular, see [115].

Pisot numeration systems are special cases of ANS. Indeed, if the numeration basis $U = (U_n)_{n \geqslant 0}$ defines a Pisot numeration system, then $\mathrm{rep}_U(\mathbb{N})$ is regular.

Example 5.3. Consider the Fibonacci sequence and the language $L = 1\{0, 01\}^* \cup \{\varepsilon\}$ defined in Example 2.1. To get the representation of an integer n, one can either decompose n using the greedy algorithm or, order genealogically the words in L and take the $(n+1)$th element.

Theorem 5.1. [81] *Let $\mathcal{S} = (L, A, <)$ be an abstract numeration system. Any ultimately periodic set is $\mathcal{S}$-recognizable.*

Note that in [78], it is in particular proved that this latter result cannot be extended to context-free languages. Specific cases of $\mathcal{S}$-recognizable sets are discussed in P. Lecomte and M. Rigo's chapter in [12, Chapter 3]. We have an extension of Theorem 3.1.

Theorem 5.2. *Let $x = (x_n)_{n \geqslant 0}$ be an infinite word over an alphabet B. This word is substitutive if and only if there exists an abstract numeration system $\mathcal{S} = (L, A, <)$ such that x is $\mathcal{S}$-automatic,* i.e.*, there exists a DFAO $(Q, A, \cdot, \{q_0\}, B, \tau)$ such that for all $n \geqslant 0$, $x_n = \tau(q_0 \cdot \mathrm{rep}_{\mathcal{S}}(n))$.*

A proof of this result is given in [105, 107] and a comprehensive treatment is given in [12, Chapter 3]. In that context, we also obtain an extension of Corollary 3.3.

Corollary 5.3. *Let $x = (x_n)_{n \geqslant 0}$ be an infinite substitutive word over an alphabet B. There exists an ANS $\mathcal{S}$ such that for all $b \in B$, $\{n \geqslant 0 \mid x_n = b\}$ is $\mathcal{S}$-recognizable. Conversely, if a set $X \subseteq \mathbb{N}$ is $\mathcal{S}$-recognizable, then its characteristic sequence is $\mathcal{S}$-automatic.*

Corollary 5.4. *A set $X \subseteq \mathbb{N}$ is substitutive if and only if there exists an ANS $\mathcal{S}$ such that X is $\mathcal{S}$-recognizable.*

5.2 Cobham's theorem for substitutive sets

In the context of substitutive sets of integers, *how could a Cobham-like theorem be expressed, i.e., what is playing the role of a base*? Assume that there exist two purely substitutive infinite words $x \in A^\omega$ and $y \in B^\omega$ respectively generated by the morphisms $\sigma : A^* \to A^*$ prolongable on $a \in A$ and $\tau : B^* \to B^*$ prolongable on $b \in B$, *i.e.*, $\sigma^\infty(a) = x$ and $\tau^\infty(b) = y$. Consider two codings $\lambda : A \to \{0,1\}$ and $\mu : B \to \{0,1\}$ such that $\lambda(x) = \mu(y)$. This situation corresponds to the case where a set (here, given by its characteristic word) is recognizable in two *a priori* different numeration systems.

If $A = B$ and $\tau = \sigma^m$ for some $m \geqslant 1$, nothing particular can be said about the infinite word $\lambda(x)$: iterating σ or σ^m from the same prolongable letter leads to the same fixed point. So we must introduce a notion analogous to the one of multiplicatively independent bases related to the substitutions σ and λ.

Definition 5.3. Let $\sigma : A^* \to A^*$ be a substitution over an alphabet A. The matrix $\mathbf{M}_\sigma \in \mathbb{N}^{A\times A}$ associated with σ is called the *incidence matrix* of σ and is defined by

$$\text{for all } a, b \in A,\ (\mathbf{M}_\sigma)_{a,b} = |\sigma(b)|_a \ .$$

A square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with entries in $\mathbb{R}_{\geqslant 0}$ is *irreducible* if, for all i, j, there exists k such that $(\mathbf{M}^k)_{i,j} > 0$. A square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with entries in $\mathbb{R}_{\geqslant 0}$ is *primitive* if there exists k such that, for all i, j, we have $(\mathbf{M}^k)_{i,j} > 0$. Similarly, a substitution over the alphabet A is *irreducible* (resp. *primitive*) if its incidence matrix is irreducible (resp. primitive). Otherwise stated, a substitution $\sigma : A^* \to A^*$ is primitive, if there exists an integer $n \geqslant 1$ such that, for all $a \in A$, all the letters in A appear in the image of $\sigma^n(a)$.

Let us denote by $\mathbf{P}$ the abelianisation map (or Parikh map) which maps a word w over $A = \{a_1, \ldots, a_r\}$ on the r-tuple ${}^t(|w|_{a_1}, \ldots, |w|_{a_r})$. The matrix $\mathbf{M}_\sigma$ can be defined by its columns:

$$\mathbf{M}_\sigma = \begin{pmatrix}\mathbf{P}(\sigma(a_1)) & \cdots & \mathbf{P}(\sigma(a_r))\end{pmatrix},$$

and it satisfies:

$$\text{for all } w \in A^*,\ \mathbf{P}(\sigma(w)) = \mathbf{M}_\sigma \mathbf{P}(w) \ .$$

Remark 5.5. If a matrix $\mathbf{M}$ is primitive, the celebrated theorem of Perron can be used, see standard textbooks like [72] or [60, 114]. A presentation is also given in [83]. To recap some of the key points, $\mathbf{M}$ has a unique dominating real eigenvalue $\beta > 0$ and there exists an eigenvector with positive entries associated with β. Also, for all i, j, there exists $c_{i,j}$ such that $(\mathbf{M}^n)_{i,j} = c_{i,j}\beta^n + o(\beta^n)$. For instance, primitiveness of $\mathbf{M}_\sigma$ implies the existence of the frequency of any factor occurring in any fixed point of σ. Note that

$$\text{if } \mathbf{P}(w) = {}^t(p_1, \ldots, p_r), \text{ then } |w| = \sum_{i=1}^{r} p_i. \tag{5.1}$$

Hence, for all $n \geqslant 0$, $|\sigma^n(a_j)|$ is obtained by summing up the entries in the jth column of $\mathbf{M}_\sigma^n$. If σ is primitive then there exists some C_j such that $|\sigma^n(a_j)| = C_j\beta^n + o(\beta^n)$. In particular, if σ is prolongable on a, then $|\sigma^n(a)| \sim C\beta^n$, for some $C > 0$.

In the general case of a matrix $\mathbf{M}$ with non-negative entries, one can use the Perron–Frobenius theorem for each of the irreducible components of $\mathbf{M}$ (they correspond to the strongly connected components of the associated graph, also called communicating classes). Thus any non-negative matrix $\mathbf{M}$ has a real eigenvalue α which is greater or equal to the modulus of any other eigenvalue. We call α the *dominating eigenvalue* of $\mathbf{M}$. Moreover, if we exclude the case where $\alpha = 1$, then there exists a positive integer p such that $\mathbf{M}^p$ has a dominating eigenvalue α^p which is a Perron number, see [83, p. 369]. A *Perron number* is an algebraic integer $\alpha > 1$ such that all its algebraic conjugates have modulus less than α. In particular, if we replace a prolongable substitution σ such that $\mathbf{M}_\sigma$ has a dominating eigenvalue $\alpha > 1$, with a convenient power σ^p of σ, we can assume that the dominating eigenvalue of σ is a Perron number.

Definition 5.4. Let $\sigma : A^* \to A^*$ be a substitution prolongable on $a \in A$ such that *all* letters of A have an occurrence in $\sigma^\infty(a)$. Let $\alpha > 1$ be the dominating eigenvalue of the incidence matrix of σ. Let $\phi : A \to B^*$ be a coding. We say $\phi(\sigma^\infty(a))$ is an *α-substitutive infinite word (with respect to σ)*. In view of Definition 5.1, this notion can be applied to subsets of $\mathbb{N}$. If moreover σ is primitive, then $\phi(\sigma^\infty(a))$ is said to be a *primitive α-substitutive infinite word (w.r.t. σ)*.

Observe that k-automatic infinite words are k-substitutive infinite words.

Example 5.4. Consider the substitution σ defined by $\sigma(a) = aa0a$, $\sigma(0) = 01$ and $\sigma(1) = 10$. Its dominating eigenvalue is 3. It is prolongable both on a, 0 and 1. The fixed point x of σ starting with 0 is the Thue-Morse sequence (see Example 3.1). Definition 5.4 does not implies that x is 3-substitutive because a does not appear in x. But the fixed point y of σ starting with a is 3-substitutive.

Example 5.5. Consider the so-called *Tribonacci word*, which is the unique fixed point of $\sigma : a \mapsto ab, b \mapsto ac, c \mapsto a$. See [117, 56]. The incidence matrix of σ is

$$\mathbf{M}_\sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

One can check that $\mathbf{M}_\sigma^3$ contains only positive entries. So the matrix is primitive. Let $\alpha_T \simeq 1.839$ be the unique real root of the characteristic polynomial $-X^3 + X^2 + X + 1$ of $\mathbf{M}_\sigma$. The Tribonacci word $T = abacabaab\cdots$ is primitive α_T-substitutive. Let $\tau : a \mapsto 1, b, c \mapsto 0$ be a coding. The word $\tau(T)$ is the characteristic sequence of a primitive α_T-substitutive set of integers $\{0, 2, 4, 6, 7, \ldots\}$.

To explain the substitutive extension of Cobham's theorem we need the following definition.

Definition 5.5. Let $\mathcal{S}$ be a set of prolongable substitutions and x be an infinite word. If x is an α-substitutive infinite word w.r.t. a substitution σ belonging to $\mathcal{S}$, then x is said to be *α-substitutive with respect to $\mathcal{S}$*.

Let us consider the following Cobham-like statement depending on two sets $\mathcal{S}$ and $\mathcal{S}'$ of prolongable substitutions. It is useful to describe chronologically known results generalizing Cobham's theorem in terms of substitutions leading to the most general statement for all substitutions.

Statement $(\mathcal{S}, \mathcal{S}')$. Let $\mathcal{S}$ and $\mathcal{S}'$ be two sets of prolongable substitutions. Let α and β be two multiplicatively independent Perron numbers. Let $x \in A^\omega$ where A is a finite alphabet. Then the following are equivalent:

(1) the infinite word x is both α-substitutive w.r.t. $\mathcal{S}$ and β-substitutive w.r.t. $\mathcal{S}'$;

(2) the infinite word x is ultimately periodic.

Note that this statement excludes 1-substitutions, *i.e.*, substitutions with a dominating eigenvalue equal to 1, because Perron numbers are larger than 1. The case of 1-substitutive infinite words will be mentioned in Subsection 5.6. Also notice that the substitutions we are dealing with can be *erasing*, *i.e.*, at least one letter is sent onto the empty word. But from a result in [34, 9, 71], we can assume that the substitutions are non-erasing. Note that α and α^k are multiplicatively dependent.

Proposition 5.6. [49] *Let x be an α-substitutive infinite word. Then, there exists an integer $k \geqslant 1$ such that x is α^k-substitutive with respect to a non-erasing substitution.*

The implication (2) $\Rightarrow$ (1) in the above general statement is not difficult to obtain as mentioned in Remark 1.3 for the uniform situation.

Proposition 5.7. [47] *Let x be an infinite word over a finite alphabet and α be a Perron number. If x is periodic (resp. ultimately periodic) then x is primitive α-substitutive (resp. α-substitutive).*

Definition 5.6. Let $\sigma : A^* \to A^*$ and $\tau : B^* \to B^*$ be two substitutions. We say that σ *projects* on τ if there exists a coding $\phi : A \to B$ such that

$$\phi \circ \sigma = \tau \circ \phi \, . \tag{5.2}$$

The implication (1) $\Rightarrow$ (2) in Statement $(\mathcal{S}, \mathcal{S}')$ is known in many cases described below:

(i) When $\mathcal{S} = \mathcal{S}'$ is the set of *uniform* substitutions, this is the classical theorem of Cobham.

(ii) In [53] S. Fabre proves the statement when $\mathcal{S}$ is the set of uniform substitutions and $\mathcal{S}'$ is a set of non-uniform substitutions related to some non-standard numeration systems.

(iii) When $\mathcal{S} = \mathcal{S}'$ is the set of *primitive substitutions*, the statement is proved in [44]. The proof is based on a characterization of primitive substitutive sequences using the notion of return word [43].

(iv) When $\mathcal{S} = \mathcal{S}'$ is the set of *substitutions projecting on primitive substitutions*, the statement is proved in [45]. This result is applied to generalize (ii). Using a characterization of U-recognizable sets of integers for a Bertrand numeration basis U

[54], the main result of [45] extends Cobham's theorem for some large family of non-standard numeration systems. This latter result includes a result obtained previously in [15] for Pisot numeration systems.

(v) Definition 5.8 and Theorem 5.17 describe the situation where $\mathcal{S} = \mathcal{S}' = \mathcal{S}_{\text{good}}$ (defined later). It includes all known and previously described situations for substitutions.

(vi) In [42], Statement $(\mathcal{S}, \mathcal{S}')$ is proven for the most general case that is $\mathcal{S}$ and $\mathcal{S}'$ are both the set of all substitutions. The final argument is based on a fine study of return words for non-primitive substitutive sequences.

Example 5.6. The Tribonacci word T is purely substitutive but is k-automatic for no integer $k \geqslant 2$. Proceed by contradiction. Assume that there exists an integer $k \geqslant 2$ such that T is k-automatic. Then T is both k-substitutive and primitive α_T-substitutive. By Theorem 5.17, T must be ultimately periodic but it is not the case. The factor complexity of T is $p_T(n) = 2n + 1$. By the Morse–Hedlund theorem, see Remark 3.5, T is not ultimately periodic.

Let $L(x)$ be the set of all factors of the infinite word x. In [55], the following generalization of Cobham's theorem is proved.

Theorem 5.8. *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. Let x be a k-automatic infinite word and y be a ℓ-automatic infinite word. If $L(x) \subset L(y)$, then x is ultimately periodic.*

The same result is valid in the primitive case.

Theorem 5.9. [44] *Let x and y be respectively a primitive α-substitutive infinite word and a primitive β-substitutive infinite word such that $L(x) = L(y)$. If α and β are multiplicatively independent, then x and y are periodic.*

Note that under the hypothesis of Theorem 5.9 x and y are primitive substitutive infinite words. Thus $L(x) = L(y)$ whenever $L(x) \subset L(y)$. Observe that if y is the fixed point starting with a and x the fixed point starting with 0 of the substitution σ defined in Example 5.4, then $L(x) \subset L(y)$ but x is not ultimately periodic.

In Sections 5.3 and 5.4 we give the main arguments to prove Statement $(\mathcal{S}_{\text{good}}, \mathcal{S}_{\text{good}})$.

5.3 Density, syndeticity and bounded gaps

The proofs of most of the generalizations of Cobham's theorem are divided into two parts.

(i) Dealing with a subset X of integers, we have to prove that X is syndetic. Equivalently, dealing with an infinite word x, we have to prove that the letters occurring infinitely many times in x appear with bounded gaps.

(ii) In the second part of the proof, the ultimate periodicity of X or x has to be carried out.

This section is devoted to the description of the main arguments that lead to the complete treatment of (i).

In the original proof of Cobham one of the main arguments is that as k and ℓ are multiplicatively independent (we refer to Theorem 1.1) the set $\{k^n/\ell^m \mid n,m \in \mathbb{N}\}$ is dense in $[0,+\infty)$. In the uniform case these powers refer to the length of the iterates of the substitutions. Indeed, suppose $\sigma : A^* \to A^*$ is a k-uniform substitution. Then for any $a \in A$ we have $|\sigma^n(a)| = k^n$. Unsurprisingly, to be able to treat the non-uniform case, it is important to know that the set

$$\left\{ \frac{\mid \sigma^n(a) \mid}{\mid \tau^m(b) \mid} \mid n,m \in \mathbb{N} \right\}$$

is dense in $[0,+\infty)$, for some $a,b \in A$. We explain below that $|\sigma^n(a)|$ and $|\tau^m(b)|$ are governed by the dominating eigenvalue of their incidence matrices. First we focus on part (i) and consider infinite words.

5.3.1 The length of the iterates

The length of the iterates are described in the following lemma. Note that it includes erasing substitutions and substitutions with a dominating eigenvalue equal to 1. Observe that for the substitution σ defined by $0 \mapsto 001$ and $1 \mapsto 11$ we have $|\sigma^n(0)| = (n+2)2^{n-1}$ and $|\sigma^n(1)| = 2^n$ showing that the situation is different from the uniform case. It can easily be described using the Jordan normal form of the incidence matrix $\mathbf{M}_\sigma$. Discussion of the following result can be found in [12, Section 4.7.3].

Lemma 5.10 (Chapter III.7 in [110]). *Let $\sigma : A \to A^*$ be a substitution. For all $a \in A$ one of the two following situations occur*

(1) there exists $N \in \mathbb{N}$ such that for all $n > N$, $|\sigma^n(a)| = 0$, or,

(2) there exist $d(a) \in \mathbb{N}$ and real numbers $c(a), \theta(a)$ such that

$$\lim_{n\to+\infty} \frac{|\sigma^n(a)|}{c(a)\, n^{d(a)}\, \theta(a)^n} = 1.$$

Moreover, in the situation (2), for all $i \in \{0,\dots,d(a)\}$ there exists a letter $b \in A$ appearing in $\sigma^j(a)$ for some $j \in \mathbb{N}$ and such that

$$\lim_{n\to+\infty} \frac{|\sigma^n(b)|}{c(b)\, n^{i}\, \theta(a)^n} = 1.$$

Definition 5.7. Let σ be a non-erasing substitution. For all $a \in A$, the pair $(d(a),\theta(a))$ defined in Lemma 5.10 is called the *growth type* of a. If (d,θ) and (e,β) are two growth types, we say that (d,θ) is *less than* (e,β) (or $(d,\theta) < (e,\beta)$) whenever $\theta < \beta$ or, $\theta = \beta$ and $d < e$.

Consequently if the growth type of $a \in A$ is less than the growth type of $b \in A$ then $\lim_{n\to+\infty} |\sigma^n(a)|/|\sigma^n(b)| = 0$. We say that $a \in A$ is a *growing letter* if $(d(a),\theta(a)) > (0,1)$ or equivalently, if $\lim_{n\to+\infty} |\sigma^n(a)| = +\infty$.

We set $\Theta := \max\{\theta(a) \mid a \in A\}$, $D := \max\{d(a) \mid \forall a \in A : \theta(a) = \Theta\}$ and $A_{max} := \{a \in A \mid \theta(a) = \Theta, d(a) = D\}$. The dominating eigenvalue of $\mathbf{M}_\sigma$ is Θ. We say that the letters of A_{max} are *of maximal growth* and that (D, Θ) is the *growth type* of σ. Consequently, we say that a substitutive infinite word y is (D, Θ)*-substitutive* if the underlying substitution is of growth type (D, Θ). Observe that, due to Lemma 5.10, any substitutive sequence is (D, Θ)-substitutive for some pair (D, Θ).

Observe that if $\Theta = 1$, then in view of the last part of Lemma 5.10, there exists at least one non-growing letter of growth type $(0, 1)$. Otherwise stated, if a letter has polynomial growth, then there exists at least one non-growing letter. Consequently σ is *growing* (*i.e.*, all its letters are growing) if and only if $\theta(a) > 1$ for all $a \in A$. We define

$$\lambda_\sigma : A^* \to \mathbb{R},\ u_0 \cdots u_{n-1} \mapsto \sum_{i=0}^{n-1} c(u_i)\, \mathbb{1}_{A_{max}}(u_i)\ ,$$

where $c : A \to \mathbb{R}_+$ is defined in Lemma 5.10. From Lemma 5.10 we deduce the following lemma.

Lemma 5.11. *For all $u \in A^*$, we have $\lim_{n\to+\infty} |\sigma^n(u)|/n^D\Theta^n = \lambda_\sigma(u)$.*

We say that the word $u \in A^*$ is of *maximal growth* if $\lambda_\sigma(u) \neq 0$.

Corollary 5.12. *Let σ be a substitution of growth type (D, Θ). For all $k \geqslant 1$, the growth type of σ^k is (D, Θ^k).*

5.3.2 Letters and words appear with bounded gaps Recall that the first step for Cobham's theorem is to prove that the letters occurring infinitely many times appear with bounded gaps. In our context, this implies the same property for words. Moreover, we can relax the multiplicative independence hypothesis in order to include 1-substitutions. Note that 1 and $\alpha > 1$ are multiplicatively dependent.

Theorem 5.13. [49] *Let $d, e \in \mathbb{N} \setminus \{0\}$ and $\alpha, \beta \in [1, +\infty)$ such that $(d, \alpha) \neq (e, \beta)$ and satisfying one of the following three conditions:*

(i) *α and β are multiplicatively independent;*
(ii) *$\alpha, \beta > 1$ and $d \neq e$;*
(iii) *$(\alpha, \beta) \neq (1, 1)$ and, $\beta = 1$ and $e \neq 0$, or, $\alpha = 1$ and $d \neq 0$.*

Let C be a finite alphabet. If $x \in C^\omega$ is both (d, α)-substitutive and (e, β)-substitutive then the words occurring infinitely many times in x appear with bounded gaps.

The main argument used to prove this in [49] is the following.

Theorem 5.14. *Let $d, e \in \mathbb{N}$ and $\alpha, \beta \in [1, +\infty)$. The set*

$$\Omega = \left\{ \frac{\alpha^n n^d}{\beta^m m^e} \mid n, m \in \mathbb{N} \right\}$$

is dense in $[0, +\infty)$ if and only if one of the following three conditions holds:

(i) *α and β are multiplicatively independent;*
(ii) *$\alpha, \beta > 1$ and $d \neq e$;*
(iii) *$\beta = 1$ and $e \neq 0$, or, $\alpha = 1$ and $d \neq 0$.*

Sketch of the proof of Theorem 5.13. We only consider the case where α and β are multiplicatively independent.

Let $\sigma : A^* \to A^*$ be a substitution prolongable on a letter a' having growth type (d, α). Let $\tau : B^* \to B^*$ be a substitution prolongable on a letter b' having growth type (e, β). Let $\phi : A \to C$ and $\psi : B \to C$ be two codings such that $\phi(\sigma^\infty(a')) = \psi(\tau^\infty(b')) = x$. Using Proposition 5.6 we may assume that σ and τ are non-erasing. Suppose there is a letter a having infinitely many occurrences in x but that appears with unbounded gaps. Then the letters in $\phi^{-1}(\{a\})$ appear with unbounded gaps. To avoid extra technicalities (a complete treatment is considered in [49]), we assume that there is a letter in $\phi^{-1}(\{a\})$ having maximal growth. Then, it is quite easy to construct, for all $n \in \mathbb{N}$, a word w_n of length $c_1 n^d \alpha^n$, appearing in y at the index $c_2 n^d \alpha^n$, that does not contain any letter of $\phi^{-1}(\{a\})$. On the other hand, using a kind of pumping lemma for substitutions, one can show that there is a letter of $\psi^{-1}(\{a\})$ in z at the index $c_3 n^e \beta^n$. Therefore, using Theorem 5.14, the letter a appears in a word $\phi(w_n)$ for some n. This is not possible.

Now let us explain how to extend this result for a single letter to words. It uses what is called in [103] the *substitutions of the words of length n*. Let u be a word of length n occurring infinitely often in x. To prove that u appears with bounded gaps in x, it suffices to prove that the letter 1 appears with bounded gaps in the infinite word $t \in \{0,1\}^{\mathbb{N}}$ defined by

$$t_i = \begin{cases} 1, & \text{if } x_i \cdots x_{i+n-1} = u; \\ 0, & \text{otherwise.} \end{cases}$$

Let A^n be the set of words of length n over A. The infinite word $y^{(n)} = (y_i \cdots y_{i+n-1})_{i \geqslant 0}$ over the alphabet A^n is a fixed point of the substitution $\sigma_n : (A^n)^* \to (A^n)^*$ defined, for all $(a_1 \cdots a_n)$ in A^n, by

$$\sigma_n((a_1 \cdots a_n)) = (b_1 \cdots b_n)(b_2 \cdots b_{n+1}) \cdots (b_{|\sigma(a_1)|} \cdots b_{|\sigma(a_1)|+n-1})$$

where $\sigma(a_1 \cdots a_n) = b_1 \cdots b_k$. For details, see Section V.4 in [103].

Let $\rho : A^n \to A^*$ be the coding defined by $\rho((b_1 \cdots b_n)) = b_1$ for all $(b_1 \cdots b_n) \in A^n$. We have $\rho \circ \sigma_n = \sigma \circ \rho$, and then $\rho \circ \sigma_n^k = \sigma^k \circ \rho$. Hence, if σ is of growth type (d, α) then $y^{(n)}$ is (d, α)-substitutive. Let $f : A^n \to \{0,1\}$ be the coding defined by

$$f((b_1 \cdots b_n)) = \begin{cases} 1, & \text{if } b_1 \cdots b_n = u; \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that $f(y^{(n)}) = t$, hence t is (d, α)-substitutive. Then one proceeds in the same way with τ and uses the result for letters to conclude the proof. □

5.4 Ultimate periodicity

Definition 5.8. Let $\sigma : A^* \to A^*$ be a substitution. If there exists a sub-alphabet $B \subseteq A$ such that for all $b \in B$, $\sigma(b) \in B^*$, then the substitution $\tau : B^* \to B^*$ defined by the restriction $\tau(b) = \sigma(b)$, for all $b \in B$, is a *sub-substitution* of σ. Note that σ is in particular a sub-substitution of itself.

The substitution σ having α as dominating eigenvalue is a *"good" substitution*, if it has a *primitive* sub-substitution whose dominating eigenvalue is α. So let us stress the fact that to be a "good" substitution, the sub-substitution has to be primitive and have the same dominating eigenvalue as the original substitution. We let $\mathcal{S}_{\text{good}}$ denote the set of good substitutions.

Remark 5.15. For all growing substitutions σ, there exists an integer k such that σ^k has a primitive sub-substitution. Hence by taking a convenient power of σ, the substitution can always be assumed to have a primitive sub-substitution.

Note that primitive substitutions and uniform substitutions are good substitutions. Now consider the substitution $\sigma : \{a, 0, 1\}^* \to \{a, 0, 1\}^*$ given by $\sigma : a \mapsto aa0, 0 \mapsto 01, 1 \mapsto 0$. Its dominating eigenvalue is 2 and it has only one primitive sub-substitution ($0 \mapsto 01,\ 1 \mapsto 0$) whose dominating eigenvalue is $(1 + \sqrt{5})/2$, hence it is not a good substitution.

Remark 5.16. Let $\sigma : A^* \to A^*$ and $\tau : B^* \to B^*$ be two substitutions such that σ projects on τ, recall (5.2) for the definition of projection. There exists a coding $\phi : A \to B$ such that $\phi \circ \sigma = \tau \circ \phi$. Note that $\phi \circ \sigma^n = \tau^n \circ \phi$. If τ is primitive, then it follows that σ belongs to $\mathcal{S}_{\text{good}}$.

Theorem 5.17. *Let α and β be two multiplicatively independent Perron numbers. Let $x \in A^\omega$ where A is a finite alphabet. Then the following are equivalent:*

(i) *the infinite word x is both α-substitutive w.r.t. $\mathcal{S}_{\text{good}}$ and β-substitutive w.r.t. $\mathcal{S}_{\text{good}}$;*

(ii) *the infinite word x is ultimately periodic.*

Proof. Let $\sigma : B^* \to B^*$ (resp. $\tau : C^* \to C^*$) be a substitution in $\mathcal{S}_{\text{good}}$ having α (resp. β) as its dominating eigenvalue and ϕ (resp. ψ) be a coding such that $x = \phi(\sigma^\infty(b))$ for some $b \in B$ (resp. $x = \psi(\tau^\infty(c))$ for some $c \in C$).

Let us first suppose that both substitutions are growing. In this way, taking a power if needed, we can suppose that they have primitive sub-substitutions.

By Theorem 5.13, the factors occurring infinitely many times in x appear with bounded gaps. Hence for any primitive and growing sub-substitutions $\overline{\sigma}$ and $\overline{\tau}$ of σ and of τ respectively, we have $\phi(L(\overline{\sigma})) = \psi(L(\overline{\tau})) = L$. Using Theorem 5.9 it follows that L is periodic, *i.e.*, there exists a shortest word u, appearing infinitely many times in x, such that $L = L(u^\omega)$. Thus u appears with bounded gaps. Let $\mathcal{R}_u$ be the set of return words to u. A word w is a *return word* to u if $wu \in L(x)$, u is a prefix of wu and u has exactly two occurrences in wu. Since u appears with bounded gaps, the set $\mathcal{R}_u$ is finite. There exists an integer N such that all words $wu \in L(x_N x_{N+1} \cdots)$ appear infinitely many times in x for all $w \in \mathcal{R}_u$. Hence these words appear with bounded gaps in x. We set $t = x_N x_{N+1} \cdots$ and we will prove that t is periodic. Consequently x would be ultimately periodic. We can suppose that u is a prefix of t. Then t is a concatenation of return words to u. Let w be a return word to u. It appears with bounded gaps hence it appears in some $\phi(\overline{\sigma}^n(a))$, where $\overline{\sigma}$ is a primitive and growing sub-substitution, and there exist two words, p and q, and an integer i such that $wu = pu^i q$. As $|u|$ is the least period of L it must be that $wu = u^i$. It follows that $t = u^\omega$.

If, for example, σ is non-growing, then a result of J.-J. Pansiot [94] asserts that either by modifying in a suitable way σ and ϕ (in that case α could be replaced by a power of α) we can suppose σ is growing or $L(\sigma^\infty(b))$ contains the language of a periodic infinite word. We have treated the first case before. For the second case it suffices to use Theorem 5.13. □

Suppose α and β are multiplicatively independent real numbers and that x is a α-substitutive infinite word w.r.t. $\mathcal{S}_{\text{good}}$ and y is a β-substitutive infinite word w.r.t. $\mathcal{S}_{\text{good}}$ satisfying $L(x) \subset L(y)$. Then the conclusion of Theorem 5.8 is far from true. It suffices to look at Example 5.4 and the observation made after Theorem 5.9.

Remark 5.18. The Statement $(\mathcal{S}, \mathcal{S}')$ remains open when $\mathcal{S}$ is the set of substitutions which are not good. Nevertheless there are cases where we can say more. For example, if x is both α-substitutive and β-substitutive (with α and β being multiplicatively independent), and, $L(x)$ contains the language of a periodic sequence then, from Theorem 5.13, we deduce that x is ultimately periodic.

Moreover, as we will see in the next section, this statement holds true in the purely substitutive context.

5.5 The case of fixed points

Now let restrict ourselves to the purely substitutive case. In this setting Cobham's theorem holds true. Note that in the statement of the following result, α and β are necessarily Perron numbers. Moreover, since the substitutions are growing, then α and β must be larger than one.

Theorem 5.19. *Let $\sigma : A^* \to A^*$ and $\tau : A^* \to A^*$ be two non-erasing growing substitutions prolongable on $a \in A$ with respective dominating eigenvalues α and β. Suppose that all letters of A appear in $\sigma^\infty(a)$ and in $\tau^\infty(a)$ and that α and β are multiplicatively independent. If $x = \sigma^\infty(a) = \tau^\infty(a)$, then x is ultimately periodic.*

Proof. Thanks to Remark 5.15, we may assume that σ has a primitive sub-substitution. Using Theorem 5.13, the letters appearing infinitely often in x appear with bounded gaps. Let $\overline{\sigma} : \overline{A} \to \overline{A}$ be a primitive sub-substitution of σ. Let $c \in \overline{A}$. Suppose that there exists a letter b, appearing infinitely many times in x, which does not belong to $\overline{A}$. Then the word $\sigma^n(c) = \overline{\sigma}^n(c)$ does not contain b and b could not appear with bounded gaps. Consequently all letters (and in particular a letter of maximal growth) appearing infinitely often in x belong to $\overline{A}$. Hence $\overline{\sigma}$ also has α as dominating eigenvalue and σ is a "good" substitution. In the same way τ is a "good" substitution. Theorem 5.17 concludes the proof. □

5.6 Back to numeration systems

Let $\mathcal{S}$ be an abstract numeration system. There is no reason for the substitutions describing characteristic words of $\mathcal{S}$-recognizable sets (see Corollary 5.4) to be primitive. To obtain a Cobham type theorem for families of abstract numeration systems, one has to interpret Theorem 5.17 in this formalism.

5.6.1 Polynomially growing abstract numeration systems Here we only mention the following result. The paper [40] is also of interest. It is well-known that the growth function counting the number of words of length n in a regular language is either *polynomial*, *i.e.*, in $\mathcal{O}(n^k)$ for some integer k or *exponential*, *i.e.*, in $\Omega(\theta^n)$ for some $\theta > 1$.

Proposition 5.20. [49] *Let $\mathcal{S} = (L, A, <)$ (resp. $\mathcal{T} = (M, B, \prec)$) be an abstract numeration system where L is a polynomial regular language (resp. M is an exponential regular language). A set X of integers is both $\mathcal{S}$-recognizable and $\mathcal{T}$-recognizable if and only if X is ultimately periodic.*

5.6.2 Bertrand basis and ω_α-substitutive words Let U be a Bertrand numeration basis such that $\mathrm{rep}_U(\mathbb{N}) = L(\alpha)$ where α is a Parry number which is not an integer. In [54] a substitution denoted by ω_α is defined. The importance of this substitution is justified by Theorem 5.21. If $d_\alpha(1) = t_1 \cdots t_n 0^\omega$, $t_n \neq 0$, then ω_α is defined on the alphabet $\{1, \ldots, n\}$ by

$$1 \mapsto 1^{t_1}2,\ \ldots,\ n-1 \mapsto 1^{t_{n-1}}n,\ n \mapsto 1^{t_n}\ .$$

If $d_\alpha(1) = t_1 \cdots t_n (t_{n+1} t_{n+2} \cdots t_{n+m})^\omega$, where n and m are minimal and where $t_{n+1} + t_{n+2} + \cdots + t_{n+m} \neq 0$, then ω_α is defined on the alphabet $\{1, \cdots, n+m\}$ by

$$1 \mapsto 1^{t_1}2,\ \ldots,\ n+m-1 \mapsto 1^{t_{n+m-1}}(n+m),\ n+m \mapsto 1^{t_{n+m}}(n+1)\ .$$

In both cases the substitution ω_α is primitive and has α as dominating eigenvalue. A substitution that projects (see Definition 5.6) on ω_α is called a ω_α*-substitution* and we call each infinite word which is the image under a coding of a fixed point of a ω_α-substitution a ω_α*-substitutive* infinite word (α-automatic infinite word in [54]).

Theorem 5.21. [54, Corollary 1] *Let U be a Bertrand numeration basis such that $\mathrm{rep}_U(\mathbb{N}) = L(\alpha)$ where α is a Parry number. A set $X \subset \mathbb{N}$ is U-recognizable if and only if its characteristic sequence $\mathbb{1}_X$ is ω_α-substitutive.*

Remark 5.16 and Theorem 5.17 imply the following result.

Theorem 5.22. [45] *Let U and V be two Bertrand numeration systems. Let α and β be two multiplicatively independent Parry numbers such that $\mathrm{rep}_U(\mathbb{N}) = L(\alpha)$ and $\mathrm{rep}_V(\mathbb{N}) = L(\beta)$. A set $X \subseteq \mathbb{N}$ is U-recognizable and V-recognizable if and only if X is ultimately periodic.*

6 Cobham's theorem in various contexts

6.1 Regular sequences

Regular sequences as presented in [6, 7, 9] are a generalization of automatic sequences for sequences taking infinitely many values. Many examples of such sequences are given in the first two references. Also see [41] for a generalization of the notion of automaticity in the framework of group actions. Let R be a commutative ring. Let $k \geqslant 2$. Consider a sequence $x = (x_n)_{n\geqslant 0}$ taking values in some R-module. If the R-module generated by all sequences in the k-kernel $N_k(x)$ is finitely generated (recall Theorem 3.2) then the sequence x is said to be (R,k)*-regular*.

Theorem 6.1 (Cobham–Bell theorem [10]). *Let R be a commutative ring*[3]*. Let k, ℓ be two multiplicatively independent integers. If a sequence $x \in R^{\mathbb{N}}$ is both (R,k)-regular and (R,ℓ)-regular, then it satisfies a linear recurrence over R.*

6.2 Algebraic setting and quasi-automatic functions

In [32] G. Christol characterized p-recognizable sets in terms of formal power series.

Theorem 6.2. *Let p be a prime number and $\mathbb{F}_p$ be the field with p elements. A subset $A \subset \mathbb{N}$ is p-recognizable if and only if $f(X) = \sum_{n\in A} X^n \in \mathbb{F}_p[[X]]$ is algebraic over $\mathbb{F}_p(X)$.*

This was applied to Cobham's theorem in [33] to obtain an algebraic version.

Theorem 6.3. *Let A be a finite alphabet, $x \in A^{\mathbb{N}}$, and, K_1 and K_2 be two finite fields with different characteristics. Let $\alpha_1 : A \to K_1$ and $\alpha_2 : A \to K_2$ be two one-to-one maps. If $f(X) = \sum_{n\in\mathbb{N}} \alpha_1(x_n)X^n \in K_1[[X]]$ is algebraic over $K_1(X)$ and $f(X) = \sum_{n\in\mathbb{N}} \alpha_2(x_n)X^n \in K_2[[X]]$ is algebraic over $K_2(X)$ then $f(X)$ is rational.*

Quasi-automatic functions are introduced by Kedlaya in [74]. Also see [75] where Christol's theorem is generalized to Hahn's generalized power series. In this algebraic setting, an extension of Cobham's theorem is proved by Adamczewski and Bell in [1]. Details are given in the chapter "Automata in number theory" of this handbook.

6.3 Real numbers and verification of infinite-state systems

Sets of numbers recognized by finite automata arise when analyzing systems with unbounded mixed variables taking integer or real values. Therefore systems such as timed or hybrid automata are considered [17]. One needs to develop data structures representing sets manipulated during the exploration of infinite state systems. For instance, it is often needed to compute the set of reachable configurations of such a system. Let $k \geqslant 2$ be an integer. Considering separately integer and fractional parts, a real number $x > 0$ can be decomposed as

$$x = \sum_{i=0}^{d} c_i\, k^i + \sum_{i=1}^{+\infty} c_{-i}\, k^{-i},\ c_i \in [\![0, k-1]\!],\ i \leqslant d, \tag{6.1}$$

and gives rise to the infinite word $c_d \cdots c_0 \star c_{-1}c_{-2}\cdots$ over $[\![0,k-1]\!] \cup \{\star\}$ which is a *k-ary representation* of x. Note that rational numbers of the kind p/k^n have two k-ary

[3] Note that in [6] the ground ring R is assumed to be Noetherian (every ideal in R is finitely generated), but this extra assumption is not needed in the above statement.

representations, one ending with 0^ω and one with $(k-1)^\omega$. For the representation of negative elements, one can consider base k-complements or signed number representations [77], the sign being determined by the most significant digit which is thus 0 or $k-1$ (and this digit may be repeated an arbitrary number of times). For definition of Büchi and Muller automata, see the first part of this handbook.

Definition 6.1. A set $X \subseteq \mathbb{R}$ is *k-recognizable* if there exists a Büchi automaton accepting all the k-ary representations of the elements in X. Such an automaton is called a *Real Number Automaton* or *RNA*.

These notions extend naturally to subsets of $\mathbb{R}^d$ and to *Real Vector Automata* or *RVA*. Also the Büchi theorem 4.5 holds for a suitable structure $\langle \mathbb{R}, \mathbb{Z}, +, <, V_k \rangle$, see [22].

Theorem 6.4. [21] *If $X \subseteq \mathbb{R}^d$ is definable by a first-order formula in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, then X written in base $k \geqslant 2$ is accepted by a weak deterministic RVA $\mathcal{A}$.*

Weakness means that each strongly connected component of $\mathcal{A}$ contains only accepting states or only non-accepting states.

Theorem 6.5. [18] *Let $k, \ell \geqslant 2$ be two multiplicatively independent integers. If $X \subseteq \mathbb{R}$ is both k- and ℓ-recognizable by two weak deterministic RVA, then it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

The extension of the Cobham–Semenov theorem for subsets of $\mathbb{R}^d$ in this setting is discussed in [20], see also [24] for a comprehensive presentation. The case of two coprime bases was first considered in [18]. Surprisingly, if the multiplicatively independent bases $k, \ell \geqslant 2$ share the same prime factors, then there exists a subset of $\mathbb{R}$ that is both k- and ℓ-recognizable but not definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, see [19]. This shows a main difference between recognizability of subsets of real numbers written in base k for (general) Büchi automata and weak deterministic RVA. Though written in a completely different language, a similar result was independently obtained in [2]. This latter paper is motivated by the study of some fractal sets.

6.4 Dynamical systems and subshifts

In this section we would like to express a Cobham-type theorem in terms of dynamical systems called substitutive subshifts. Theorem 5.9 will appear as a direct corollary of these developments.

We first need some definitions.

A *dynamical system* is a pair (X, S) where X is a compact metric space and S a continuous map from X onto itself. The dynamical system (X, S) is *minimal* whenever X and the empty set are the only S-invariant closed subsets of X, that is, $S(X) = X$. We say that a minimal system (X, S) is *periodic* whenever X is finite.

Let (X, S) and (Y, T) be two dynamical systems. We say that (Y, T) is a *factor* of (X, S) if there is a continuous and onto map $\phi : X \to Y$ such that $\phi \circ S = T \circ \phi$ (ϕ is called a *factor map*). If ϕ is one-to-one we say that ϕ is an *isomorphism* and that (X, S) and (Y, T) are *isomorphic*.

Let A be an alphabet. We endow A^ω with the infinite product of the discrete topologies. It is a metric space where the metric is given by

$$d(x, y) = \frac{1}{2^n} \text{ with } n = \inf\{k \mid x_k \neq y_k\}, \tag{6.2}$$

where $x = (x_n)_{n\geqslant 0}$ and $y = (y_n)_{n\geqslant 0}$ are two elements of A^ω. A *subshift* on A is a pair $(X, T_{|X})$ where X is a closed T-invariant subset of A^ω and T is the *shift transformation* $T : A^\omega \to A^\omega,\ (x_n)_{n\geqslant 0} \mapsto (x_{n+1})_{n\geqslant 0}$.

Let u be a word over A. The set $[u]_X = \{x \in X \mid x_0 \cdots x_{|u|-1} = u\}$ is a *cylinder*. The family of these sets is a base of the induced topology on X. When there is no misunderstanding, we write $[u]$ and T instead of $[u]_X$ and $T_{|X}$.

Let $x \in A^\omega$. The set $\{y \in A^\omega \mid L(y) \subseteq L(x)\}$ is denoted $\Omega(x)$. It is clear that $(\Omega(x), T)$ is a subshift. We say that $(\Omega(x), T)$ is the *subshift generated* by x. When x is a sequence, we have $\Omega(x) = \overline{\{T^n x \mid n \in \mathbb{N}\}}$. Observe that $(\Omega(x), T)$ is minimal if and only if x is uniformly recurrent, i.e., all its factors occur infinitely often in x and for each factor u of x, there exists a constant K such that the distance between two consecutive occurrences of u in x is bounded by K.

Let ϕ be a factor map from the subshift (X, T) on the alphabet A onto the subshift (Y, T) on the alphabet B. Here $x_{[i,j]}$ denotes the word $x_i \cdots x_j$, $i \leqslant j$. The Curtis–Hedlund–Lyndon theorem [83, Thm. 6.2.9] asserts that ϕ is a *sliding block code*: there exists an r-*block map* $f : A^r \to B$ such that $(\phi(x))_i = f(x_{[i,i+r-1]})$ for all $i \in \mathbb{N}$ and $x \in X$. We shall say that f is a *block map associated to* ϕ and that f *defines* ϕ. If $u = u_0 u_1 \cdots u_{n-1}$ is a word of length $n \geqslant r$ we define $f(u)$ by $(f(u))_i = f(u_{[i,i+r-1]})$, $i \in \{0, 1, \cdots, n-r+1\}$. Let C denote the alphabet A^r and $Z = \{(x_{[i,r+i-1]})_{i\geqslant 0} \mid (x_n)_{n\geqslant 0} \in X\}$. It is easy to check that the subshift (Z, T) is isomorphic to (X, T) and that f induces a 1-block map (a coding) from C onto B which defines a factor map from (Z, T) onto (Y, T).

We can now state a Cobham-type theorem for subshifts generated by substitutive sequences. Observe that it implies Theorem 5.9 and Statement $(\mathcal{S}, \mathcal{S}')$ when $\mathcal{S} = \mathcal{S}'$ is the set of primitive substitutions.

Theorem 6.6. *Let (X, T) and (Y, T) be two subshifts generated respectively by a primitive α-substitutive sequence x and by a primitive β-substitutive sequence y. Suppose (X, T) and (Y, T) both factorize to the subshift (Z, T). If α and β are multiplicatively independent then (Z, T) is periodic.*

Below we give a sketch of the proof, which involves the concept of an ergodic measure. An *invariant measure* for the dynamical system (X, S) is a probability measure μ, on the σ-algebra $\mathcal{B}(X)$ of Borel sets, with $\mu(S^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}(X)$; the measure is *ergodic* if every S-invariant Borel set has measure 0 or 1. The set of invariant measures for (X, S) is denoted by $\mathcal{M}(X, S)$. The system (X, S) is *uniquely ergodic* if $\#(\mathcal{M}(X, S)) = 1$. For expository books on subshifts and/or ergodic theory, see [37, 76, 83, 103, 79].

It is well known that the subshifts generated by primitive substitutive sequences are uniquely ergodic [103].

Let $\phi : X \to Z$ and $\psi : Y \to Z$ be two factor maps. Suppose that (Z,T) is not periodic. We will prove that α and β are multiplicatively independent.

Let μ and λ be the unique ergodic measures of (X,T) and (Y,T) respectively. It is not difficult to see that (Z,T) is also generated by a primitive substitutive sequence and consequently is uniquely ergodic. Let δ be its unique ergodic measure. We notice that $\phi\mu$ defined by $\phi\mu(A) = \mu(\phi^{-1}(A))$, for all Borel sets A of Z, and $\psi\lambda$ defined by $\psi\lambda(A) = \mu(\psi^{-1}(A))$, for all Borel sets A of Z, are invariant measures for (Z,T). Hence $\phi\mu = \delta = \psi\lambda$. Let us give more details about these measures in order to conclude the proof.

Theorem 6.7. [69] *Let (Ω,T) be a subshift generated by a primitive purely γ-substitutive sequence and m be its unique ergodic measure. Then, the measures of cylinders in Ω lie in a finite union of geometric progressions. There exists a finite set $\mathcal{F}$ of positive real numbers such that*

$$\{m(C) \mid C \textit{ cylinder of } X\} \subset \bigcup_{n\in\mathbb{N}} \gamma^{-n}\mathcal{F}\,.$$

In conjunction with the next result and using the pigeon hole principle we will conclude the proof.

Proposition 6.8. [46] *Let (Ω,T) be a subshift generated by a primitive substitutive sequence on the alphabet A. There exists a constant K such that for any block map $f : A^{2r+1} \to B$, we have $\#(f^{-1}(\{u\})) \leqslant K$ for all u appearing in some sequences of $f(\Omega)$.*

¿From these last two results we deduce that there exist two sets of numbers $\mathcal{F}_X$ and $\mathcal{F}_Y$ such that

$$\begin{aligned}\{\delta(C) \mid C \text{ cylinder of } Z\} =& \{\mu(\phi^{-1}(C)) \mid C \text{ cylinder of } Z\}\\ =& \{\lambda(\psi^{-1}(C)) \mid C \text{ cylinder of } X\}\\ \subset& \left(\bigcup_{n\in\mathbb{N}} \alpha^{-n}\mathcal{F}_X\right) \bigcap \left(\bigcup_{n\in\mathbb{N}} \beta^{-n}\mathcal{F}_Y\right).\end{aligned}$$

The sets $\mathcal{F}_X$ and $\mathcal{F}_Y$ being finite, there exist two cylinder sets U and V of Z, $a \in \mathcal{F}_X$, $b \in \mathcal{F}_Y$ and n, m, r, s four distinct positive integers, such that

$$a\alpha^{-n} = \delta(U) = b\beta^{-m} \text{ and } a\alpha^{-r} = \delta(V) = b\beta^{-s}\,.$$

Consequently α and β are multiplicatively dependent.

6.5 Tilings

6.5.1 From definable sets Let A be a finite alphabet. An *array* in $\mathbb{N}^d$ is a map $\mathcal{T} : \mathbb{N}^d \to A$. It can be viewed as a tiling of $\mathbb{R}^d_+$. The collection of all these arrays is $A^{\mathbb{N}^d}$. For all $\mathbf{x} \in \mathbb{N}^d$, let $|\mathbf{x}|$ denote the sum of the coordinates of $\mathbf{x}$ and $B(\mathbf{x}, r)$ be the set $\{(y_1,\ldots,y_d) \in \mathbb{N}^d \mid 0 \leqslant y_i - x_i < r, 1 \leqslant i \leqslant d\}$.

We say $\mathcal{T}$ is *periodic* (resp. *ultimately periodic*) if there exists $\mathbf{p} \in \mathbb{N}^d$ such that $\mathcal{T}(\mathbf{x}+\mathbf{p}) = \mathcal{T}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{N}^d$ (resp. for all large enough $\mathbf{x}$). We also need another notion of periodicity. We say that $Z \subset \mathbb{N}^d$ is $\mathbf{p}$-*periodic inside* $X \subset \mathbb{N}^d$ if for any $\mathbf{x} \in X$ with $\mathbf{x}+\mathbf{p} \in X$ we have

$$\mathbf{x} \in Z \text{ if and only if } \mathbf{x}+\mathbf{p} \in Z \, .$$

We say that Z is *locally periodic* if there exists a non-empty finite set $V \subset \mathbb{N}^d$ of non-zero vectors such that for some $K > \max\{|\mathbf{v}| \mid \mathbf{v} \in V\}$ and $L \geqslant 0$ one has:

$$(\forall \mathbf{x} \in \mathbb{N}^d, |\mathbf{x}| \geqslant L)(\exists \mathbf{v} \in V)(Z \text{ is } \mathbf{v}\text{-periodic inside } B(\mathbf{x}, K)) \, .$$

Observe that for $d = 1$, local periodicity is equivalent to ultimate periodicity. We say $\mathcal{T}$ is *pseudo-periodic* if for all $a \in A$, $\mathcal{T}^{-1}(a)$ is locally periodic and every $(d-1)$-section of $\mathcal{T}^{-1}(a)$, say $S(i,n) = \{\mathbf{x} \in \mathcal{T}^{-1}(a) \mid x_i = n\}$, $1 \leqslant i \leqslant d$ and $n \in \mathbb{N}$, is pseudo-periodic (ultimately periodic when $d-1=1$). The following criterion is due to Muchnik, see [91] for the proof.

Proposition 6.9. *Let $E \subset \mathbb{N}^d$ and $\mathcal{T} : \mathbb{N}^d \to \{0,1\}$ be its characteristic function. The following are equivalent:*

(*i*) *E is definable in the Presburger arithmetic;*

(*ii*) *$\mathcal{T}$ is pseudo-periodic;*

(*iii*) *for all $a \in \{0,1\}$, there exist $n \in \mathbb{N}$, $\mathbf{v}_i \in \mathbb{N}^d$ and finite sets $V_i \subset \mathbb{N}^d$, $0 \leqslant i \leqslant n$ such that*

$$\mathcal{T}^{-1}(a) = V_0 \cup \left(\bigcup_{1 \leqslant i \leqslant n} \left(\mathbf{v}_i + \sum_{\mathbf{v} \in V_i} \mathbb{N}\mathbf{v} \right) \right) \, .$$

Let p be a positive integer and A be a finite alphabet. A *p-substitution* (or substitution if we do not need to specify p) is a map $S : A \to A^{B_p}$ where $B_p = B(\mathbf{0}, p) = \Pi_{i=1}^d \{0, \cdots, p-1\}$. The substitution S can be considered as a function from $A^{\mathbb{N}^d}$ into itself by setting

$$S((\mathcal{T}(\mathbf{x})) = [S(\mathcal{T}(\mathbf{y}))](\mathbf{z}), \text{ for all } \mathcal{T} \in A^{\mathbb{N}^d}$$

where $\mathbf{y} \in \mathbb{N}^d$ and $\mathbf{z} \in B_p$ are the unique vectors satisfying $\mathbf{x} = p\mathbf{y} + \mathbf{z}$.

In the same way, we can define $S : A^{B_{p^n}} \to A^{B_{p^{n+1}}}$. We remark that $S^n(a) = S(S^{n-1}(a))$ for all $a \in A$ and $n > 0$. We say $\mathcal{T}$ is *generated by a p-substitution* if there exist a coding ϕ and a fixed point $\mathcal{T}_0$ of a p-substitution such that $\mathcal{T} = \phi \circ \mathcal{T}_0$.

In [30] the authors proved the following theorem, which is analogous to Theorem 3.1.

Theorem 6.10. *Let $p \geqslant 2$ and $d \geqslant 1$. A set $E \subset \mathbb{N}^d$ is p-recognizable if and only if the characteristic function of E is generated by a p-substitution.*

Hence we can reformulate the Cobham–Semenov theorem as follows [113].

Theorem 6.11 (Cobham–Semenov theorem, Version 2). *Let p and q be two multiplicatively independent integers greater or equal to 2. Then, the array $\mathcal{T}$ is generated by both a p-substitution and a q-substitution if and only if $\mathcal{T}$ is pseudo-periodic.*

A dynamical proof of this can be given as for the unidimensional case, see [48] for the primitive case.

6.5.2 Self-similar tilings In [38], a Cobham-like theorem is expressed in terms of self-similar tilings of $\mathbb{R}^d$ with a proof using ergodic measures, see [116] for more about self-similar tilings. From the point of view of dynamical systems, the main result in [93] is also a Cobham-like theorem for self-similar tilings.

6.6 Toward Cobham's theorem for the Gaussian integers

I. Kátai and J. Szabó proved in [73] that the sequences $((-p+i)^n)_{n\geqslant 0}$ and $((-p-i)^n)_{n\geqslant 0}$ give rise to numeration systems whose set of digits is $\{0,1,\ldots,p^2\}$, $p\in\mathbb{N}\setminus\{0\}$. It is an exercise to check that when $p\in\mathbb{N}\setminus\{0\}$ and $q\in\mathbb{N}\setminus\{0\}$ are different then $-p+i$ and $-q+i$ are multiplicatively independent. Therefore one could expect a Cobham-type theorem for the set of Gaussian integers $\mathbb{G}=\{a+ib \mid a,b\in\mathbb{Z}\}$. A subset $S\subset\mathbb{G}$ is *periodic* if there exists $h\in\mathbb{G}$ such that, for all $g\in\mathbb{G}$, $s\in S$ if and only if $s+gh\in S$. G. Hansel and T. Safer conjectured in [65] the following:

Conjecture 6.12. *Let p and q be two different positive integers and $S\in\mathbb{G}$. Then the following are equivalent.*

(*i*) *The set S is $(-p+i)$-recognizable and $(-q+i)$-recognizable;*
(*ii*) *There exists a periodic set P such that the symmetric difference set $S\Delta P$ is finite.*

The proof that (ii) implies (i) is easy. They tried to prove the other implication using the following (classical) steps:

(1) $D_{p,q}=\left\{\frac{(-p+i)^n}{(-q+i)^m} \mid n,m\in\mathbb{Z}\right\}$ is dense in $\mathbb{C}$.
(2) S is syndetic
(3) S is periodic up to some finite set.

They succeeded in proving (ii) as given by the next result.

Theorem 6.13. *Let p and q be two positive integers such that the set $D_{p,q}$ is dense in $\mathbb{C}$. Let $S\subset\mathbb{G}$ be $(-p+i)$-recognizable and $(-q+i)$-recognizable. Then, S is syndetic.*

Let us make some observations about the density of the set $D_{p,q}$. Let $-p+i=ae^{i\theta}$ and $-q+i=be^{i\phi}$.

Proposition 6.14. *The following are equivalent.*

(*i*) *The set $D_{p,q}$ is dense in $\mathbb{C}$;*
(*ii*) *The set $D_{p,q}$ is dense on the circle: $\{e^{i\theta}\mid\theta\in\mathbb{R}\}\subset\overline{D_{p,q}}$;*
(*iii*) *The following numbers are rationally independent (or linearly dependent over $\mathbb{Q}$):*

$$\frac{\ln b}{\ln a},\ \frac{\theta}{2\pi}\frac{\ln b}{\ln a}-\frac{\phi}{2\pi},\ 1\,.$$

The equivalence between (i) and (iii) is proven in [65] from an easy computation. The equivalence between (i) and (ii) comes from the fact that $p^2 + 1$ and $q^2 + 1$ are multiplicatively independent, see [65, Prop. 2]. As an example, take $p = 1$ and $p = 2$. Then, $a = \sqrt{2}$, $b = \sqrt{5}$, $\theta = \frac{3\pi}{4}$ and $\phi = \arctan(-\frac{1}{2})$. Proving the density of $D_{1,2}$ is equivalent to proving that $\ln 5/\ln 2$, $\arctan(1/2)/\pi$ and 1 are rationally independent. In [65] the authors observe that the Four Exponential Conjecture, see [120], would imply that $D_{p,q}$ is dense in $\mathbb{C}$.

Conjecture 6.15 (Four Exponential Conjecture). *Let $\{\lambda_1, \lambda_2\}$ and $\{x_1, x_2\}$ be two pairs of rationally independent complex numbers. Then, one of the numbers $e^{\lambda_1 x_1}$, $e^{\lambda_1 x_2}$, $e^{\lambda_2 x_1}$, $e^{\lambda_2 x_2}$ is transcendental.*

6.7 Recognizability over $\mathbb{F}_q[X]$

Using the analogy existing between $\mathbb{Z}$ and the ring of polynomials over a finite field $\mathbb{F}_q$ of positive characteristic, one can easily define B-recognizable sets of polynomials [106]. In [121] characterization of these sets in a convenient logical structure analogous to Theorem 4.5 is given. A family of sets of polynomials recognizable in all polynomial bases is described in [106, 121]. We can again conjecture a Cobham-like theorem.

7 Decidability issues

So far we have seen that ultimately periodic sets have a very special status in the context of numeration systems (recall Proposition 2.6, Theorem 5.1 or Theorems 5.17 and 5.19). They can be described using a finite amount of data (two finite words for the preperiodic and the periodic parts). Let us settle down once more to the usual integer base numeration system. Let $X \subseteq \mathbb{N}$ be a k-recognizable set of integers given by a DFA accepting $\mathrm{rep}_k(X)$. Is there an algorithmic decision procedure which permits one to decide for any such set X, whether or not X is ultimately periodic? For an integer base, the problem was solved positively in [70]. The main ideas are the following ones. Given a DFA $\mathcal{A}$ accepting a k-recognizable set $X \subseteq \mathbb{N}$, the number of states of $\mathcal{A}$ gives an upper bound on the possible index and period for X. Consequently, there are finitely many candidates to check. For each such pair (i, p) of candidates, produce a DFA for all possible corresponding ultimately periodic sets and compare it with $\mathcal{A}$. Using non-deterministic finite automata, the same problem was solved in [5]. With the formalism of first order logic the problem becomes trivial. If a set $X \subseteq \mathbb{N}$ is k-recognizable, then using Theorem 4.5 it is definable by a formula $\varphi(x)$ in $\langle \mathbb{N}, +, V_k \rangle$ and X is ultimately periodic if and only if $(\exists p)(\exists N)(\forall x)(x \geqslant N \wedge (\varphi(x) \leftrightarrow \varphi(x + p)))$. Since we have a decidable theory, it is decidable whether this latter sentence is true [28, Prop. 8.2]. The problem can be extended to $\mathbb{Z}^d$ and was discussed in [91]. It is solved in polynomial time in [82]. In view of Theorem 5.1 the question is extended to any abstract numeration system. Let $\mathcal{S}$ be an abstract numeration system. Given a DFA accepting an $\mathcal{S}$-recognizable set $X \subseteq \mathbb{N}$. Decide whether or not X is ultimately periodic. Some special cases have been solved positively in [31, 11]. Using Corollary 5.3, the same question can be asked in terms of morphisms. Given a morphism $\sigma : A^* \to A^*$ prolongable on a letter a and a coding $\tau : A \to B$, decide whether or not $\tau(\sigma^\infty(a))$ is ultimately periodic. It is the *HD0L (ultimate) periodicity problem*. The purely substitutive case was solved independently in [95] and [67]. Note that the general substitutive case is still open (one has to give a decision procedure for any abstract numeration system). Also see [86, 87] where decidability questions about almost-periodicity are considered. A word is *almost periodic* if factors occurring infinitely often have a bounded distance between occurrences (but some factors may occur only finitely often).

8 Acknowledgments

We would like to warmly thank Valérie Berthé, Alexis Bès, Véronique Bruyère, Christiane Frougny, Julien Leroy and Narad Rampersad for their useful comments after reading a preliminary version of this work.

References

[1] B. Adamczewski and J. Bell. Function fields in positive characteristic: expansions and Cobham's theorem. *J. Algebra*, 319:2337–2350, 2008. 881

[2] B. Adamczewski and J. Bell. An analogue of Cobham's theorem for fractals. *Trans. Amer. Math. Soc.*, 363:4421–4442, 2011. 882

[3] B. Alexeev. Minimal DFAs for testing divisibility. *J. Comput. Syst. Sci.*, 69:235–243, 2004. 860

[4] J.-P. Allouche. Sur la complexité des suites infinies. *Bull. Belg. Math. Soc.*, 1:133–143, 1994. 865

[5] J.-P. Allouche, N. Rampersad, and J. Shallit. Periodicity, repetitions, and orbits of an automatic sequence. *Theoret. Comput. Sci.*, 410(30-32):2795–2803, 2009. 887

[6] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. *Theoret. Comput. Sci.*, 98(2):163–197, 1992. 880, 881

[7] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. II. *Theoret. Comput. Sci.*, 307(1):3–29, 2003. Words. 880

[8] J.-P. Allouche and J. O. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and Their Applications, Proceedings of SETA '98*, pages 1–16. Springer-Verlag, 1999. 865

[9] J.-P. Allouche and J. O. Shallit. *Automatic Sequences, Theory, Applications, Generalizations*. Cambridge University Press, 2003. 863, 864, 865, 873, 880

[10] J. P. Bell. A generalization of Cobham's theorem for regular sequences. *Sém. Lothar. Combin.*, 54A:Art. B54Ap, 15 pp. (electronic), 2005/07. 881

[11] J. P. Bell, E. Charlier, A. S. Fraenkel, and M. Rigo. A decision problem for ultimately periodic sets in non-standard numeration systems. *Internat. J. Algebra Comput.*, 19:809–839, 2009. 888

[12] V. Berthé and M. Rigo, editors. *Combinatorics, Automata and Number Theory*, volume 135 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 2009. 860, 862, 864, 865, 870, 875

[13] A. Bertrand-Mathis. Développement en base θ; répartition modulo un de la suite $(x\theta^n)_{n\geqslant 0}$; langages codés et θ-shift. *Bull. Soc. Math. France*, 114:271–323, 1986. 861, 862

[14] A. Bertrand-Mathis. Comment écrire les nombres entiers dans une base que n'est pas entière. *Acta Math. Hung.*, 54:237–241, 1989. 861, 863

[15] A. Bès. An extension of the Cobham-Semënov theorem. *J. Symbolic Logic*, 65(1):201–211, 2000. 868, 869, 874

[16] A. Bès. A survey of arithmetical definability. *Bull. Belg. Math. Soc. Simon Stevin*, (suppl.):1–54, 2001. A tribute to Maurice Boffa. 867

[17] B. Boigelot, L. Bronne, and S. Rassart. An improved reachability analysis method for strongly linear hybrid systems. In *Proc. 9th CAV*, volume 1254 of *Lecture Notes in Computer Science*, pages 167–177, 1997. 881

[18] B. Boigelot and J. Brusten. A generalization of Cobham's theorem to automata over real numbers. *Theoret. Comput. Sci.*, 410(18):1694–1703, 2009. 882

[19] B. Boigelot, J. Brusten, and V. Bruyère. On the sets of real numbers recognized by finite automata in multiple bases. In *Proc. 35th ICALP (Reykjavik)*, volume 5126 of *Lecture Notes in Computer Science*, pages 112–123. Springer-Verlag, 2008. 882

[20] B. Boigelot, J. Brusten, and J. Leroux. A generalization of Semenov's theorem to automata over real numbers. In R. A. Schmidt, editor, *Automated Deduction, 22nd International Conference, CADE 2009, McGill University, Montreal*, volume 5663 of *Lecture Notes in Computer Science*, pages 469–484, 2009. 882

[21] B. Boigelot, S. Jodogne, and P. Wolper. An effective decision procedure for linear arithmetic over the integers and reals. *ACM Trans. Comput. Log.*, 6(3):614–633, 2005. 882

[22] B. Boigelot, S. Rassart, and P. Wolper. On the expressiveness of real and integer arithmetic automata. In *Proc. 25th ICALP (Aalborg)*, volume 1443 of *Lecture Notes in Computer Science*, pages 152–163. Springer, 1998. 882

[23] S. Brlek. Enumeration of factors in the Thue-Morse word. *Discrete Appl. Math.*, 24:83–96, 1989. 866

[24] J. Brusten. *On the sets of real vectors recognized by finite automata in multiple bases*. PhD thesis, University of Liège, Fac. Sciences Appliquées, 2011. 882

[25] V. Bruyère. Automata and numeration systems. *Sém. Lothar. Combin.*, 35:Art. B35b, approx. 19 pp. (electronic), 1995. 859

[26] V. Bruyère. On Cobham's theorem. Thematic term on semigroups, algorithms, automata and languages, School on automata and languages, Coimbra, Portugal, 2001. 859

[27] V. Bruyère and G. Hansel. Bertrand numeration systems and recognizability. *Theoret. Comput. Sci.*, 181(1):17–43, 1997. Latin American Theoretical INformatics (Valparaíso, 1995). 861, 868

[28] V. Bruyère, G. Hansel, C. Michaux, and R. Villemaire. Logic and p-recognizable sets of integers. *Bull. Belg. Math. Soc.*, 1:191–238, 1994. Corrigendum, *Bull. Belg. Math. Soc.* **1** (1994), 577. 859, 867, 868, 869, 887

[29] J. R. Büchi. Weak secord-order arithmetic and finite automata. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 6:66–92, 1960. Reprinted in S. Mac Lane and D. Siefkes, eds., *The Collected Works of J. Richard Büchi*, Springer-Verlag, 1990, pp. 398–424. 868

[30] A. Černý and J. Gruska. Modular trellises. In G. Rozenberg and A. Salomaa, editors, *The Book of L*, pages 45–61. Springer-Verlag, 1986. 885

[31] E. Charlier and M. Rigo. A decision problem for ultimately periodic sets in non-standard numeration systems. In *Proc. of the 33th International Symposium: MFCS (Torun)*, volume 5162 of *Lecture Notes in Computer Science*, pages 241–252, 2008. 888

[32] G. Christol. Ensembles presque périodiques k-reconnaissables. *Theoret. Comput. Sci.*, 9:141–145, 1979. 881

[33] G. Christol, T. Kamae, M. Mendès France, and G. Rauzy. Suites algébriques, automates et substitutions. *Bull. Soc. Math. France*, 108:401–419, 1980. 881

[34] A. Cobham. On the Hartmanis-Stearns problem for a class of tag machines. In *IEEE Conference Record of 1968 Ninth Annual Symposium on Switching and Automata Theory*, pages 51–60, 1968. Also appeared as IBM Research Technical Report RC-2178, August 23 1968. 873

[35] A. Cobham. On the base-dependence of sets of numbers recognizable by finite automata. *Math. Systems Theory*, 3:186–192, 1969. 859, 863

[36] A. Cobham. Uniform tag sequences. *Math. Systems Theory*, 6:164–192, 1972. 863, 864, 865

[37] I. P. Cornfeld, S. V. Fomin, and Y. G. Sinaĭ. *Ergodic theory*, volume 245 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, New York, 1982. 883

[38] M. I. Cortez and F. Durand. Self-similar tiling systems, topological factors and stretching factors. *Discrete Comput. Geom.*, 40:622–640, 2008. 886

[39] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theoret. Comput. Sci.*, 63(3):333–348, 1989. 866

[40] V. Diekert and D. Krieger. Some remarks about stabilizers. *Theoret. Comput. Sci.*, 410(30-32):2935–2946, 2009. 880

[41] A. W. M. Dress and F. von Haeseler. A semigroup approach to automaticity. *Ann. Comb.*, 7(2):171–190, 2003. 880

[42] F. Durand. Cobham's theorem for substitutions. *J. Eur. Math. Soc. (JEMS)*. to appear. 874

[43] F. Durand. A characterization of substitutive sequences using return words. *Discrete Math.*, 179(1-3):89–101, 1998. 873

[44] F. Durand. A generalization of Cobham's theorem. *Theory Comput. Syst.*, 31(2):169–185, 1998. 873, 874

[45] F. Durand. Sur les ensembles d'entiers reconnaissables. *J. Théor. Nombres Bordeaux*, 10(1):65–84, 1998. 873, 874, 880

[46] F. Durand. Linearly recurrent subshifts have a finite number of non-periodic subshift factors. *Ergodic Theory Dynam. Systems*, 20(4):1061–1078, 2000. 884

[47] F. Durand. A theorem of Cobham for non primitive substitutions. *Acta Arith.*, 104:225–241, 2002. 873

[48] F. Durand. Cobham-Semenov theorem and $\mathbb{N}^d$-subshifts. *Theoret. Comput. Sci.*, 391(1-2):20–38, 2008. 886

[49] F. Durand and M. Rigo. Syndeticity and independent substitutions. *Adv. in Appl. Math.*, 42:1–22, 2009. 873, 876, 877, 880

[50] H.-D. Ebbinghaus, J. Flum, and W. Thomas. *Mathematical logic*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1994. 868

[51] S. Eilenberg. *Automata, Languages, and Machines*, volume A. Academic Press, 1974. 863, 864

[52] H. B. Enderton. *A mathematical introduction to logic*. Academic Press, New York, 1972. 868

[53] S. Fabre. Une généralisation du théorème de Cobham. *Acta Arith.*, 67:197–208, 1994. 873

[54] S. Fabre. Substitutions et β-systèmes de numération. *Theoret. Comput. Sci.*, 137(2):219–236, 1995. 874, 880

[55] I. Fagnot. Sur les facteurs des mots automatiques. *Theoret. Comput. Sci.*, 172:67–89, 1997. 874

[56] N. P. Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel. 865, 872

[57] A. S. Fraenkel. Systems of numeration. *Amer. Math. Monthly*, 92:105–114, 1985. 860

[58] C. Frougny. Representations of numbers and finite automata. *Math. Systems Theory*, 25(1):37–60, 1992. 861, 868

[59] C. Frougny. Non-standard number representation: computer arithmetic, beta-numeration and quasicrystals. In *Physics and theoretical computer science*, volume 7 of *NATO Secur. Sci. Ser. D Inf. Commun. Secur.*, pages 155–169. IOS, Amsterdam, 2007. 858

[60] F. R. Gantmacher. *The Theory of Matrices*. Chelsea, 1960. 871

[61] S. Ginsburg and E. H. Spanier. Semigroups, Presburger formulas, and languages. *Pacific J. Math.*, 16:285–296, 1966. 869

[62] W. H. Gottschalk and G. A. Hedlund. *Topological Dynamics*, volume 36 of *AMS Colloquium Publications*. Amer. Math. Soc., 1955. 863

[63] G. Hansel. A propos d'un théorème de Cobham. In D. Perrin, editor, *Actes de la Fête des Mots*, pages 55–59. Greco de Programmation, CNRS, Rouen, 1982. 863

[64] G. Hansel. Systèmes de numération indépendants et syndéticité. *Theoret. Comput. Sci.*, 204:119–130, 1998. 863

[65] G. Hansel and T. Safer. Vers un théorème de Cobham pour les entiers de Gauss. *Bull. Belg. Math. Soc. Simon Stevin*, 10:723–735, 2003. 886, 887

[66] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, 5th edition, 1985. 859

[67] T. Harju and M. Linna. On the periodicity of morphisms on free monoids. *RAIRO Inform. Théor. App.*, 20:47–54, 1986. 888

[68] M. Hollander. Greedy numeration systems and regularity. *Theory Comput. Systems*, 31:111–133, 1998. 862

[69] C. Holton and L. Q. Zamboni. Directed graphs and substitutions. *Theory Comput. Syst.*, 34(6):545–564, 2001. 884

[70] J. Honkala. A decision method for the recognizability of sets defined by number systems. *RAIRO Inform. Théor. App.*, 20:395–403, 1986. 887

[71] J. Honkala. On the simplification of infinite morphic words. *Theoret. Comput. Sci.*, 410(8-10):997–1000, 2009. 873

[72] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original. 871

[73] I. Kátai and J. Szabó. Canonical number systems for complex integers. *Acta Sci. Math. (Szeged)*, 37(3-4):255–260, 1975. 886

[74] K. S. Kedlaya. The algebraic closure of the power series field in positive characteristic. *Proc. Amer. Math. Soc.*, 129(12):3461–3470, 2001. 881

[75] K. S. Kedlaya. Finite automata and algebraic extensions of function fields. *J. Théor. Nombres Bordeaux*, 18(2):379–420, 2006. 881

[76] B. P. Kitchens. *Symbolic dynamics*. Universitext. Springer-Verlag, Berlin, 1998. One-sided, two-sided and countable state Markov shifts. 883

[77] D. E. Knuth. *The art of computer programming. Vol. 2*. Addison-Wesley Publishing Co., Reading, Mass., second edition, 1981. Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing. 882

[78] D. Krieger, A. Miller, N. Rampersad, B. Ravikumar, and J. Shallit. Decimations of languages and state complexity. *Theoret. Comput. Sci.*, 410:2401–2409, 2009. 870

[79] P. Kůrka. *Topological and symbolic dynamics*, volume 11 of *Cours Spécialisés*. Société Mathématique de France, Paris, 2003. 883

[80] L. Latour. *Presburger Arithmetic: from Automata to Formulas*. PhD thesis, University of Liège, Fac. Sciences Appliquées, 2006. 868

[81] P. B. A. Lecomte and M. Rigo. Numeration systems on a regular language. *Theory Comput. Systems*, 34:27–44, 2001. 870

[82] J. Leroux. A polynomial time Presburger criterion and synthesis for number decision diagrams. In *IEEE Symposium on Logic in Computer Science (LICS 2005)*, IEEE Computer Society, pages 147–156, 2005. 887

[83] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995. 871, 872, 883

[84] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, 1983. 864, 865

[85] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 2002. 860, 861, 862, 864

[86] A. Maes. Morphisms and almost-periodicity. *Discrete Appl. Math.*, 86(2-3):233–248, 1998. 888

[87] A. Maes. More on morphisms and almost-periodicity. *Theoret. Comput. Sci.*, 231(2):205–215, 2000. Universal machines and computations (Metz, 1998). 888

[88] C. Michaux and R. Villemaire. Cobham's theorem seen through Büchi's theorem. In *Automata, languages and programming (Lund, 1993)*, volume 700 of *Lecture Notes in Computer Science*, pages 325–334. Springer, Berlin, 1993. 869

[89] C. Michaux and R. Villemaire. Presburger arithmetic and recognizability of sets of natural numbers by automata: New proofs of Cobham's and Semenov's theorems. *Ann. Pure Appl. Logic*, 77:251–277, 1996. 869

[90] M. Morse and G. A. Hedlund. Symbolic dynamics. *Amer. J. Math.*, 60:815–866, 1938. 865

[91] A. A. Muchnik. The definable criterion for definability in Presburger arithmetic and its applications. *Theoret. Comput. Sci.*, 290(3):1433–1444, 2003. 869, 885, 887

[92] M. Nivat. invited talk at ICALP, bologna, 1997. 867

[93] N. Ormes, C. Radin, and L. Sadun. A homeomorphism invariant for substitution tiling spaces. *Geom. Dedicata*, 90:153–182, 2002. 886

[94] J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés. In *Automata, languages and programming (Antwerp, 1984)*, volume 172 of *Lecture Notes in Computer Science*, pages 380–389. Springer, Berlin, 1984. 878

[95] J.-J. Pansiot. Decidability of periodicity for infinite words. *RAIRO Inform. Théor. App.*, 20:43–46, 1986. 888

[96] W. Parry. On the β-expansions of real numbers. *Acta Math. Acad. Sci. Hung.*, 11:401–416, 1960. 861

[97] B. Pascal. *Œuvres complètes*. Seuil, 1963. The treatise *De numeris multiplicibus*, written with the other arithmetical treatises before 1654, was published by Guillaume Desprez in 1665. 860

[98] D. Perrin. Finite automata. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B:Formal Models and Semantics*, pages 1–57. Elsevier — MIT Press, 1990. 859, 863

[99] F. Point and V. Bruyère. On the Cobham-Semenov theorem. *Theory Comput. Syst.*, 30(2):197–220, 1997. 869

[100] M. Presburger. Über die volständigkeit eines gewissen systems der arithmetik ganzer zahlen, in welchem die addition als einzige operation hervortritt. *C. R. Premier congrès des Mathématiciens des pays slaves, Varsovie*, pages 92–101, 1929. 867, 868

[101] M. Presburger. On the completeness of a certain system of arithmetic of whole numbers in which addition occurs as the only operation. *Hist. Philos. Logic*, 12:225–233, 1991. 867, 868

[102] A. Quas and L. Zamboni. Periodicity and local complexity. *Theoret. Comput. Sci.*, 319:229–240, 2004. 867

[103] M. Queffélec. *Substitution dynamical systems—spectral analysis*, volume 1294 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1987. 877, 883

[104] A. Rényi. Representations for real numbers and their ergodic properties. *Acta Math. Acad. Sci. Hung.*, 8:477–493, 1957. 861

[105] M. Rigo. Generalization of automatic sequences for numeration systems on a regular language. *Theoret. Comput. Sci.*, 244:271–281, 2000. 870

[106] M. Rigo. Syntactical and automatic properties of sets of polynomials over finite fields. *Finite Fields Appl.*, 14(1):258–276, 2008. 887

[107] M. Rigo and A. Maes. More on generalized automatic sequences. *J. Automata, Languages, and Combinatorics*, 7:351–376, 2002. 870

[108] M. Rigo and L. Waxweiler. A note on syndeticity, recognizable sets and Cobham's theorem. *Bull. European Assoc. Theor. Comput. Sci.*, 88:169–173, February 2006. 863

[109] J. Sakarovitch. *Éléments de théorie des automates*. Vuibert, 2003. English corrected edition: *Elements of Automata Theory*, Cambridge University Press, 2009. 860

[110] A. Salomaa and M. Soittola. *Automata-theoretic aspects of formal power series*. Springer-Verlag, New York, 1978. Texts and Monographs in Computer Science. 875

[111] O. Salon. Suites automatiques à multi-indices. In *Séminaire de Théorie des Nombres de Bordeaux*, pages 4.01–4.27, 1986-1987. 866

[112] O. Salon. Suites automatiques à multi-indices et algébricité. *C. R. Acad. Sci. Paris*, 305:501–504, 1987. 866

[113] A. L. Semenov. The Presburger nature of predicates that are regular in two number systems. *Sibirsk. Mat. Ž.*, 18(2):403–418, 479, 1977. In Russian. English translation in *Siberian J. Math.* **18** (1977), 289–300. 869, 885

[114] E. Seneta. *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York, 2006. 871

[115] J. O. Shallit. Numeration systems, linear recurrences, and regular sets. *Inform. Comput.*, 113:331–347, 1994. 861, 870

[116] B. Solomyak. Dynamics of self-similar tilings. *Ergodic Theory Dynam. Systems*, 17(3):695–738, 1997. 886

[117] B. Tan and Z.-Y. Wen. Some properties of the Tribonacci sequence. *European J. Combin.*, 28(6):1703–1719, 2007. 872

[118] R. Villemaire. Joining k- and l-recognizable sets of natural numbers. In *STACS 92 (Cachan, 1992)*, volume 577 of *Lecture Notes in Computer Science*, pages 83–94. Springer, Berlin, 1992. 869

[119] R. Villemaire. The theory of $\langle \mathbf{N}, +, V_k, V_l \rangle$ is undecidable. *Theoret. Comput. Sci.*, 106(2):337–349, 1992. 869

[120] M. Waldschmidt. *Diophantine approximation on linear algebraic groups*, volume 326 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2000. Transcendence properties of the exponential function in several variables. 887

[121] L. Waxweiler. *Caractère reconnaissable d'ensembles de polynômes à coefficients dans un corps fini*. PhD thesis, University of Liège, 2009. http://hdl.handle.net/2268/11381. 887

[122] P. Wolper and B. Boigelot. Verifying systems with infinite but regular state spaces. In *Computer aided verification (Vancouver, BC, 1998)*, volume 1427 of *Lecture Notes in Computer Science*, pages 88–97. Springer, Berlin, 1998. 858

[123] E. Zeckendorf. Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres Lucas. *Bull. Soc. Roy. Liège*, 41:179–182, 1972. 861

Symbolic dynamics

Marie-Pierre Béal[1], *Jean Berstel*[1], *Søren Eilers*[2], *Dominique Perrin*[1]

[1]LIGM (Laboratoire d'Informatique Gaspard-Monge), Université Paris-Est
email: {beal,berstel,perrin}@univ-mlv.fr

[2]Institut for Matematiske Fag, Københavns Universitet
email: eilers@math.ku.dk

2010 Mathematics Subject Classification: 68Q45, 37B10

Key words: Symbolic dynamics,

Contents

1 Introduction

Symbolic dynamics is part of dynamical systems theory. It studies discrete dynamical systems called shift spaces and their relations under appropriately defined morphisms, in particular isomorphisms called conjugacies. A special emphasis has been put on the classification of shift spaces up to conjugacy or flow equivalence.

There is a considerable overlap between symbolic dynamics and automata theory. Actually, one of the basic objects of symbolic dynamics, the sofic systems, are essentially the same as finite automata. In addition, the morphisms of shift spaces are a particular case of rational transductions, that is functions defined by finite automata with output. The difference is that symbolic dynamics considers mostly infinite words and that all states of the automata are initial and final. Also, the morphisms are particular transductions which are given by local maps.

This chapter presents some of the links between automata theory and symbolic dynamics. The emphasis is on two particular points. The first one is the interplay between some particular classes of automata, such as local automata and results on embeddings of shifts of finite type. The second one is the connection between syntactic semigroups and the classification of sofic shifts up to conjugacy.

The chapter is organized as follows. In Section 2, we introduce the basic notions of symbolic dynamics: shift spaces, conjugacy and flow equivalence. We state without proof two important results: the Decomposition Theorem and the Classification Theorem.

In Section 3, we introduce automata in relation to sofic shifts. In Section 4, we define two kinds of minimal automata for shift spaces: the Krieger automaton and the Fischer automaton. We also relate these automata with the syntactic semigroup of a shift space.

In Section 5, we state and prove an analogue due to Nasu of the Decomposition Theorem and of the Classification Theorem.

In Section 6 we consider two special families of automata: local automata and automata with finite delay. We show that they are related to shifts of finite type and of almost finite type, respectively. We prove an embedding theorem (Theorem 6.4) which is a counterpart for automata of a result known as Nasu's masking lemma.

In Section 7 we study syntactic invariants of sofic shifts. We introduce the syntactic graph of an automaton. We show that that the syntactic graph of an automaton is invariant under conjugacy (Theorem 7.4) and also under flow equivalence. We finally state some results concerning the shift spaces corresponding to some pseudovarieties of ordered semigroups.

We follow the notation of the book of Doug Lind and Brian Marcus [19]. In general, we have not not reproduced the proofs of the results which can be found there. We thank Mike Boyle and Alfredo Costa for their help.

2 Shift spaces

This section contains basic definitions concerning symbolic dynamics.

The first subsection gives the definition of shift spaces, and the important case of edge shifts.

The next subsection and thus also under (Section 2.2) introduces conjugacy, and the basic notion of state splitting and merging. It contains the statement of two important theorems, the Decomposition Theorem (Theorem 2.12) and the Classification Theorem (Theorem 2.14).

The last subsection (Section 2.3) introduces flow equivalence, and states Frank's characterization of flow equivalent edge shifts (Theorem 2.16).

2.1 Shift spaces

Let A be a finite alphabet. We denote by A^* the set of words on A and by A^+ the set of nonempty words. A word v is a *factor* of a word t if $t = uvw$ for some words u, w.

We denote by $A^{\mathbb{Z}}$ the set of biinfinite sequences of symbols from A. This set is a topological space in the product topology of the discrete topology on A. The *shift transformation* on $A^{\mathbb{Z}}$ is the map σ_A from $A^{\mathbb{Z}}$ onto itself defined by $y = \sigma_A(x)$ if $y_n = x_{n+1}$ for $n \in \mathbb{Z}$. A set $X \subset A^{\mathbb{Z}}$ is *shift invariant* if $\sigma(X) = X$. A *shift space* on the alphabet A is a shift-invariant subset of $A^{\mathbb{Z}}$ which is closed in the topology. The set $A^{\mathbb{Z}}$ itself is a shift space called the *full shift*.

For a set $W \subset A^*$ of words (whose elements are called the *forbidden factors*), we denote by $X^{(W)}$ the set of $x \in A^{\mathbb{Z}}$ such that no $w \in W$ is a factor of x.

Proposition 2.1. *The shift spaces on the alphabet A are the sets $X^{(W)}$, for $W \subset A^*$.*

A shift space X is of *finite type* if there is a finite set $W \subset A^*$ such that $X = X^{(W)}$.

Example 2.1. Let $A = \{a, b\}$, and let $W = \{bb\}$. The shift $X^{(W)}$ is composed of the sequences without two consecutive b's. It is a shift of finite type, called the *golden mean shift*.

Recall that a set $W \subset A^*$ is said to be *recognizable* if it can be recognized by a finite automaton or, equivalently, defined by a regular expression. A shift space X is said to be *sofic* if there is a recognizable set W such that $X = X^{(W)}$. Since a finite set is recognizable, any shift of finite type is sofic.

Example 2.2. Let $A = \{a, b\}$, and let $W = a(bb)^*ba$. The shift $X^{(W)}$ is composed of the sequences where two consecutive occurrences of the symbol a are separated by an even number of b's. It is a sofic shift called the *even shift*. It is not a shift of finite type. Indeed, assume that $X = X^{(V)}$ for a finite set $V \subset A^*$. Let n be the maximal length of the words of V. A biinfinite repetition of the word ab^n has the same blocks of length at most n as a biinfinite repetition of the word ab^{n+1}. However, one is in X if and only if the other is not in X, a contradiction.

Example 2.3. Let $A = \{a, b\}$ and let $W = \{ba^n b^m a \mid n, m \geqslant 1, n \neq m\}$. The shift $X^{(W)}$ is composed of infinite sequences of the form $\ldots a^{n_i} b^{n_i} a^{n_{i+1}} b^{n_{i+1}} \ldots$. The set W is not recognizable and it can be shown that X is not sofic.

Edge shifts. In this chapter, a *graph* $G = (Q, \mathcal{E})$ is a pair composed of a finite set Q of *vertices* (or *states*), and a finite set $\mathcal{E}$ of *edges*. The graph is equipped with two maps $i, t : \mathcal{E} \to Q$ which associate, to an edge e, its *initial* and *terminal* vertex[1]. We say that e starts in $i(e)$ and ends in $t(e)$. Sometimes, $i(e)$ is called the *source* and $t(e)$ is called the *target* of e.

We also say that e is an incoming edge for $t(e)$, and an outgoing edge for $i(e)$. Two edges $e, e' \in \mathcal{E}$ are *consecutive* if $t(e) = i(e')$.

For $p, q \in Q$, we denote by $\mathcal{E}_p^q$ the set of edges of a graph $G = (Q, \mathcal{E})$ starting in state p and ending in state q. The *adjacency matrix* of a graph $G = (Q, \mathcal{E})$ is the $Q \times Q$-matrix $M(G)$ with elements in $\mathbb{N}$ defined by

$$M(G)_{pq} = \mathrm{Card}(\mathcal{E}_p^q)\,.$$

A (finite or biinfinite) *path* is a (finite or biinfinite) sequence of consecutive edges. The *edge shift* on the graph G is the set of biinfinite paths in G. It is denoted by X_G and is a shift of finite type on the alphabet of edges. Indeed, it can be defined by taking the set of non-consecutive edges for the set of forbidden factors. The converse does not hold, since the golden mean shift is not an edge shift. However, we shall see below (Proposition 2.5) that every shift of finite type is conjugate to an edge shift.

A graph is *essential* if every state has at least one incoming and one outgoing edge. This implies that every edge is on a biinfinite path. The *essential part* of a graph G is the subgraph obtained by restricting to the set of vertices and edges which are on a biinfinite path.

2.2 Conjugacy

Morphisms. Let X be a shift space on an alphabet A, and let Y be a shift space on an alphabet B.

A *morphism* φ from X into Y is a continuous map from X into Y which commutes with the shift. This means that $\varphi \circ \sigma_A = \sigma_B \circ \varphi$.

Let k be a positive integer. A *k-block* of X is a factor of length k of an element of X. We denote by $\mathcal{B}(X)$ the set of all blocks of X and by $\mathcal{B}_k(X)$ the set of k-blocks of X. A function $f : \mathcal{B}_k(X) \to B$ is called a *k-block substitution*. Let now m, n be fixed nonnegative integers with $k = m + 1 + n$. Then the function f defines a map φ called *sliding block map* with *memory* m and *anticipation* n as follows. The image of $x \in X$ is the element $y = \varphi(x) \in B^{\mathbb{Z}}$ given by

$$y_i = f(x_{i-m} \cdots x_i \cdots x_{i+n})\,.$$

We denote $\varphi = f_\infty^{[m,n]}$. It is a sliding block map from X into Y if y is in Y for all x in X. We also say that φ is a k-block map from X into Y. The simplest case occurs when $m = n = 0$. In this case, φ is a 1-block map.

The following result is Theorem 6.2.9 in [19].

[1] We avoid the use of the terms 'initial state' or 'terminal state' of an edge to avoid confusion with the initial or terminal states of an automaton

Theorem 2.2 (Curtis–Lyndon–Hedlund). *A map from a shift space X into a shift space Y is a morphism if and only if it is a sliding block map.*

Conjugacies of shifts. A morphism from a shift X onto a shift Y is called a *conjugacy* if it is one-to-one from X onto Y. Note that in this case, using standard topological arguments, one shows that the inverse mapping is also a morphism, and thus a conjugacy.

We define the n-th *higher block shift* $X^{[n]}$ of a shift X over the alphabet A as follows. The alphabet of $X^{[n]}$ is the set $B = \mathcal{B}_n(X)$ of blocks of length n of X.

Proposition 2.3. *The shifts X and $X^{[n]}$ for $n \geqslant 1$ are conjugate.*

Proof. Let $f : \mathcal{B}_n(X) \to B$ be the n-block substitution which maps the factor $x_1 \cdots x_n$ to itself, viewed as a symbol of the alphabet B. By construction, the shift $X^{[n]}$ is the image of X by the map $f_\infty^{[n-1,0]}$. This map is a conjugacy since it is bijective, and its inverse is the 1-block map g_∞ corresponding to the 1-block map which associates to the symbol $x_1 \cdots x_n$ of B the symbol x_n of A. □

Let $G = (Q, \mathcal{E})$ be a graph. For an integer $n \geqslant 1$, denote by $G^{[n]}$ the following graph called the n-th *higher edge graph* of G. For $n = 1$, one has $G^{[1]} = G$. For $n > 1$, the set of states of $G^{[n]}$ is the set of paths of length $n - 1$ in G. The edges of $G^{[n]}$ are the paths of length n of G. The start state of an edge $(e_1, e_2, \ldots, e_n)$ is $(e_1, e_2, \ldots, e_{n-1})$ and its end state is $(e_2, e_3, \ldots, e_n)$.

The following result shows that the higher block shifts of an edge shift are again edge shifts.

Proposition 2.4. *Let G be a graph. For $n \geqslant 1$, one has $X_G^{[n]} = X_{G^{[n]}}$.*

A shift of finite type need not be an edge shift. For example the golden mean shift of Example 2.1 is not an edge shift. However, any shift of finite type comes from an edge shift in the following sense.

Proposition 2.5. *Every shift of finite type is conjugate to an edge shift.*

Proof. We show that for every shift of finite type X there is an integer n such that $X^{[n]}$ is an edge shift. Let $W \subset A^*$ be a finite set of words such that $X = X^{(W)}$, and let n be the maximal length of the words of W. If $n = 0$, X is the full shift. Thus we assume $n \geqslant 1$. Define a graph G whose vertices are the blocks of length $n - 1$ of X, and whose edges are the block of length n of X. For $w \in \mathcal{B}_n(X)$, the initial (resp. terminal) vertex of w is the prefix (resp. suffix) of length $n - 1$ of w.

We show that $X_G = X^{[n]}$. An element of $X^{[n]}$ is always an infinite path in G. To show the other inclusion, consider an infinite path y in G. It is the sequence of n-blocks of an element x of $A^{\mathbb{Z}}$ which does not contain any block on W. Since $X = X^{(W)}$, we get that x is in X. Consequently, y is in $X^{[n]}$. This proves the equality. □

Proposition 2.6. *A shift space that is conjugate to a shift of finite type is itself of finite type.*

Proof. Let $\varphi : X \to Y$ be a conjugacy from a shift of finite type X onto a shift space Y. By Proposition 2.5, we may assume that $X = X_G$ for some graph G. Changing G into some higher edge graph, we may assume that φ is 1-block. We may consider G as a graph labeled by φ. Suppose that φ^{-1} has memory m and anticipation n. Set $\varphi^{-1} = f_\infty^{[m,n]}$. Let W be the set of words of length $m + n + 2$ which are not the label of a path in G. We show that $Y = X^{(W)}$, which implies that Y is of finite type. Indeed, the inclusion $Y \subset X^{(W)}$ is clear. Conversely, consider y in $X^{(W)}$. For each $i \in \mathbb{Z}$, set $x_i = f(y_{i-m} \cdots y_i \cdots y_{i+n})$. Since $y_{i-m} \cdots y_i \cdots y_{i+n} y_{i+n+1}$ is the label of a path in G, the edges x_i and x_{i+1} are consecutive. Thus $x = (x_i)_{i \in \mathbb{Z}}$ is in X and $y = \varphi(x)$ is in Y. □

Conjugacy invariants. No effective characterization of conjugate shift spaces is known, even for shifts of finite type. There are however several quantities that are known to be invariant under conjugacy.

The *entropy* of a shift space X is defined by

$$h(X) = \lim_{n\to\infty} \frac{1}{n} \log s_n \,,$$

where $s_n = \mathrm{Card}(\mathcal{B}_n(X))$. The limit exists because the sequence s_n is sub-additive (see [19] Lemma 4.1.7). Note that since $\mathrm{Card}(\mathcal{B}_n(X)) \leqslant \mathrm{Card}(A)^n$, we have $h(X) \leqslant \log \mathrm{Card}(A)$. If X is nonempty, then $0 \leqslant h(X)$.

The following statement shows that the entropy is invariant under conjugacy (see [19] Corollary 4.1.10).

Theorem 2.7. *If X, Y are conjugate shift spaces, then* $h(X) = h(Y)$.

Example 2.4. Let X be the golden mean shift of Example 2.1. Then a block of length $n + 1$ is either a block of length $n - 1$ followed by ab or a block of length n followed by a. Thus $s_{n+1} = s_n + s_{n-1}$. As a classical result, $h(X) = \log \lambda$ where $\lambda = (1 + \sqrt{5})/2$ is the golden mean.

An element x of a shift space X over the alphabet A has *period* n if $\sigma_A^n(x) = x$. If $\varphi : X \to Y$ is a conjugacy, then an element x of X has period n if and only if $\varphi(x)$ has period n.

The *zeta function* of a shift space X is the power series

$$\zeta_X(z) = \exp \sum_{n \geqslant 0} \frac{p_n}{n} z^n \,,$$

where p_n is the number of elements x of X of period n.

It follows from the definition that the sequence $(p_n)_{n\in\mathbb{N}}$ is invariant under conjugacy, and thus the zeta function of a shift space is invariant under conjugacy.

Several other conjugacy invariants are known. One of them is the Bowen-Franks group of a matrix which defines an invariant of the associated shift space. This will be defined below.

Example 2.5. Let $X = A^{\mathbb{Z}}$. Then $\zeta_X(z) = \frac{1}{1-kz}$, where $k = \mathrm{Card}(A)$. Indeed, one has

$p_n = k^n$, since an element x of $A^{\mathbb{Z}}$ has period n if and only if it is a biinfinite repetition of a word of length n over A.

State splitting. Let $G = (Q, \mathcal{E})$ and $H = (R, \mathcal{F})$ be graphs. A pair (h, k) of surjective maps $k : R \to Q$ and $h : \mathcal{F} \to \mathcal{E}$ is called a *graph morphism* from H onto G if the two diagrams in Figure 1 are commutative.

$$\begin{tikzcd}
\mathcal{F} \arrow[r, "h"] \arrow[d, "i"] & \mathcal{E} \arrow[d, "i"] & & \mathcal{F} \arrow[r, "h"] \arrow[d, "t"] & \mathcal{E} \arrow[d, "t"] \\
R \arrow[r, "k"] & Q & & R \arrow[r, "k"] & Q
\end{tikzcd}$$

Figure 1. Graph morphism.

A graph morphism (h, k) from H onto G is an *in-merge* from H onto G if for each $p, q \in Q$ there is a partition $(\mathcal{E}_p^q(t))_{t \in k^{-1}(q)}$ of the set $\mathcal{E}_p^q$ such that for each $r \in k^{-1}(p)$ and $t \in k^{-1}(q)$, the map h is a bijection from $\mathcal{F}_r^t$ onto $\mathcal{E}_p^q(t)$. If this holds, then G is called an *in-merge* of H, and H is an *in-split* of G.[2]

Thus an in-split H is obtained from a graph G as follows: each state $q \in Q$ is split into copies which are the states of H in the set $k^{-1}(q)$. Each of these states t receives a copy of $\mathcal{E}_p^q(t)$ starting in r and ending in t for each r in $k^{-1}(p)$.

Each r in $k^{-1}(p)$ has the same number of edges going out of r and coming in s, for any $s \in R$.

Moreover, for any $p, q \in Q$ and $e \in \mathcal{E}_p^q$, all edges in $h^{-1}(e)$ have the same terminal vertex, namely the state t such that $e \in \mathcal{E}_p^q(t)$.

Example 2.6. Let G and H be the graphs represented on Figure 2. Here $Q = \{1, 2\}$ and $R = \{3, 4, 5\}$. The graph H is an in-split of the graph G. The graph morphism (h, k)

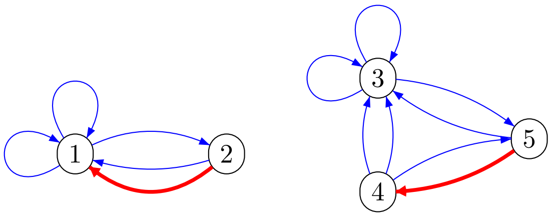


Figure 2. An in-split from G (on the left) onto H (on the right).

is defined by $k(3) = k(4) = 1$ and $k(5) = 2$. Thus the state 1 of G is split into two states 3 and 4 of H, and the map h is associated to the partition obtained as follows: the edges from 2 to 1 are partitioned into two classes, indexed by 3 and 4 respectively, and containing each one edge from 2 to 1.

The following result is well-known (see [19]). It shows that if H is an in-split of a graph G, then X_G and X_H are conjugate.

[2] In this chapter, a *partition* of a set X is a family $(X_i)_{i \in I}$ of pairwise disjoint, possibly empty subsets of X, indexed by a set I, such that X is the union of the sets X_i for $i \in I$.

Proposition 2.8 ([19, Theorem 2.4.10]). *If (h, k) is an in-merge of a graph H onto a graph G, then h_∞ is a 1-block conjugacy from X_H onto X_G and its inverse is 2-block.*

The map h_∞ from X_H to X_G is called an *edge in-merging map* and its inverse an *edge in-splitting map*.

A *column division matrix* over two sets R, Q is an $R \times Q$-matrix D with elements in $\{0, 1\}$ such that each column has at least one 1 and each row has exactly one 1. Thus, the columns of such a matrix represent a partition of R into $\mathrm{Card}(Q)$ sets.

The following result is Theorem 2.4.14 of [19].

Proposition 2.9. *Let G and H be essential graphs. The graph H is an in-split of the graph G if and only if there is an $R \times Q$-column division matrix D and a $Q \times R$-matrix E with nonnegative integer entries such that*

$$M(G) = ED, \quad M(H) = DE. \tag{2.1}$$

Example 2.7. For the graphs G, H of Example 2.6, one has $M(G) = DE$ and $M(H) = ED$ with

$$E = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Observe that a particular case of a column division matrix is a permutation matrix. The corresponding in-split (or merge) is a renaming of the states of a graph.

The notion of an *out-merge* is defined symmetrically. A graph morphism (h, k) from H onto G is an *out-merge* from H onto G if for each $p, q \in Q$ there is a partition $(\mathcal{E}_p^q(r))_{r \in k^{-1}(p)}$ of the set $\mathcal{E}_p^q$ such that for each $r \in k^{-1}(p)$ and $t \in k^{-1}(q)$, the map h is a bijection from $\mathcal{F}_r^t$ onto $\mathcal{E}_p^q(r)$. If this holds, then G is called an *out-merge* of H, and H is an *out-split* of G.

Proposition 2.8 also has a symmetrical version. Thus if (h, k) is an out-merge from G onto H, then h_∞ is a 1-block conjugacy from X_H onto X_G whose inverse is 2-block. The conjugacy h_∞ is called an *edge out-merging map* and its inverse an *edge out-splitting map*.

Symmetrically, a *row division matrix* is a matrix with elements in the set $\{0, 1\}$ such that each column has at least one 1 and each row has exactly one 1.

The following statement is symmetrical to Proposition 2.9.

Proposition 2.10. *Let G and H be essential graphs. The graph H is an out-split of the graph G if and only if there is a row division matrix D and a matrix E with nonnegative integer entries such that*

$$M(G) = DE, \quad M(H) = ED. \tag{2.2}$$

Example 2.8. Let G and H be the graphs represented on Figure 3. Here $Q = \{1, 2\}$ and $R = \{3, 4, 5\}$. The graph H is an out-split of the graph G. The graph morphism (h, k) is defined by $k(3) = k(4) = 1$ and $k(5) = 2$. The map h is associated with the partition

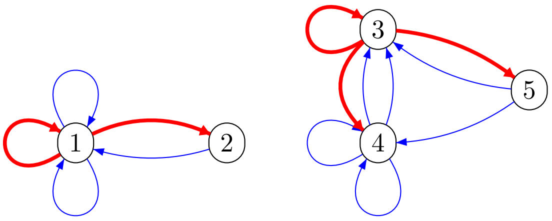


Figure 3. The graphs G and H.

indicated by the colors. One has $M(G) = ED$ and $M(H) = DE$ with

$$D = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 1 & 0 \end{bmatrix}.$$

We use the term *split* to mean either an in-split or an out-split. The same convention holds for a *merge*.

Proposition 2.11. *For $n \geqslant 2$, the graph $G^{[n-1]}$ is an in-merge of the graph $G^{[n]}$.*

Proof. Consider for $n \geqslant 2$ the equivalence on the states of $G^{[n]}$ which relates two paths of length $n-1$ which differ only by the first edge. It is clear that this equivalence is such that two equivalent elements have the same output. Thus $G^{[n-1]}$ is an in-merge of $G^{[n]}$. □

The Decomposition Theorem. The following result is known as the *Decomposition Theorem* (Theorem 7.1.2 in [19]).

Theorem 2.12. *Every conjugacy from an edge shift onto another is the composition of a sequence of edge splitting maps followed by a sequence of edge merging maps.*

The statement of Theorem 2.12 given in [19] is less precise, since it does not specify the order of splitting and merging maps.

The proof relies on the following statement (Lemma 7.1.3 in [19]).

Lemma 2.13. *Let G, H be graphs and let $\varphi : X_G \to X_H$ be a 1-block conjugacy whose inverse has memory $m \geqslant 1$ and anticipation $n \geqslant 0$. There are in-splittings $\overline{G}, \overline{H}$ of the graphs G, H and a 1-block conjugacy with memory $m-1$ and anticipation n $\overline{\varphi} : X_{\overline{G}} \to X_{\overline{H}}$ such that the following diagram commutes.*

The horizontal edges in the above diagram represent the edge in-splitting maps from X_G to $X_{\overline{G}}$ and from X_H to $X_{\overline{H}}$ respectively.

The Classification Theorem. Two nonnegative integral square matrices M, N are *elementary equivalent* if there exists a pair R, S of nonnegative integral matrices such that

$$M = RS\,, \quad N = SR\,.$$

$$\begin{tikzcd}
X_G \arrow[r] \arrow[d, "\varphi"] & X_{\overline{G}} \arrow[d, "\overline{\varphi}"] \\
X_H \arrow[r] & X_{\overline{H}}
\end{tikzcd}$$

Thus if a graph H is a split of a graph G, then, by Proposition 2.9, the matrices $M(G)$ and $M(H)$ are elementary equivalent. The matrices M and N are *strong shift equivalent* if there is a sequence $(M_0, M_1, \ldots, M_n)$ of nonnegative integral matrices such that M_i and M_{i+1} are elementary equivalent for $0 \leqslant i < n$ with $M_0 = M$ and $M_n = N$.

The following theorem is Williams' Classification Theorem (Theorem 7.2.7 in [19]).

Theorem 2.14. *Let G and H be two graphs. The edge shifts X_G and X_H are conjugate if and only if the matrices $M(G)$ and $M(H)$ are strong shift equivalent.*

Note that one direction of this theorem is contained in the Decomposition Theorem. Indeed, if X_G and X_H are conjugate, there is a sequence of edge splitting and edge merging maps from X_G to X_H. And if G is a split or a merge of H, then $M(G)$ and $M(H)$ are elementary equivalent, whence the result in one direction follows. Note also that, in spite of the easy definition of strong shift equivalence, it is not even known whether there exists a decision procedure for determining when two nonnegative integral matrices are strong shift equivalent.

2.3 Flow equivalence

In this section, we give basic definitions and properties concerning flow equivalence of shift spaces. The notion comes from the notion of equivalence of continuous flows, see Section 13.6 of [19]. A characterization of flow equivalence for shift spaces (which we will take below as our definition of flow equivalence for shift spaces) is due to Parry and Sullivan [23]. It is noticeable that the flow equivalence of irreducible shifts of finite type has an effective characterization, by Franks' Theorem (Theorem 2.16).

Let A be an alphabet and a be a letter in A. Let ω be a letter which does not belong to A. Set $B = A \cup \omega$. The *symbol expansion* of a set $W \subset A^+$ relative to a is the image of W by the semigroup morphism $\varphi : A^+ \to B^+$ such that $\varphi(a) = a\omega$ and $\varphi(b) = b$ for all $b \in A \setminus a$. Recall that a *semigroup morphism* $f : A^+ \to B^+$ is a map satisfying $f(xy) = f(x)f(y)$ for all words x, y. It should not be confused with the morphisms of shift spaces defined earlier. The semigroup morphism φ is also called a symbol expansion. Let X be a shift space on the alphabet A. The *symbol expansion* of X relative to a is the least shift space X' on the alphabet $B = A \cup \omega$ which contains the symbol expansion of $\mathcal{B}(X)$. Note that if φ is a symbol expansion, it defines a bijection from $\mathcal{B}(X)$ onto $\mathcal{B}(X')$. The inverse of a symbol expansion is called a *symbol contraction*.

Two shift spaces X, Y are said to be *flow equivalent* if there is a sequence $X_0, \ldots, X_n$ of shift spaces such that $X_0 = X$, $Y_n = Y$ and for $0 \leqslant i \leqslant n-1$, either X_{i+1} is the image of X_i by a conjugacy, a symbol expansion or a symbol contraction.

Example 2.9. Let $A = \{a, b\}$. The symbol expansion of the full shift $A^{\mathbb{Z}}$ relative to b

is conjugate to the golden mean shift. Thus the full shift on two symbols and the golden mean shift are flow equivalent.

For edge shifts, symbol expansion can be replaced by another operation. Let G be a graph and let p be a vertex of G. The *graph expansion* of G relative to p is the graph G' obtained by replacing p by an edge from a new vertex p' to p to and replacing all edges coming in p by edges coming in p' (see Figure 4). The inverse of a graph expansion is called a *graph contraction*. Note that graph expansion (relative to vertex 1) changes the adjacency matrix of a graph as indicated below.

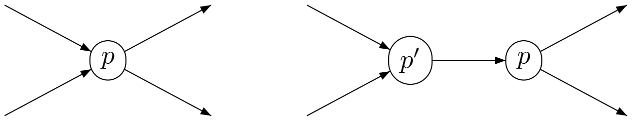


Figure 4. Graph expansion

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & a_{11} & a_{12} & \dots & a_{1n} \\ 1 & 0 & 0 & \dots & 0 \\ 0 & a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & & \\ 0 & a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Proposition 2.15. *The flow equivalence relation on edge shifts is generated by conjugacies and graph expansions.*

Proof. Let $G = (Q, E)$ be a graph and let p be a vertex of G. The graph expansion of G relative to p can be obtained by a symbol expansion of each of the edges coming into p followed by a conjugacy which merges all the new symbols into one new symbol. Conversely, let e be an edge of G. The symbol expansion of X_G relative to e can be obtained by a input split which makes e the only edge going into its end vertex q followed by a graph expansion relative to q. □

The *Bowen-Franks group* of a square $n \times n$-matrix M with integer elements is the Abelian group

$$BF(M) = \mathbb{Z}^n / \mathbb{Z}^n (I - M)$$

where $\mathbb{Z}^n(I - M)$ is the image of $\mathbb{Z}^n$ under the matrix $I - M$ acting on the right. In other terms, $\mathbb{Z}^n(I - M)$ is the Abelian group generated by the rows of the matrix $I - M$. This notion is due to Bowen and Franks [5], who have shown that it is an invariant for flow equivalence.

The following result is due to Franks [14]. We say that a graph is *trivial* if it is reduced to one cycle.

Theorem 2.16. *Let G, G' be two strongly connected nontrivial graphs and let M, M' be their adjacency matrices. The edge shifts $X_G, X_{G'}$ are flow equivalent if and only if* $\det(I - M) = \det(I - M')$ *and the groups $BF(M)$, $BF(M')$ are isomorphic.*

In the case trivial graphs, the theorem is false. Indeed, any two edge shifts on strongly connected trivial graphs are flow equivalent and are not flow equivalent to any edge shift on a nontrivial irreducible graph. For any trivial graph G with adjacency matrix M, one has $\det(I - M) = 0$ and $BF(M) \sim \mathbb{Z}$. However there are nontrivial strongly connected graphs such that $\det(I - M) = 0$ and $BF(M) \sim \mathbb{Z}$.

The case of arbitrary shifts of finite type has been solved by Huang (see [6, 8]). A similar characterization for sofic shifts is not known (see [7]).

Example 2.10. Let

$$M = \begin{bmatrix} 4 & 1 \\ 1 & 0 \end{bmatrix}, \quad M' = \begin{bmatrix} 3 & 2 \\ 1 & 0 \end{bmatrix}.$$

One has $\det(I - M) = \det(I - M') = -4$. Moreover $BF(M) \sim \mathbb{Z}/4\mathbb{Z}$. Indeed, the rows of the matrix $I - M$ are $\begin{bmatrix} -3 & -1 \end{bmatrix}$ and $\begin{bmatrix} -1 & 1 \end{bmatrix}$. They generate the same group as $\begin{bmatrix} 4 & 0 \end{bmatrix}$ and $\begin{bmatrix} -1 & 1 \end{bmatrix}$. Thus $BF(M) \sim \mathbb{Z}/4\mathbb{Z}$. In the same way, $BF(M') \sim \mathbb{Z}/4\mathbb{Z}$. Thus, according to Theorem 2.16, the edge shifts X_G and $X_{G'}$ are flow equivalent.

Actually X_G and $X_{G'}$ are both flow equivalent to the full shift on 5 symbols.

3 Automata

In this section, we start with the definition and notation for automata recognizing shifts, and we show that sofic shifts are precisely the shifts recognized by finite automata (Proposition 3.3).

We introduce the notion of labeled conjugacy; it is a conjugacy preserving the labeling. We extend the Decomposition Theorem and the Classification Theorem to labeled conjugacies (Theorems 3.8 and 3.9).

3.1 Automata and sofic shifts

The automata considered in this section are finite automata. We do not mention the initial and final states in the notation when all states are both initial and final. Thus, an automaton is denoted by $\mathcal{A} = (Q, E)$ where Q is the finite set of *states* and $E \subset Q \times A \times Q$ is the set of *edges*. The edge (p, a, q) has initial state p, label a and terminal state q. The underlying graph of $\mathcal{A}$ is the same as $\mathcal{A}$ except that the labels of the edges are not used.

An automaton is *essential* if its underlying graph is essential. The *essential part* of an automaton is its restriction to the essential part of its underlying graph.

We denote by $X_{\mathcal{A}}$ the set of biinfinite paths in $\mathcal{A}$. It is the edge shift of the underlying graph of $\mathcal{A}$. Note that since the automaton is supposed finite, the shift space $X_{\mathcal{A}}$ is on a finite alphabet, as required for a shift space. We denote by $L_{\mathcal{A}}$ the set of labels of biinfinite paths in $\mathcal{A}$. We denote by $\lambda_{\mathcal{A}}$ the 1-block map from $X_{\mathcal{A}}$ into the full shift $A^{\mathbb{Z}}$ which assigns to a path its label. Thus $L_{\mathcal{A}} = \lambda_{\mathcal{A}}(X_{\mathcal{A}})$. If this holds, we say that $X_{\mathcal{A}}$ is the shift space *recognized* by $\mathcal{A}$.

The following propositions describe how this notion of recognition is related to that for finite words. In the context of finite words, we denote by $\mathcal{A} = (Q, I, E, T)$ an automaton with distinguished subsets I (resp. T) of initial (resp. terminal) states. A word w is *recognized* by $\mathcal{A}$ if there is a path from a state in I to a state in T labeled w. Recall that a set is recognizable if it is the set of words recognized by a finite automaton. An automaton $\mathcal{A} = (Q, I, T)$ is *trim* if, for every state p in Q, there is a path from a state in I to p and a path from p to a state in T.

Proposition 3.1. *Let $W \subset A^*$ be a recognizable set and let $\mathcal{A} = (Q, I, T)$ be a trim finite automaton recognizing the set $A^* \setminus A^*WA^*$. Then $L_{\mathcal{A}} = X^{(W)}$.*

Proof. The label of a biinfinite path in the automaton $\mathcal{A}$ does not contain a factor w in W. Otherwise, there is a finite path $p \xrightarrow{w} q$ which is a segment of this infinite path. The path $p \xrightarrow{w} q$ can be extended to a path $i \xrightarrow{u} p \xrightarrow{w} q \xrightarrow{v} t$ for some $i \in I, t \in T$, and uwv is accepted by $\mathcal{A}$, which is a contradiction.

Next, consider a biinfinite word $x = (x_i)_{i \in \mathbb{Z}}$ in $X^{(W)}$. For every $n \geqslant 0$, there is a path π_n in the automaton $\mathcal{A}$ labeled $w_n = x_{-n} \cdots x_0 \cdots x_n$ because the word w_n has no factor in W. By compactness (König's lemma) there is an infinite path in $\mathcal{A}$ labeled x. Thus x is in $L_{\mathcal{A}}$. □

The following proposition states in some sense the converse.

Proposition 3.2. *Let X be a sofic shift over A, and let $\mathcal{A} = (Q, I, T)$ be a trim finite automaton recognizing the set $\mathcal{B}(X)$ of blocks of X. Then $L_{\mathcal{A}} = X$.*

Proof. Set $W = A^* \setminus \mathcal{B}(X)$. Then one easily checks that $X = X^{(W)}$. Next, $\mathcal{A}$ recognizes $A^* \setminus A^*WA^*$. By Proposition 3.1, one has $L_{\mathcal{A}} = X$. □

Proposition 3.3. *A shift X over A is sofic if and only if there is a finite automaton $\mathcal{A}$ such that $X = L_{\mathcal{A}}$.*

Proof. The forward implication results from Proposition 3.1. Conversely, assume that $X = L_{\mathcal{A}}$ for some finite automaton $\mathcal{A}$. Let W be the set of finite words which are not labels of paths in $\mathcal{A}$. Clearly $X \subset X^{(W)}$. Conversely, if $x \in X^{(W)}$, then all its factors are labels of paths in $\mathcal{A}$. Again by compactness, x itself is the label of a biinfinite path in $\mathcal{A}$. □

Example 3.1. The golden mean shift of Example 2.1 is recognized by the automaton of Figure 5 on the left while the even shift of Example 2.2 is recognized by the automaton of Figure 5 on the right.

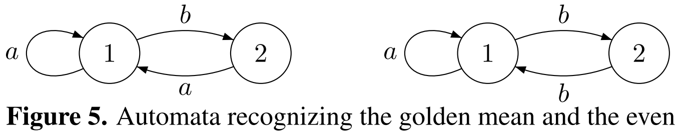


Figure 5. Automata recognizing the golden mean and the even shift

The *adjacency matrix* of the automaton $\mathcal{A} = (Q, E)$ is the $Q \times Q$-matrix $M(\mathcal{A})$ with elements in $\mathbb{N}\langle A\rangle$ defined by

$$(M(\mathcal{A})_{pq}, a) = \begin{cases} 1 & \text{if } (p, a, q) \in E\,, \\ 0 & \text{otherwise.} \end{cases}$$

We write M for $M(\mathcal{A})$ when the automaton is understood. The entries in the matrix M^n, for $n \geqslant 0$, have an easy combinatorial interpretation: for each word w of length n, the coefficient $(M^n_{p,q}, w)$ is the number of distinct paths from p to q carrying the label w.

A matrix M is called *alphabetic* over the alphabet A if its elements are homogeneous polynomials of degree 1 over A with nonnegative coefficients. Adjacency matrices are special cases of alphabetic matrices. Indeed, its elements are homogeneous polynomials of degree 1 with coefficients 0 or 1.

3.2 Labeled conjugacy

Let $\mathcal{A}$ and $\mathcal{B}$ be two automata on the alphabet A. A *labeled conjugacy* from $X_{\mathcal{A}}$ onto $X_{\mathcal{B}}$ is a conjugacy φ such that $\lambda_{\mathcal{A}} = \lambda_{\mathcal{B}}\varphi$, that is such that the following diagram is commutative. We say that $\mathcal{A}$ and $\mathcal{B}$ are *conjugate* if there exists a labeled conjugacy from $X_{\mathcal{A}}$ to $X_{\mathcal{B}}$. The aim of this paragraph is to give two characterizations of labeled conjugacy.

$$\begin{tikzcd} X_{\mathcal{A}} \arrow[rr, "\varphi"] \arrow[dr, "\lambda_{\mathcal{A}}"'] & & X_{\mathcal{B}} \arrow[dl, "\lambda_{\mathcal{B}}"] \\ & A^{\mathbb{Z}} & \end{tikzcd}$$

Labeled split and merge. Let $\mathcal{A} = (Q, E)$ and $\mathcal{B} = (R, F)$ be two automata. Let G, H be the underlying graphs of $\mathcal{A}$ and $\mathcal{B}$ respectively.

A *labeled in-merge* from $\mathcal{B}$ onto $\mathcal{A}$ is an in-merge (h, k) from H onto G such that for each $f \in F$ the labels of f and $h(f)$ are equal. We say that $\mathcal{B}$ is a *labeled in-split* of $\mathcal{A}$, or that $\mathcal{A}$ is a *labeled in-merge* of $\mathcal{B}$.

The following statement is the analogue of Proposition 2.8 for automata.

Proposition 3.4. *If (h, k) is a labeled in-merge from the automaton $\mathcal{B}$ onto the automaton $\mathcal{A}$, then the map h_∞ is a labeled conjugacy from $X_{\mathcal{B}}$ onto $X_{\mathcal{A}}$.*

Proof. Let (h, k) be a labeled in-merge from $\mathcal{B}$ onto $\mathcal{A}$. By Proposition 2.8, the map h_∞ is a 1-block conjugacy from $X_{\mathcal{B}}$ onto $X_{\mathcal{A}}$. Since the labels of f and $h(f)$ are equal for each edge f of $\mathcal{B}$, this map is a labeled conjugacy. □

The next statement is the analogue of Proposition 2.9 for automata.

Proposition 3.5. *An automaton $\mathcal{B} = (R, F)$ is a labeled in-split of the automaton $\mathcal{A} = (Q, E)$ if and only if there is an $R \times Q$-column division matrix D and an alphabetic $Q \times R$-matrix N such that*

$$M(\mathcal{A}) = ND, \quad M(\mathcal{B}) = DN. \tag{3.1}$$

Proof. Suppose first that D and N are as described in the statement, and define a map $k : R \to Q$ by $k(r) = q$ if $D_{rq} = 1$. We define $h : F \to E$ as follows. Consider an edge $(r, a, s) \in F$. Set $p = k(r)$ and $q = k(s)$. Since $M(\mathcal{B}) = DN$, we have $(N_{ps}, a) = 1$. Since $M(\mathcal{A}) = ND$, this implies that $(M(\mathcal{A})_{pq}, a) = 1$ or, equivalently, that $(p, a, q) \in E$. We set $h(r, a, s) = (p, a, q)$. Then (h, k) is a labeled in-merge. Indeed h is associated with the partitions defined by

$$E_p^q(t) = \{(p, a, q) \in E \mid (N_{pt}, a) = 1 \text{ and } k(t) = q\}.$$

Suppose conversely that (h, k) is a labeled in-merge from $\mathcal{B}$ onto $\mathcal{A}$. Let D be the $R \times Q$-column division matrix defined by

$$D_{rq} = \begin{cases} 1 & \text{if } k(r) = q \\ 0 & \text{otherwise} \end{cases}$$

For $p \in Q$ and $t \in R$, we define N_{rt} as follows. Set $q = k(t)$. By definition of an in-merge, there is a partition $(E_p^q(t))_{t \in k^{-1}(q)}$ of E_p^q such that h is a bijection from F_r^t onto $E_p^q(t)$. For $a \in A$, set

$$(N_{pt}, a) = \begin{cases} 1 & \text{if } (p, a, q) \in E_p^q(t) \\ 0 & \text{otherwise} \end{cases}$$

Then $M(\mathcal{A}) = ND$ and $M(\mathcal{B}) = DN$. □

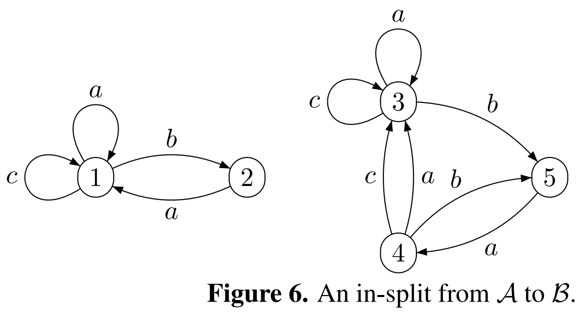


Figure 6. An in-split from $\mathcal{A}$ to $\mathcal{B}$.

Example 3.2. Let $\mathcal{A}$ and $\mathcal{B}$ be the automata represented on Figure 6. Here $Q = \{1, 2\}$ and $R = \{3, 4, 5\}$. One has $M(\mathcal{A}) = ND$ and $M(\mathcal{B}) = DN$ with

$$N = \begin{bmatrix} a + c & 0 & b \\ 0 & a & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

A *labeled out-merge* from $\mathcal{B}$ onto $\mathcal{A}$ is an out-merge (h,k) from H onto G such that for each $f \in F$ the labels of f and $h(f)$ are equal.

We say that $\mathcal{B}$ is a *labeled out-split* of $\mathcal{A}$, or that $\mathcal{A}$ is a *labeled in-merge* of $\mathcal{B}$.

Thus if $\mathcal{B}$ is a labeled out-split of $\mathcal{A}$, there is a labeled conjugacy from $X_{\mathcal{B}}$ onto $X_{\mathcal{A}}$.

Proposition 3.6. *The automaton $\mathcal{B} = (R,F)$ is a labeled out-split of the automaton $\mathcal{A} = (Q,E)$ if and only if there is a $Q \times R$-row division matrix D and an alphabetic $R \times Q$-matrix N such that*

$$M(\mathcal{A}) = DN\,, \quad M(\mathcal{B}) = ND\,. \tag{3.2}$$

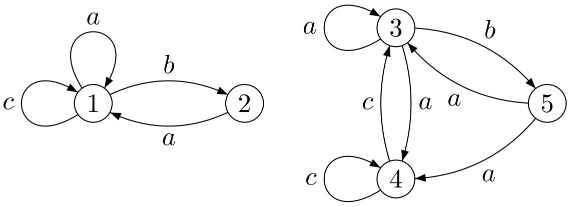


Figure 7. An out-split from $\mathcal{A}$ to $\mathcal{B}$.

Example 3.3. Let $\mathcal{A}$ and $\mathcal{B}$ be the automata represented on Figure 7. Here $Q = \{1,2\}$ and $R = \{3,4,5\}$. One has $M(\mathcal{A}) = ND$ and $M(\mathcal{B}) = DN$ with

$$N = \begin{bmatrix} a & b \\ c & 0 \\ a & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $\mathcal{A} = (Q,E)$ be an automaton. For a pair of integers $m,n \geqslant 0$, denote by $\mathcal{A}^{[m,n]}$ the following automaton called the (m,n)-th *extension* of $\mathcal{A}$. The underlying graph of $\mathcal{A}^{[m,n]}$ is the higher edge graph $G^{[k]}$ for $k = m+n+1$. The label of an edge

$$p_0 \xrightarrow{a_1} p_1 \xrightarrow{a_2} \cdots \xrightarrow{a_m} p_m \xrightarrow{a_{m+1}} p_{m+1} \xrightarrow{a_{m+2}} \cdots \xrightarrow{a_{m+n}} p_{m+n} \xrightarrow{a_{m+n+1}} p_{m+n+1}$$

is the letter a_{m+1}. Observe that $\mathcal{A}^{[0,0]} = \mathcal{A}$. By this construction, each graph $G^{[k]}$ produces k extensions according to the choice of the labeling.

Proposition 3.7. *For $m \geqslant 1, n \geqslant 0$, the automaton $\mathcal{A}^{[m-1,n]}$ is a labeled in-merge of the automaton $\mathcal{A}^{[m,n]}$ and for $m \geqslant 0, n \geqslant 1$, the automaton $\mathcal{A}^{[m,n-1]}$ is a labeled out-merge of the automaton $\mathcal{A}^{[m,n]}$.*

Proof. Suppose that $m \geqslant 1, n \geqslant 0$. Let k be the map from the paths of length $m+n$ in $\mathcal{A}$ onto the paths of length $m+n-1$ which erases the first edge of the path. Let h be the map from the set of edges of $\mathcal{A}^{[m,n]}$ to the set of edges of $\mathcal{A}^{[m-1,n]}$ defined by $h(\pi,a,\rho) = (k(\pi),a,k(\rho))$. Then (h,k) is a labeled in-merge from $\mathcal{A}^{[m,n]}$ onto $\mathcal{A}^{[m-1,n]}$. The proof that, for $m \geqslant 0, n \geqslant 1$, the automaton $\mathcal{A}^{[m,n-1]}$ is an out-merge of the automaton $\mathcal{A}^{[m,n]}$ is symmetrical. □

The following result is the analogue, for automata, of the Decomposition Theorem.

Theorem 3.8. *Every conjugacy of automata is a composition of labeled splits and merges.*

Proof. Let $\mathcal{A}$ and $\mathcal{B}$ be two conjugate automata. Let φ be a labeled conjugacy from $\mathcal{A}$ onto $\mathcal{B}$. Let G_0 and H_0 be the underlying graphs of $\mathcal{A}$ and $\mathcal{B}$, respectively. By the Decomposition Theorem 2.12, there are sequences $(G_1, \ldots, G_n)$ and $(H_1, \ldots, H_m)$ of graphs with $G_n = H_m$ and such that G_{i+1} is a split of G_i for $0 \leqslant i < n$ and H_{j+1} is a split of H_j for $0 \leqslant j < m$. Moreover, φ is the composition of the sequence of edge splitting maps from G_i onto G_{i+1} followed by the sequence of edge merging maps from H_{j+1} onto H_j. Let (h_i, k_i), for $1 \leqslant i \leqslant n$, be a merge from G_i onto G_{i-1} and (u_j, v_j), for $1 \leqslant j \leqslant m$ be a merge from H_j onto H_{j-1}. Then we may define labels on the edges of $G_1, \ldots, G_n$ in such a way that G_i becomes the underlying graph of an automaton $\mathcal{A}_i$ and (h_i, k_i) is a labeled merge from $\mathcal{A}_i$ onto $\mathcal{A}_{i-1}$. In the same way, we may define labels on the edges of H_j in such a way that H_j becomes the underlying graph of an automaton $\mathcal{B}_j$ and (u_j, v_j) is a labeled merge from $\mathcal{B}_j$ onto $\mathcal{B}_{j-1}$.

$$G_0 \xleftarrow{(h_1,k_1)} G_1 \cdots \xleftarrow{(h_n,k_n)} G_n = H_m \xrightarrow{(u_m,v_m)} \cdots H_1 \xrightarrow{(u_1,v_1)} H_0 \,.$$

Let $h = h_1 \cdots h_n$ and $u = u_1 u_2 \cdots u_m$. Since $\varphi = u_\infty h_\infty^{-1}$, and φ is a labeled conjugacy, we have $\lambda_{\mathcal{A}} h_\infty = \lambda_{\mathcal{B}} u_\infty$. This shows that the automata $\mathcal{A}_n$ and $\mathcal{B}_m$ are equal. Thus there is a sequence of labeled splitting maps followed by a sequence of labeled merging maps which is a equal to φ. □

Let M and M' be two alphabetic square matrices over the same alphabet A. We say that M and M' are *elementary equivalent* if there exists a nonnegative integral matrix D and an alphabetic matrix N such that

$$M = DN \,, \quad M' = ND \quad \text{or vice-versa.}$$

By Proposition 3.5, if $\mathcal{B}$ is an in-split of $\mathcal{A}$, then $M(\mathcal{B})$ and $M(\mathcal{A})$ are elementary equivalent. We say that M, M' are *strong shift equivalent* if there is a sequence $(M_0, M_1, \ldots, M_n)$ such that M_i and M_{i+1} are elementary equivalent for $0 \leqslant i < n$ with $M_0 = M$ and $M_n = M'$. The following result is the version, for automata, of the Classification Theorem.

Theorem 3.9. *Two automata are conjugate if and only if their adjacency matrices are strong shift equivalent.*

Note that when D is a column division matrix, the statement results from Propositions 3.4 and 2.9. The following statement proves the theorem in one direction.

Proposition 3.10. *Let $\mathcal{A}$ and $\mathcal{B}$ be two automata. If $M(\mathcal{A})$ is elementary equivalent to $M(\mathcal{B})$, then $\mathcal{A}$ and $\mathcal{B}$ are conjugate.*

Proof. Let $\mathcal{A} = (Q, E)$ and $\mathcal{B} = (R, F)$. Let D be an $R \times Q$ nonnegative integral matrix and let N be an alphabetic $Q \times R$ matrix such that

$$M(\mathcal{A}) = ND, \quad M(\mathcal{B}) = DN.$$

Consider the map f from the set of paths of length 2 in $\mathcal{A}$ into F defined as follows (see Figure 8 on the left). Let $p \xrightarrow{a} q \xrightarrow{b} r$ be a path of length 2 in $\mathcal{A}$. Since $(M(\mathcal{A})_{pq}, a) = 1$ and $M(\mathcal{A}) = ND$ there is a unique $t \in R$ such that $(N_{pt}, a) = D_{tq} = 1$. In the same way, since $(M(\mathcal{A})_{qr}, b) = 1$, there is a unique $u \in R$ such that $(N_{qu}, b) = D_{ur} = 1$. Since $M(\mathcal{B}) = DN$, we have $(M(\mathcal{B})_{tu}, b) = D_{tq} = (N_{qu}, b) = 1$ and thus (t, u, b) is an edge of $\mathcal{B}$. We set

$$f(p \xrightarrow{a} q \xrightarrow{b} r) = t \xrightarrow{b} u$$

Similarly, we may define a map g from the set of paths of length 2 in $\mathcal{B}$ into E by

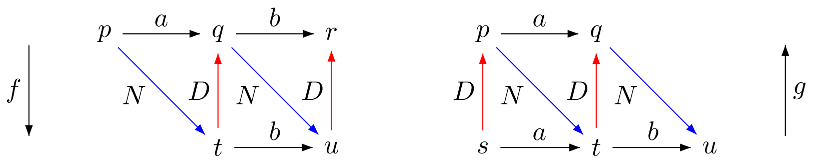


Figure 8. The maps f and g.

$$g(s \xrightarrow{a} t \xrightarrow{b} u) = p \xrightarrow{a} q$$

if $D_{sp} = (N_{pt}, a) = D_{tq} = 1$. Let $\varphi = f_\infty^{[1,0]}$ and $\gamma = g_\infty^{[0,1]}$ (see Figure 8 on the right). We verify that

$$\varphi\gamma = \mathrm{Id}_F, \quad \gamma\varphi = \mathrm{Id}_E$$

where Id_E and Id_F are the identities on $E^{\mathbb{Z}}$ and $F^{\mathbb{Z}}$. Let indeed π be a path in $X_{\mathcal{A}}$ and let $\rho = \varphi(\pi)$. Set $\pi_i = (p_i, a_i, p_{i+1})$ and $\rho_i = (r_i, b_i, r_{i+1})$ (see Figure 9). Then, by definition of φ, we have for all $i \in \mathbb{Z}$, $b_i = a_i$ and $(N_{p_i r_{i+1}}, a_i) = D_{r_i p_i} = 1$. Let $\sigma = \gamma(\rho)$ and $\sigma = (s_i, c_i, s_{i+1})$. By definition of γ, we have $c_i = b_i$ and $D_{r_i s_i} = (N_{s_i r_{i+1}}, b_i) = 1$. Thus we have simultaneously $D_{r_i p_i} = (N_{p_i r_{i+1}}, a_i) = 1$ and $D_{r_i s_i} = (N_{s_i r_{i+1}}, a_i) = 1$. Since $M(\mathcal{A}) = DN$, this forces $p_i = s_i$. Thus $\sigma = \pi$ and this shows that $\gamma\varphi = \mathrm{Id}_E$. The fact that $\varphi\gamma = \mathrm{Id}_F$ is proved in the same way.

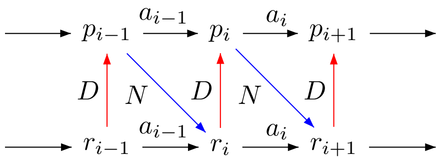


Figure 9. Conjugacy of automata.

□

Proof of Theorem 3.9. In one direction, the above statement is a direct consequence of the Decomposition Theorem 2.12. Indeed, if $\mathcal{A}$ and $\mathcal{B}$ are conjugate, there is a sequence $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_n$ of automata such that $\mathcal{A}_i$ is a split or a merge of $\mathcal{A}_{i+1}$ for $0 \leqslant i < n$ with $\mathcal{A}_0 = \mathcal{A}$ and $\mathcal{A}_n = \mathcal{B}$. The other direction follows from Proposition 3.10. □

4 Minimal automata

In this section, we define two notions of minimal automaton for sofic shifts: the Krieger automaton and the Fischer automaton. The first is defined for any sofic shift, and the second for irreducible ones.

The main result is that the Fischer automaton has the minimal number of states among all deterministic automata recognizing a given sofic shift (Proposition 4.6).

We then define the syntactic semigroup of a sofic shift, as an ordered semigroup. We show that this semigroup is isomorphic to the transition semigroup of the Krieger automaton and, for irreducible shifts, to the transition semigroup of the Fischer automaton (Proposition 4.8).

Minimal automata of sets of finite words. Recall that an automaton $\mathcal{A} = (Q, E)$ recognizes a shift X if $X = L_{\mathcal{A}}$. There should be no confusion with the notion of acceptance for sets of finite words in the usual sense: if $\mathcal{A}$ has an initial state i and a set of terminal states T, the set of finite words recognized by $\mathcal{A}$ is the set of labels of finite paths from i to a terminal state t in T. In this chapter[3], an automaton is called *deterministic* if, for each state p and each letter a, there is at most one edge starting in p and carrying the label a. We write, as usual, $p \cdot u$ for the unique end state, provided it exists, of a path starting in p and labeled u. For a set W of A^*, there exists a unique deterministic minimal automaton (this time with a unique initial state) recognizing W. Its states are the nonempty sets $u^{-1}W$ for $u \in A^*$, called the *right contexts* of u, and the edges are the triples $(u^{-1}W, a, (ua)^{-1}W)$, for $a \in A$ (see the chapter of J.-É. Pin).

Let $\mathcal{A} = (Q, E)$ be a finite automaton. For a state $p \in Q$, we denote by $L_p(\mathcal{A})$ or simply L_p the set of labels of finite paths starting from p. The automaton $\mathcal{A}$ is said to be *reduced* if $p \neq q$ implies $L_p \neq L_q$.

A word w is *synchronizing* for a deterministic automaton $\mathcal{A}$ if the set of paths labeled w is nonempty and all paths labeled w end in the same state. An automaton is *synchronized* if there is a synchronizing word. The following result holds because all states are terminal.

Proposition 4.1. *A reduced deterministic automaton is synchronized.*

Proof. Let $\mathcal{A} = (Q, E)$ be a reduced deterministic automaton. Given any word x, we denote by $Q \cdot X$ the set $Q \cdot x = \{q \cdot x \mid q \in Q\}$.

Let x be a word such that $Q \cdot x$ has minimal nonzero cardinality. Let p, q be two elements of the set $Q \cdot x$. If u is a word such that $p \cdot u$ is nonempty, then $q \cdot u$ is also nonempty since otherwise $Q \cdot xu$ would be of nonzero cardinality less than $Q \cdot x$. This implies that $L_p = L_q$ and thus $p = q$ since $\mathcal{A}$ is reduced. Thus x is synchronizing. □

4.1 Krieger automata and Fischer automata

Krieger automata. We denote by $A^{-\mathbb{N}}$ the set of left infinite words $x = \cdots x_{-1}x_0$. For $y = \cdots y_{-1}y_0 \in A^{-\mathbb{N}}$ and $z = z_0z_1 \cdots \in A^{\mathbb{N}}$, we denote by $y \cdot z = (w_i)_{i \in \mathbb{Z}}$ the biinfinite word defined by $w_i = y_{i+1}$ for $i < 0$ and $w_i = z_i$ for $i \geqslant 0$. Let X be a shift space. For

[3] This contrasts the more traditional definition which assumes in addition that there is a unique initial state.

$y \in A^{-\mathbb{N}}$, the set of *right contexts* of y is the set $C_X(y) = \{z \in A^{\mathbb{N}} \mid y \cdot z \in X\}$. For $u \in A^+$, we denote $u^\omega = uu\cdots$ and $u^{-\omega} = \cdots uu$.

The *Krieger automaton* of a shift space X is the deterministic automaton whose states are the nonempty sets of the form $C_X(y)$ for $y \in A^{-\mathbb{N}}$, and whose edges are the triples (p, a, q) where $p = C_X(y)$ for some left infinite word, $a \in A$ and $q = C_X(ya)$.

The definition of the Krieger automaton uses infinite words. One could use instead of the sets $C_X(y)$ for $y \in A^{-\mathbb{N}}$, the sets

$$D_X(y) = \{u \in A^* \mid \exists z \in A^{\mathbb{N}} : yuz \in X\}.$$

Indeed $C_X(y) = C_X(y')$ if and only if $D_X(y) = D_X(y')$. However, one cannot dispense completely with infinite words (see Proposition 4.2).

Example 4.1. Let $A = \{a, b\}$, and let $X = X^{(ba)}$. The Krieger automaton of X is represented in Figure 10. The states are the sets $1 = C_X(a^{-\omega}) = a^\omega \cup a^*b^\omega$ and $2 = C_X(a^{-\omega}b) = b^\omega$.

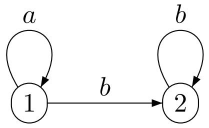


Figure 10. The Krieger automaton of $X^{(ba)}$.

Proposition 4.2. *The Krieger automaton of a shift space X is reduced and recognizes X. It is finite if and only if X is sofic.*

Proof. Let $\mathcal{A} = (Q, E)$ be the Krieger automaton of X. Let $p, q \in Q$ and let $y, z \in A^{-\mathbb{N}}$ be such that $p = C_X(y)$, $q = C_X(z)$. If $L_p = L_q$, then the labels of infinite paths starting from p and q are the same. Thus $p = q$. This shows that $\mathcal{A}$ is reduced. If $\mathcal{A}$ finite, then X is sofic by Proposition 3.3. Conversely, if X is sofic, let $\mathcal{A}$ be a finite automaton recognizing X. The set of right contexts of a left infinite word y only depends on the set of states p such that there is a path in the automaton $\mathcal{A}$ labeled y ending in state p. Thus the family of sets of right contexts is finite. □

We say that a deterministic automaton $\mathcal{A} = (Q, E)$ over the alphabet A is a *subautomaton* of a deterministic automaton $\mathcal{A}' = (Q', E')$ if $Q \subset Q'$ and if for each edge $(p, a, q) \in E$ such that $p \in Q$ one has $q \in Q$ and $(p, a, q) \in E'$.

The following proposition appears in [22] and in [11] where an algorithm to compute the states of the minimal automaton which are in the Krieger automaton is described.

Proposition 4.3. *The Krieger automaton of a sofic shift X is, up to an isomorphism, a subautomaton of the minimal automaton of the set of blocks of X.*

Proof. Let X be a sofic shift. Let $y \in A^{-\mathbb{N}}$ and set $y = \cdots y_{-1}y_0$ with $y_i \in A$ for $i \leqslant 0$. Set $u_i = y_{-i} \cdots y_0$ and $U_i = u_i^{-1}\mathcal{B}(X)$. Since $\mathcal{B}(X)$ is regular, the chain

$$\ldots \subset U_i \subset \ldots \subset U_1 \subset U_0$$

is stationary. Thus there is an integer $n \geqslant 0$ such that $U_{n+i} = U_n$ for all $i \geqslant 0$. We define $s(y) = U_n$.

We show that the map $C_X(y) \mapsto s(y)$ is well-defined and injective. Suppose first that $C_X(y) = C_X(y')$ for some $y, y' \in A^{-\mathbb{N}}$. Let $u \in A^*$ be such that $y_{-m} \cdots y_0 u \in \mathcal{B}(X)$ for all $m \geqslant n$. By compactness, there exists a $z \in A^{\mathbb{N}}$ such that $yuz \in X$. Then $y' \cdot uz \in X$ implies $u \in s(y')$. Symmetrically $u \in s(y')$ implies $u \in s(y)$. This shows that the map is well-defined.

To show that it is injective, consider $y, y' \in A^{-N}$ such that $s(y) = s(y')$. Let $z \in C_X(y)$. For each integer $m \geqslant 0$, we have $z_0 \cdots z_m \in s(y)$ and thus $z_0 \cdots z_m \in s(y')$. Since X is closed, this implies that $y' \cdot z \in X$ and thus $z \in C_X(y')$. The converse implication is proved in the same way. □

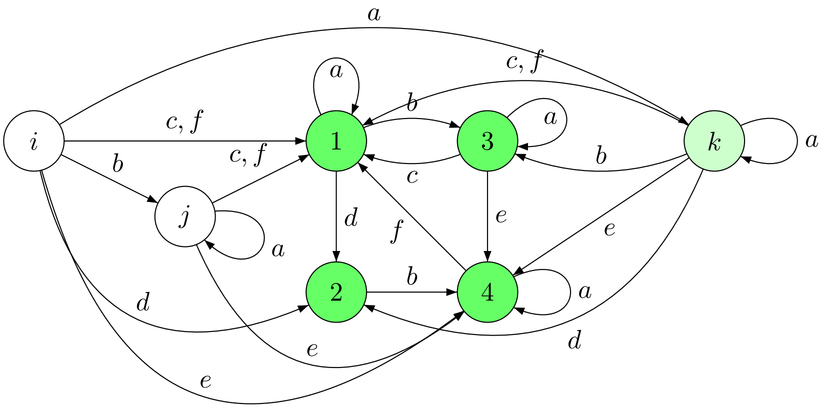


Figure 11. An example of Krieger automaton.

Example 4.2. Consider the automaton $\mathcal{A}$ on 7 states $i, j, k, 1, 2, 3, 4$ given in Figure 11. It can obtained, starting with the subautomaton over the states $1, 2, 3, 4$, using the subset construction computing the accessible nonempty sets of states, starting from the state $i = \{1, 2, 3, 4\}$ with $j = \{3, 4\}$, $k = \{1, 3, 4\}$.

The subautomaton with dark shaded states $1, 2, 3, 4$ is strongly connected and recognizes an irreducible sofic shift denoted by X. The whole automaton is the minimal automaton (with initial state i) of the set $\mathcal{B}(X)$. We identify each residual of $\mathcal{B}(X)$ with one of the 7 states of $\mathcal{A}$. Thus, for example, $i = \mathcal{B}(X)$ and $k = a^{-1}\mathcal{B}(X)$. The Krieger automaton of X is the automaton on the five shaded states. Indeed, with the map s defined in the proof of Proposition 4.3, there is no left infinite word y such that $s(y) = i$ or $s(y) = j$. On the contrary, since $i \cdot a = k$ and $k \cdot a = k$, one has $s(a^{-\omega}) = k$.

Fischer automata of irreducible shift spaces. A shift space $X \subset A^{\mathbb{Z}}$ is called *irreducible* if for any $u, v \in \mathcal{B}(X)$ there exists a $w \in \mathcal{B}(X)$ such that $uwv \in \mathcal{B}(X)$.

An automaton is said to be strongly connected if its underlying graph is strongly connected. Clearly a shift recognized by a strongly connected automaton is irreducible.

A strongly connected component of an automaton $\mathcal{A}$ is *minimal* if all successors of vertices of the component are themselves in the component. One may verify that a minimal strongly connected component is the same as a strongly connected subautomaton.

The following result is due to Fischer [13] (see also [19, Section 3]). It implies in particular that an irreducible sofic shift can be recognized by a strongly connected automaton.

Proposition 4.4. *The Krieger automaton of an irreducible sofic shift X is synchronized and has a unique minimal strongly connected component.*

Proof. Let $\mathcal{A} = (Q, E)$ be the Krieger automaton of X. By Proposition 4.2, $\mathcal{A}$ is reduced and by Proposition 4.1, it follows that it is synchronized.

Let x be a synchronizing word. Let R be the set of states reachable from the state $q = Q \cdot x$. The set R is a minimal strongly connected component of $\mathcal{A}$. Indeed, for any $r \in R$ there is a path $q \xrightarrow{y} r$. Since X is irreducible there is a word z such that $yzx \in \mathcal{B}(X)$. Since $q \cdot yzx = q$, r belongs to the same strongly connected component as q. Next, if p belongs to a minimal strongly connected component S of $\mathcal{A}$, since X is irreducible, there is a word y such that $p \cdot yx$ is not empty. Thus q is in S, which implies $S = R$. Thus R is the only minimal strongly component of $\mathcal{A}$. □

Example 4.3. Let X be the even shift. The Krieger and Fischer automata of X are represented on Figure 12. The word a is synchronizing.

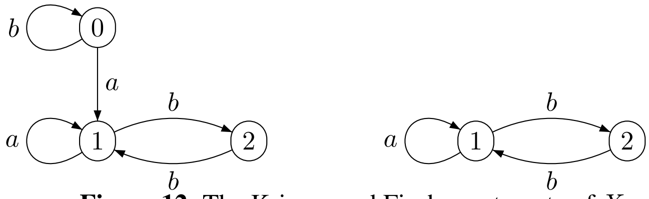


Figure 12. The Krieger and Fischer automata of X.

Example 4.4. The Fischer automaton of the irreducible shift of Example 4.2 is the subautomaton on states $1, 2, 3, 4$ represented with dark shaded states in Figure 11.

Let X be an irreducible sofic shift X. The minimal strongly connected component of the Krieger automaton of X is called its *Fischer automaton*.

Proposition 4.5. *The Fischer automaton of an irreducible sofic shift X recognizes X.*

Proof. The Fischer automaton $\mathcal{F}$ of X is a subautomaton of the Krieger automaton of X which in turn is a subautomaton of the minimal automaton $\mathcal{A}$ of the set $\mathcal{B}(X)$. Let i be the initial state of $\mathcal{A}$. Since $\mathcal{A}$ is trim, there is a word w such that $i \cdot w$ is a state of $\mathcal{F}$. Let v be any block of X. Since X is irreducible, there is a word u such that wuv is a block of X. This shows that v is a label of a path in $\mathcal{F}$. Thus every block of X is a label of a path in $\mathcal{F}$ and conversely. In view of Proposition 3.2, the automaton $\mathcal{F}$ recognizes X. □

Let $\mathcal{A} = (Q, E)$ and $B = (R, F)$ be two deterministic automata. A *reduction* from $\mathcal{A}$ onto $\mathcal{B}$ is a map h from Q onto R such that for any letter $a \in A$, one has $(p, a, q) \in E$ if and only if $(h(p), a, h(q)) \in F$. Thus any labeled in or out-merge is a reduction. However the converse is not true since a reduction is not, in general, a conjugacy.

For any automaton $\mathcal{A} = (Q, E)$, there is reduction from $\mathcal{A}$ onto a reduced automaton $\mathcal{B}$. It is obtained by identifying the pairs of states $p, q \in Q$ such that $L_p = L_q$.

The following statement is Corollary 3.3.20 of [19].

Proposition 4.6. *Let X be an irreducible shift space. For any strongly connected deterministic automaton $\mathcal{A}$ recognizing X there is a reduction from $\mathcal{A}$ onto the Fischer automaton of X.*

Proof. Let $\mathcal{A} = (Q, E)$ be a strongly connected automaton recognizing X. Let $\mathcal{B} = (R, F)$ be the reduced automaton obtained from $\mathcal{A}$ identifying the pairs $p, q \in Q$ such that $L_p = L_q$. By Proposition 4.1, $\mathcal{B}$ is synchronized.

We now show that $\mathcal{B}$ can be identified with the Fischer automaton of X. Let w be a synchronizing word for $\mathcal{B}$. Set $s = Q \cdot w$. Let r be a state such that $r \cdot w = s$. and let $y \in A^{-\mathbb{N}}$ be the label of a left infinite path ending in the state s. For any state t in R, let u be a word such that $s \cdot u = t$. The set $C_X(ywu)$ depends only on the state t, and not on the word u such that $s \cdot u = t$. Indeed, for each right infinite word z, one has $ywuz$ in X if and only if there is a path labeled z starting at t. This holds because w is synchronizing.

Thus the map $t \mapsto C_X(ywu)$ is well-defined and defines a reduction from $\mathcal{B}$ onto the Fischer automaton of X. □

This statement shows that the Fischer automaton of an irreducible shift X is minimal in the sense that it has the minimal number of states among all deterministic strongly connected automata recognizing X.

The statement also gives the following practical method to compute the Fischer automaton of an irreducible shift. We start with a strongly connected deterministic automaton recognizing X and merge the pairs of states p, q such that $L_p = L_q$. By the above result, the resulting automaton is the Fischer automaton of X.

4.2 Syntactic semigroup

Recall that a *preorder* on a set is a relation which is reflexive and transitive. The equivalence associated to a preorder is the equivalence relation defined by $u \equiv v$ if and only if $u \leqslant v$ and $v \leqslant u$.

Let S be a semigroup. A preorder on S is said to be *stable* if $s \leqslant s'$ implies $us \leqslant us'$ and $su \leqslant s'u$ for all $s, s', u \in S$. An *ordered semigroup* S is a semigroup equipped with a stable preorder. Any semigroup can be considered as an ordered semigroup equipped with the equality order.

A *congruence* in an ordered semigroup S is the equivalence associated to a stable preorder which is coarser than the preorder of S. The quotient of an ordered semigroup by a congruence is the ordered semigroup formed by the classes of the congruence.

The *set of contexts* of a word u with respect to a set $W \subset A^+$ is the set $\Gamma_W(u)$ of pairs of words defined by $\Gamma_W(u) = \{(\ell, r) \in A^* \times A^* \mid \ell ur \in W\}$. The preorder on

A^+ defined by $u \leqslant_W v$ if $\Gamma_W(u) \subset \Gamma_W(v)$ is stable and thus defines a congruence of the semigroup A^+ equipped with the equality order called the *syntactic congruence*. The *syntactic semigroup* of a set $W \subset A^*$ is the quotient of the semigroup A^+ by the syntactic congruence.

Let $\mathcal{A} = (Q, E)$ be a deterministic automaton on the alphabet A. Recall that for $p \in Q$ and $u \in A^+$, there is at most one path π labeled u starting in p. We set $p \cdot u = q$ if q is the end of π and $p \cdot u = \emptyset$ if π does not exist. The preorder defined on A^+ by $u \leqslant_{\mathcal{A}} v$ if $p \cdot u \subset p \cdot v$ for all $p \in Q$ is stable. The quotient of A^+ by the congruence associated to this preorder is the *transition semigroup* of $\mathcal{A}$.

The following property is standard, see the chapter of J.-É Pin.

Proposition 4.7. *The syntactic semigroup of a set $W \subset A^+$ is isomorphic to the transition semigroup of the minimal automaton of W.*

The *syntactic semigroup* of a shift space X is by definition the syntactic semigroup of $\mathcal{B}(X)$.

Proposition 4.8. *Let X be a sofic shift and let S be its syntactic semigroup. The transition semigroup of the Krieger automaton of X is isomorphic to S. Moreover, if X is irreducible, then it is isomorphic to the transition semigroup of its Fischer automaton.*

Proof. Let $\mathcal{A}$ be the minimal automaton of $\mathcal{B}(X)$, and let $\mathcal{K}$ be the Krieger automaton of X. We have to show that for any $u, v \in A^+$, one has $u \leqslant_{\mathcal{A}} v$ if and only if $u \leqslant_{\mathcal{K}} v$. Since, by Proposition 4.3, $\mathcal{K}$ is isomorphic to a subautomaton of $\mathcal{A}$, the direct implication is clear. Indeed, if p is a state of $\mathcal{K}$, then $L_p(\mathcal{K})$ is equal to the set $L_p(\mathcal{A})$. Consequently, if $u \leqslant_{\mathcal{A}} v$ then $u \leqslant_{\mathcal{K}} v$. Conversely, suppose that $u \leqslant_{\mathcal{K}} v$. We prove that $u \leqslant_{\mathcal{B}(X)} v$. For this, let $(\ell, r) \in \Gamma_{\mathcal{B}(X)}(u)$. Then $\ell ur \in \mathcal{B}(X)$. Then $y \cdot \ell urz \in X$ for some $y \in A^{-\mathbb{N}}$ and $z \in A^{\mathbb{N}}$. But since $C_X(y\ell u) \subset C_X(y\ell v)$, this implies $rz \in C_X(y\ell v)$ and thus $\ell vr \in \mathcal{B}(X)$. Thus $u \leqslant_{\mathcal{B}(X)} v$ which implies $u \leqslant_{\mathcal{A}} v$.

Next, suppose that X is irreducible. We have to show that $u \leqslant_{\mathcal{A}} v$ if and only if $u \leqslant_{\mathcal{F}(X)} v$. Since $\mathcal{F}(X)$ is a subautomaton of $\mathcal{K}(X)$ and $\mathcal{K}(X)$ is a subautomaton of $\mathcal{A}$, the direct implication is clear. Conversely, assume that $u \leqslant_{\mathcal{F}(X)} v$. Suppose that $\ell ur \in \mathcal{B}(X)$. Let i be the initial state of $\mathcal{A}$ and let w be such that $i \cdot w$ is a state of $\mathcal{F}(X)$. Since X is irreducible, there is a word s such that $ws\ell ur \in \mathcal{B}(X)$. But then $i \cdot ws\ell ur \neq \emptyset$ implies $i \cdot ws\ell vr \neq \emptyset$. Thus $\ell vr \in \mathcal{B}(X)$. This shows that $u \leqslant_{\mathcal{B}(X)} v$ and thus $u \leqslant_{\mathcal{A}} v$. □

5 Symbolic conjugacy

This section is concerned with a new notion of conjugacy between automata called symbolic conjugacy. It extends the notion of labeled conjugacy and captures the fact that the automata may be over different alphabets. The table below summarizes the various notions.

object type	isomorphism	elementary transformation
shift spaces	conjugacy	split/merge
edge shifts	conjugacy	edge split/merge
integer matrices	strong shift equivalence	elementary equivalence
automata (same alphabet)	labeled conjugacy	labeled split/merge
automata	symbolic conjugacy	split/merge
alphabetic matrices	symbolic strong shift	elementary symbolic

There are two main results in this section. Theorem 5.7 due to Nasu is a version of the Classification Theorem for sofic shifts. It implies in particular that conjugate sofic shifts have symbolic conjugate Krieger or Fisher automata.The proof uses the notion of bipartite automaton, which corresponds to the symbolic elementary equivalence of adjacency matrices. Theorem 5.8 is due to Hamachi and Nasu: it characterizes symbolic conjugate automata by means of their adjacency matrices.

In this section, we will use for convenience automata in which several edges with the same source and target can have the same label. Formally, such an automaton is a pair $\mathcal{A} = (G, \lambda)$ of a graph $G = (Q, \mathcal{E})$ and a map assigning to each edge $e \in \mathcal{E}$ of a label $\lambda(e) \in A$. The adjacency matrix of $\mathcal{A}$ is the $Q \times Q$-matrix $M(\mathcal{A})$ with elements in $\mathbb{N}\langle A\rangle$ defined by

$$(M(\mathcal{A})_{pq}, a) = \mathrm{Card}\{e \in \mathcal{E} \mid \lambda(e) = a\}. \tag{5.1}$$

Note that $M(\mathcal{A})$ is alphabetic but may have arbitrary nonnegative coefficients. The advantage of this version of automata is that for any alphabetic $Q \times Q$-matrix M there is an automaton $\mathcal{A}$ such that $M(\mathcal{A}) = M$.

We still denote by $X_{\mathcal{A}}$ the edge shift X_G and by $L_{\mathcal{A}}$ the set of labels of infinite paths in G.

Symbolic conjugate automata. Let $\mathcal{A}, \mathcal{B}$ be two automata. A *symbolic conjugacy* from $\mathcal{A}$ onto $\mathcal{B}$ is a pair (φ, ψ) of conjugacies $\varphi : X_{\mathcal{A}} \to X_{\mathcal{B}}$ and $\psi : L_{\mathcal{A}} \to L_{\mathcal{B}}$ such that the following diagram is commutative.

$$\begin{tikzcd}
X_{\mathcal{A}} \arrow[r, "\varphi"] \arrow[d, "\lambda_{\mathcal{A}}"] & X_{\mathcal{B}} \arrow[d, "\lambda_{\mathcal{B}}"] \\
L_{\mathcal{A}} \arrow[r, "\psi"] & L_{\mathcal{B}}
\end{tikzcd}$$

5.1 Splitting and merging maps

Let A, B be two alphabets and let $f : A \to B$ be a map from A onto B. Let X be a shift space on the alphabet A. We consider the set of words $A' = \{f(a_1)a_2 \mid a_1a_2 \in \mathcal{B}_2(X)\}$ as a new alphabet. Let $g : \mathcal{B}_2(X) \to A'$ be the 2-block substitution defined by $g(a_1a_2) = f(a_1)a_2$.

The *in-splitting map* defined on X and relative to f or to g is the sliding block map $g_\infty^{1,0}$ corresponding to g. It is a conjugacy from X onto its image by $X' = g_\infty^{1,0}(X)$ since its inverse is 1-block. The shift space X', is called the *in-splitting* of X, relative to f or g. The inverse of an in-splitting map is called an *in-merging map*.

In addition, any renaming of the alphabet of a shift space is also considered to be an in-splitting map (and an in-merging map).

Example 5.1. Let $A = B$ and let f be the identity on A. The out-splitting of a shift X relative to f is the second higher block shift of X.

The following proposition relates splitting maps to edge splittings as defined in Section 2.2.

Proposition 5.1. *An in-splitting map on an edge shift is an edge in-splitting map, and conversely.*

Proof. Let first $G = (Q, \mathcal{E})$ be a graph, and let $f : \mathcal{E} \to I$ be a map from $\mathcal{E}$ onto a set I. Set $\mathcal{E}' = \{f(e_1)e_2 \mid e_1e_2 \in \mathcal{B}_2(X_G)\}$. Let $g : \mathcal{B}_2(X_G) \to \mathcal{E}'$ be the 2-block substitution defined by $g(e_1e_2) = f(e_1)e_2$. Let $G' = (Q', \mathcal{E}')$ be the graph on the set of states $Q' = I \times Q$ defined for $e' = f(e_1)e_2$ by $i(e') = (f(e_1), i(e_2))$ and $t(e') = (f(e_2), t(e_2))$. Define $h : \mathcal{E}' \to \mathcal{E}$ and $k : Q' \to Q$ by $h(f(e_1)e_2) = e_2$ for $e_1e_2 \in \mathcal{B}_2(X_G)$ and $k(i, q) = q$ for $(i, q) \in I \times Q$. Then the pair (h, k) is an in-merge from G' onto G and h_∞ is the inverse of $g_\infty^{1,0}$. Indeed, one may verify that (h, k) is a graph morphism from G' onto G. Next it is an in-merge because for each $p, q \in Q$, the partition $(\mathcal{E}_p^q(t))_{t\in k^{-1}(q)}$ of $\mathcal{E}_p^q$ is defined by $\mathcal{E}_p^q(i, q) = E_p^q \cap f^{-1}(i)$.

Conversely, set $G = (Q, \mathcal{E})$ and $G' = (Q', \mathcal{E}')$. Let (h, k) be an in-merge from G' onto G. Consider the map $f : \mathcal{E} \to Q'$ defined by $f(e) = r$ if r is the common end of the edges in $h^{-1}(e)$. The map α from $\mathcal{E}'$ to $Q' \times \mathcal{E}$ defined by $\alpha(i) = (r, h(i))$ where r is the origin of i is a bijection by definition of an in-merge.

Let us show that, up to the bijection α, the in-splitting map relative to f is inverse of the map h_∞. For $e_1, e_2 \in \mathcal{E}$, let $r = f(e_1)$ and $e' = \alpha^{-1}(r, e_2)$. Then $h(e') = e_2$ and thus h_∞ is the inverse of the map $g_\infty^{1,0}$ corresponding to the 2-block substitution $g(e_1e_2) = (r, e_2)$.

□

Symmetrically an *out-splitting map* is defined by the substitution $g(ab) = af(b)$. Its inverse is an out-merging map.

We use the term splitting to mean either a in-splitting or out-splitting. The same convention holds for a merging.

The following result, from [21], is a generalization of the Decomposition Theorem (Theorem 2.12) to arbitrary shift spaces.

Theorem 5.2. *Any conjugacy between shift spaces is a composition of splitting and merging maps.*

The proof is similar to the proof of Theorem 2.12. It relies on the following lemma, similar to Lemma 2.13.

Lemma 5.3. *Let $\varphi : X \to Y$ be a 1-block conjugacy whose inverse has memory $m \geqslant 1$ and anticipation $n \geqslant 0$. There are in-splitting maps from X, Y to $\tilde{X}, \tilde{Y}$ respectively such that the 1-block conjugacy $\tilde{\varphi}$ making the diagram below commutative has an inverse with memory $m - 1$ and anticipation n.*

$$\begin{tikzcd}
X \arrow[r] \arrow[d, "\varphi"] & \tilde{X} \arrow[d, "\tilde{\varphi}"] \\
Y \arrow[r] & \tilde{Y}
\end{tikzcd}$$

Proof. Let A, B the alphabets of X and Y respectively. Let $h : A \to B$ be the 1-block substitution such that $\varphi = h_\infty$. Let $\tilde{X}$ be the in-splitting of X relative to the map h. Set $A' = \{h(a_1)a_2 \mid a_1a_2 \in \mathcal{B}_2(X)\}$. Let $\tilde{Y} = Y^{[2]}$ be the second higher block shift of Y and let $B' = \mathcal{B}_2(Y)$. Let $\tilde{h} : A' \to B'$ be the 1-block substitution defined by $\tilde{h}(h(a_1)a_2) = h(a_1)h(a_2)$. Then the 1-block map $\tilde{\varphi} = \tilde{h}_\infty$ has the required properties. □

Lemma 5.3 has a dual where φ is a 1-block map whose inverse has memory $m \geqslant 0$ and anticipation $n \geqslant 1$ and where in-splits are replaced by out-splits.

Proof of Theorem 5.2. Let $\varphi : X \to Y$ be a conjugacy from X onto Y. Replacing X by a higher block shift, we may assume that φ is a 1-block map. Using iteratively Lemma 5.3, we can replace φ by a 1-block map whose inverse has memory 0. Using then iteratively the dual of Lemma 5.3, we finally obtain a 1-block map whose inverse is also 1-block and is thus just a renaming of the symbols. □

Symbolic strong shift equivalence. Let M and M' be two alphabetic $Q \times Q$-matrices over the alphabets A and B, respectively. We say that M and M' are *similar* if they are equal up to a bijection of A onto B. We write $M \leftrightarrow M'$ when M and M' are similar. We say that two alphabetic square matrices M and M' over the alphabets A and B respectively are *symbolic elementary equivalent* if there exist two alphabetic matrices R, S over the alphabets C and D respectively such that

$$M \leftrightarrow RS, \quad M' \leftrightarrow SR.$$

In this definition, the sets CD and DC of two letter words are identified with alphabets in bijection with A and B, respectively.

We say that two matrices M, M' are *symbolic strong shift equivalent* if there is a sequence $(M_0, M_1, \ldots, M_n)$ of alphabetic matrices such that M_i and M_{i+1} are symbolic elementary equivalent for $0 \leqslant i < n$ with $M_0 = M$ and $M_n = M'$.

We introduce the following notion. An automaton $\mathcal{A}$ on the alphabet A is said to be *bipartite* if there are partitions $Q = Q_1 \cup Q_2$ of the set of states and $A = A_1 \cup A_2$ of the alphabet such that all edges labeled in A_1 go from Q_1 to Q_2 and all edges labeled in A_2 go from Q_2 to Q_1.

Let $\mathcal{A}$ be a bipartite automaton. Its adjacency matrix has the form

$$M(\mathcal{A}) = \begin{bmatrix} 0 & M_1 \\ M_2 & 0 \end{bmatrix}$$

where M_1 is a $Q_1 \times Q_2$-matrix with elements in $\mathbb{N}\langle A_1\rangle$ and M_2 is a $Q_2 \times Q_1$-matrix with elements in $\mathbb{N}\langle A_2\rangle$ The automata $\mathcal{A}_1$ and $\mathcal{A}_2$ which have M_1M_2 and M_2M_1 respectively as adjacency matrix are called the *components* of $\mathcal{A}$ and the pair $\mathcal{A}_1, \mathcal{A}_2$ is a *decomposition* of $\mathcal{A}$. We denote $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ a bipartite automaton $\mathcal{A}$ with components $\mathcal{A}_1, \mathcal{A}_2$. Note that $\mathcal{A}_1, \mathcal{A}_2$ are automata on the alphabets A_1A_2 and A_2A_1 respectively.

Proposition 5.4. *Let $\mathcal{A} = (Q, E)$ be a bipartite deterministic essential automaton. Its components $\mathcal{A}_1, \mathcal{A}_2$ are deterministic essential automata which are symbolic conjugate. If moreover $\mathcal{A}$ is strongly connected (resp. reduced, resp. synchronized), then $\mathcal{A}_1, \mathcal{A}_2$ are strongly connected (resp.reduced, resp. synchronized).*

Proof. Let $Q = Q_1 \cup Q_2$ and $A = A_1 \cup A_2$ be the partitions of the set Q and the alphabet A corresponding to the decomposition $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$. It is clear that $\mathcal{A}_1, \mathcal{A}_2$ are deterministic and that they are strongly connected if $\mathcal{A}$ is strongly connected.

Let $\varphi : X_{\mathcal{A}_1} \to X_{\mathcal{A}_2}$ be the conjugacy defined as follows. For any $y = (y_n)_{n\in\mathbb{Z}}$ in $X_{\mathcal{A}_1}$ there is an $x = (x_n)_{n\in\mathbb{Z}}$ in $X_{\mathcal{A}}$ such that $y_n = x_{2n}x_{2n+1}$. Then $z = (z_n)_{n\in\mathbb{Z}}$ with $z_n = x_{2n+1}x_{2n}$ is an element of $X_{\mathcal{A}_2}$. We define $\varphi(y) = z$. The analogous map $\psi : L_{\mathcal{A}_1} \to L_{\mathcal{A}_2}$ is such that (φ, ψ) is a symbolic conjugacy from $\mathcal{A}_1$ onto $\mathcal{A}_2$.

Assume that $\mathcal{A}$ is reduced. For $p, q \in Q_1$, there is a word w such that $w \in L_p(\mathcal{A})$ and $w \notin L_q(\mathcal{A})$ (or conversely). Set $w = a_1a_2\cdots a_n$ with $a_i \in A$. If n is even, then $(a_1a_2)\cdots(a_{n-1}a_n)$ is in $L_p(\mathcal{A}_1)$ but not in $L_q(\mathcal{A}_1)$. Otherwise, since $\mathcal{A}$ is essential, there is a letter a_{n+1} such that wa_{n+1} is in $L_p(\mathcal{A})$. Then $(a_1a_2)\cdots(a_na_{n+1})$ is in $L_p(\mathcal{A}_1)$ but not in $L_q(\mathcal{A}_1)$. Thus $\mathcal{A}_1$ is reduced. One proves in the same way that $\mathcal{A}_2$ is reduced.

Suppose finally that $\mathcal{A}$ is synchronized. Let x be a synchronizing word and set $x = a_1a_2\cdots a_n$ with $a_i \in A$. Suppose that all paths labeled x end in $q \in Q_1$. Let a_{n+1} be a letter such that $q \cdot a_{n+1} \neq \emptyset$ and let a_0 be a letter such that a_0x is the label of at least one path. If n is even, then $(a_1a_2)\cdots(a_{n-1}a_n)$ is synchronizing for $\mathcal{A}_1$ and $(a_0a_1)\cdots(a_na_{n+1})$ is synchronizing for $\mathcal{A}_2$. Otherwise, $(a_0a_1)\cdots(a_{n-1}a_n)$ is synchronizing for $\mathcal{A}_1$ and $(a_1a_2)\cdots(a_na_{n+1})$ is synchronizing for $\mathcal{A}_2$. □

Proposition 5.5. *Let $\mathcal{A}, \mathcal{B}$ be two automata such that $M(\mathcal{A})$ and $M(\mathcal{B})$ are symbolic elementary equivalent. Then there is a bipartite automaton $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2)$ such that $M(\mathcal{C}_1), M(\mathcal{C}_2)$ are similar to $M(\mathcal{A}), M(\mathcal{B})$ respectively.*

Proof. Let R, S be alphabetic matrices over alphabets C and D respectively such that $M(\mathcal{A}) \leftrightarrow RS$ and $M(\mathcal{B}) \leftrightarrow SR$. Let $\mathcal{C}$ be the bipartite automaton on the alphabet $C \cup D$ which is defined by the adjacency matrix

$$M(\mathcal{C}) = \begin{bmatrix} 0 & R \\ S & 0 \end{bmatrix}$$

Then $M(\mathcal{A})$ is similar to $M(\mathcal{C}_1)$ and $M(\mathcal{B})$ is similar to $M(\mathcal{C}_2)$. □

Proposition 5.6. *If the adjacency matrices of two automata are symbolic strong shift equivalent, the automata are symbolic conjugate.*

Proof. Since a composition of conjugacies is a conjugacy, it is enough to consider the case where the adjacency matrices are symbolic elementary equivalent. Let $\mathcal{A}, \mathcal{B}$ be such that $M(\mathcal{A}), M(\mathcal{B})$ are symbolic elementary equivalent. By Proposition 5.5, there is a bipartite automaton $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2)$ such that $M(\mathcal{C}_1), M(\mathcal{C}_2)$ are similar to $M(\mathcal{A})$ and $M(\mathcal{B})$ respectively. By Proposition 5.4, the automata $\mathcal{C}_1, \mathcal{C}_2$ are symbolic conjugate. Since automata with similar adjacency matrices are obviously symbolic conjugate, the result follows. □

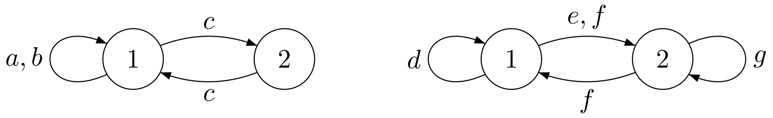


Figure 13. Two symbolic conjugate automata.

Example 5.2. Let $\mathcal{A}, \mathcal{B}$ be the automata represented on Figure 13. The matrices $M(\mathcal{A})$ and $M(\mathcal{B})$ are symbolic elementary equivalent. Indeed, we have $M(\mathcal{A}) \leftrightarrow RS$ and $M(\mathcal{B}) \leftrightarrow SR$ for

$$R = \begin{bmatrix} x & y \\ 0 & x \end{bmatrix}, \quad S = \begin{bmatrix} z & t \\ t & 0 \end{bmatrix}.$$

Indeed, one has

$$RS = \begin{bmatrix} xz + yt & xt \\ xt & 0 \end{bmatrix}, \quad SR = \begin{bmatrix} zx & zy + tx \\ tx & ty \end{bmatrix}.$$

Thus the following tables give two bijections between the alphabets.

a	b	c
xz	yt	xt

,

d	e	f	g
zx	zy	tx	ty

.

The following result is due to Nasu [21]. The equivalence between conditions (i) and (ii) is a version, for sofic shifts, of the Classification Theorem (Theorem 7.2.12 in [19]). The equivalence between conditions (i) and (iii) is due to Krieger [18].

Theorem 5.7. *Let X, X' be two sofic shifts (resp. irreducible sofic shifts) and let $\mathcal{A}, \mathcal{A}'$ be their Krieger (resp. Fischer) automata. The following conditions are equivalent.*

(i) X, X' are conjugate.
(ii) The adjacency matrices of $\mathcal{A}, \mathcal{A}'$ are symbolic strong shift equivalent.
(iii) $\mathcal{A}, \mathcal{A}'$ are symbolic conjugate.

Proof. We prove the result for irreducible shifts. The proof of the general case is in [21].

Assume that X, X' are conjugate. By the Decomposition Theorem (Theorem 5.2), it is enough to consider the case where X' is an in-splitting of X. Let $f : A \to B$ be a map and let $A' = \{f(a_1)a_2 \mid a_1a_2 \in \mathcal{B}_2(X)\}$ in such a way that X' is the in-splitting of X relative to f. Let $C = A \cup B$ and let Z be the shift space composed of all biinfinite sequences $\cdots a_i f(a_i) a_{i+1} f(a_{i+1}) \cdots$ such that $\cdots a_i a_{i+1} \cdots$ is in X. Then Z is an irreducible sofic shift. Let $\mathcal{A}$ be the Fischer automaton of Z. Then $\mathcal{A}$ is bipartite and its components recognize, up to a bijection of the alphabets, X and X' respectively. By Proposition 5.4 the components are the Fischer automata of X and X' respectively. Since the components of a bipartite automaton have symbolic elementary equivalent adjacency matrices, this proves that (i) implies (ii).

That (ii) implies (iii) is Proposition 5.6. Finally, (iii) implies (i) by definition of symbolic conjugacy. □

5.2 Symbolic conjugate automata

The following result is due to Hamachi and Nasu [16]. It shows that, in Theorem 5.7, the equivalence between conditions (ii) and (iii) holds for automata which are not reduced.

Theorem 5.8. *Two essential automata are symbolic conjugate if and only if their adjacency matrices are symbolic strong shift equivalent.*

The first element of the proof is a version of the Decomposition Theorem for automata.

Let $\mathcal{A}, \mathcal{A}'$ be two automata. An *in-split* from $\mathcal{A}$ onto $\mathcal{A}'$ is a symbolic conjugacy (φ, ψ) such that $\varphi : X_{\mathcal{A}} \to X_{\mathcal{A}'}$ and $\psi : L_{\mathcal{A}} \to L_{\mathcal{A}'}$ are in-splitting maps. A similar definition holds for out-splits.

Theorem 5.9. *Any symbolic conjugacy between automata is a composition of splits and merges.*

The proof relies on the following variant of Lemma 5.3.

Lemma 5.10. *Let α, β be 1-block maps and φ, ψ be 1-block conjugacies such such that the diagram below on the left is commutative.*

If the inverses of φ, ψ have memory $m \geqslant 1$ and anticipation $n \geqslant 0$, there exist in-splits $\tilde{X}, \tilde{Y}, \tilde{Z}, \tilde{T}$ of X, Y, Z, T respectively and 1-block maps $\tilde{\alpha} : \tilde{X} \to \tilde{Z}$, $\tilde{\beta} : \tilde{Y} \to \tilde{T}$ such that the 1-block conjugacies $\tilde{\varphi}, \tilde{\psi}$ making the diagram below on the right commutative have inverses with memory $m - 1$ and anticipation n.

$$\begin{tikzcd}
X \arrow[r, "\varphi"] \arrow[d, "\alpha"] & Y \arrow[d, "\beta"] \\
Z \arrow[r, "\psi"] & T
\end{tikzcd}
\qquad
\begin{tikzcd}
X \arrow[rrr, "\varphi"] \arrow[dr] \arrow[ddd, "\alpha"] & & & Y \arrow[ddd, "\beta"] \\
& \tilde{X} \arrow[r, "\tilde{\varphi}"] \arrow[d, "\tilde{\alpha}"] & \tilde{Y} \arrow[ur] \arrow[d, "\tilde{\beta}"] & \\
& \tilde{Z} \arrow[r, "\tilde{\psi}"] & \tilde{T} \arrow[dr] & \\
Z \arrow[ur] \arrow[rrr, "\psi"] & & & T
\end{tikzcd}$$

Proof. Let A, B, C, D be the alphabets of X, Y, Z and T respectively. Let $h : A \to B$ and $k : C \to D$ be the 1-block substitutions such that $\varphi = h_\infty$ and $\psi = k_\infty$. Set

$\tilde{A} = \{h(a_1)a_2 \mid a_1a_2 \in \mathcal{B}_2(X)\}$ and $\tilde{C} = \{k(c_1)c_2 \mid c_1c_2 \in \mathcal{B}_2(Z)\}$. Let $\tilde{X}$ (resp. $\tilde{Z}$) be the image of X (resp. of Z) under the in-splitting map relative to h (resp. k). Set $\tilde{Y} = Y^{[2]}$, $\tilde{B} = \mathcal{B}_2(Y)$, $\tilde{T} = T^{[2]}$ and $\tilde{D} = \mathcal{B}_2(T)$. Define $\tilde{\alpha}$ and $\tilde{\beta}$ by

$$\tilde{\alpha}(h(a_1)a_2) = k\alpha(a_1)\alpha(a_2), \quad \tilde{\beta}(b_1b_2) = \beta(b_1)\beta(b_2)$$

and $\tilde{h} : \tilde{A} \to \tilde{B}$, $\tilde{k} : \tilde{C} \to \tilde{D}$ by

$$\tilde{h}(h(a_1)a_2) = h(a_1)h(a_2), \quad \tilde{k}(k(c_1)c_2) = k(c_1)k(c_2)$$

Then the 1-block conjugacies $\tilde{\varphi} = \tilde{h}_\infty$ and $\tilde{\psi} = \tilde{k}_\infty$ satisfy the conditions of the statement. □

Proof of Theorem 5.9. Let $\mathcal{A} = (G, \lambda)$ and $\mathcal{A}' = (G', \lambda')$ be two automata with $G = (Q, \mathcal{E})$ and $G' = (Q', \mathcal{E}')$. Let (φ, ψ) be a symbolic conjugacy from $\mathcal{A}$ onto $\mathcal{A}'$. Replacing $\mathcal{A}$ and $\mathcal{B}$ by some extension $\mathcal{A}^{[m,n]}$ and $\mathcal{B}^{[m,n]}$ we may reduce to the case where φ, ψ are 1-block conjugacies. By using repeatedly Lemma 5.10, we may reduce to the case where the inverses of φ, ψ have memory 0. Using repeatedly the dual version of Lemma 5.10, we are reduced to the case where φ, ψ are renaming of the alphabets. □

The second step for the proof of Theorem 5.8 is the following statement.

Proposition 5.11. *Let $\mathcal{A}, \mathcal{A}'$ be two essential automata. If $\mathcal{A}'$ is an in-split of $\mathcal{A}$, the matrices $M(\mathcal{A})$ and $M(\mathcal{A}')$ are symbolic elementary equivalent.*

Proof. Set $\mathcal{A} = (G, \lambda)$ and $\mathcal{A}' = (G', \lambda')$. Let $A' = \{f(a)b \mid ab \in \mathcal{B}_2(L_{\mathcal{A}})\}$ be the alphabet of $\mathcal{A}'$ for a map $f : A \to B$. By Proposition 5.1, the symbolic in-splitting map from X_G onto $X_{G'}$ is also an in-splitting map. Thus there is an in-merge (h, k) from G' onto G such that the in-split from $\mathcal{A}$ onto $\mathcal{A}'$ has the form (h_∞^{-1}, ψ). We define an alphabetic $Q' \times Q$-matrix R and a $Q \times Q'$-matrix S as follows. Let $r, t \in Q'$ and let $p = k(r)$, $q = k(t)$. Let e be an edge of $\mathcal{A}'$ ending in r, and set $a = \lambda(h(e))$. Then the label of any edge going out of r is of the form $f(a)b$ for some $b \in A$. Thus $f(a)$ does not depend on e but only on r. We define a map $\pi : Q' \to B$ by $\pi(r) = f(a)$. Then, we set

$$R_{rp} = \begin{cases} \pi(r) & \text{if } k(r) = p \\ 0 & \text{otherwise} \end{cases}, \quad S_{pt} = M(\mathcal{A})_{pq}$$

Let us verify that $M(\mathcal{A}') = RS$ and $M(\mathcal{A}) \leftrightarrow SR$. We first have for $r, t \in Q'$

$$(RS)_{rt} = \sum_{p \in Q} R_{rp} S_{pt} = \pi(r) M_{k(r)k(q)} = M(\mathcal{A}')_{rt}$$

and thus $RS = M(\mathcal{A}')$. Next, for $p, q \in Q$

$$(SR)_{pq} = \sum_{p \in Q} R_{rp} S_{pt} = \sum_{t \in k^{-1}(q)} M(\mathcal{A})_{pq} \pi(t) = \sum_{a \in A} (M(\mathcal{A})_{pq}, a) a f(a)$$

and thus $SR \leftrightarrow M(\mathcal{A})$ using the bijection $a \to af(a)$ between A and AB. □

Proof of Theorem 5.8. The condition is sufficient by Proposition 5.6. Conversely, let $\mathcal{A}, \mathcal{A}'$ be two symbolic conjugate essential automata. By Theorem 5.9, we may assume that $\mathcal{A}'$ is a split of $\mathcal{A}$. We assume that $\mathcal{A}'$ is an in-split of $\mathcal{A}$. By Proposition 5.11, the adjacency matrices of $\mathcal{A}$ and $\mathcal{A}'$ are symbolic elementary equivalent. □

6 Special families of automata

In this section, we consider two particular families of automata: local automata and automata with finite delay. Local automata are closely related to shifts of finite type. The main result is an embedding theorem (Theorem 6.4) related to Nasu's Masking Lemma (Proposition 6.5). Automata with finite left and right delay are related to a class of shifts called shifts of almost finite type (Proposition 6.10).

6.1 Local automata

Let $m, n \geqslant 0$. An automaton $\mathcal{A} = (Q, E)$ is said to be (m, n)-*local* if whenever $p \xrightarrow{u} q \xrightarrow{v} r$ and $p' \xrightarrow{u} q' \xrightarrow{v} r'$ are two paths with $|u| = m$ and $|v| = n$, then $q = q'$. It is *local* if it is (m, n)-local for some m, n.

Example 6.1. The automaton represented in Figure 14 is $(3, 0)$-local. Indeed, a simple inspection shows that each of the six words of length 3 which are labels of paths uniquely determines its terminal vertex. It is also $(0, 3)$-local. It is not $(2, 0)$-local (check the word ab), but it is $(2, 1)$-local and also $(1, 2)$-local.

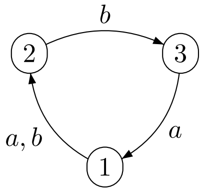


Figure 14. A local automaton.

We say that an automaton $\mathcal{A} = (Q, E)$ is *contained* in an automaton $\mathcal{A}' = (Q', E')$ if $Q \subset Q'$ and $E \subset E'$. We note that if $\mathcal{A}$ is contained in $\mathcal{A}'$ and if $\mathcal{A}'$ is local, then $\mathcal{A}$ is local.

Proposition 6.1. *An essential automaton $\mathcal{A}$ is local if and only if the map $\lambda_{\mathcal{A}} : X_{\mathcal{A}} \to L_{\mathcal{A}}$ is a conjugacy from $X_{\mathcal{A}}$ onto $L_{\mathcal{A}}$.*

Proof. Suppose first that $\mathcal{A}$ is (m, n)-local. Consider an $m{+}1{+}n$-block $w = uav$ of $L_{\mathcal{A}}$, with $|u| = m$, $|v| = n$. All finite paths of $\mathcal{A}$ labeled w have the form $r \xrightarrow{u} p \xrightarrow{a} q \xrightarrow{v} s$ and share the same edge $p \xrightarrow{a} q$. This shows that $\lambda_{\mathcal{A}}$ is injective and that $\lambda_{\mathcal{A}}^{-1}$ is a map with memory m and anticipation n.

Conversely, assume that $\lambda_{\mathcal{A}}^{-1}$ exists, and that it has memory m and anticipation n. We show that $\mathcal{A}$ is $(m+1,n)$-local. Let

$$r \xrightarrow{u} p \xrightarrow{a} q \xrightarrow{v} s \quad \text{and} \quad r' \xrightarrow{u} p' \xrightarrow{a} q' \xrightarrow{v} s'$$

and be two paths of length $m+1+n$, with $|u| = m$, $|v| = n$ and a a letter. Since $\mathcal{A}$ is essential, there exist two biinfinite paths which contain these finite paths, respectively. Since $\lambda_{\mathcal{A}}^{-1}$ has memory m and anticipation n, the blocks uav of the biinfinite words carried by these paths are mapped by $\lambda_{\mathcal{A}}^{-1}$ onto the edges $p \xrightarrow{a} q$ and $p' \xrightarrow{a} q'$ respectively. This shows that $p = p'$ and $q = q'$. □

The next statement is Proposition 10.3.10 in [4].

Proposition 6.2. *The following conditions are equivalent for a strongly connected finite automaton $\mathcal{A}$.*

(i) *$\mathcal{A}$ is local;*

(ii) *distinct cycles have distinct labels.*

Two cycles in this statement are considered to be distinct if, viewed as paths, they are distinct.

The following result shows the strong connection between shifts of finite type and local automata. It gives an effective method to verify whether or not a shift space is of finite type.

Proposition 6.3. *A shift space (resp. an irreducible shift space) is of finite type if and only if its Krieger automaton (resp. its Fischer automaton) is local.*

Proof. Let $X = X^{(W)}$ for a finite set $W \subset A^*$. We may assume that all words of W have the same length n. Let $\mathcal{A} = (Q, i, Q)$ be the $(n,0)$-local deterministic automaton defined as follows. The set of states is $Q = A^n \setminus W$ and there is an edge (u, a, v) for every $u, v \in Q$ and $a \in A$ such that $ua \in Av$. Then $\mathcal{A}$ recognizes the set $\mathcal{B}(X)$. Since the reduction of a local automaton is local, the minimal automaton of $\mathcal{B}(X)$ is local. Since the Krieger automaton of X is contained in the minimal automaton of $\mathcal{B}(X)$, it is local. If X is irreducible, then its Fischer automaton is also local since it is contained in the Krieger automaton.

Conversely, Proposition 6.1 implies that a shift space recognized by a local automaton is conjugate to a shift of finite type and thus is of finite type. □

Example 6.2. Let X be the shift of finite type on the alphabet $A = \{a, b\}$ defined by the forbidden factor ba. The Krieger automaton of X is represented on Figure 15. It is $(1,0)$-local.

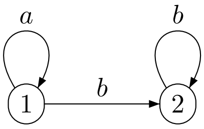


Figure 15. The Krieger automaton of a reducible shift of finite type.

For $m, n \geqslant 0$, the *standard* (m, n)-local automaton is the automaton with states the set of words of length $m+n$ and edges the triples $(uv, a, u'v')$ for $u, u' \in A^m$, $a \in A$ and $v, v' \in A^n$ such that for some letters $b, c \in A$, one has $uvc = bu'v'$ and a is the first letter of vc.

The standard $(m, 0)$-local automaton is also called the De Bruijn automaton of order m.

Example 6.3. The standard $(1, 1)$-local automaton on the alphabet $\{a, b\}$ is represented on Figure 16.

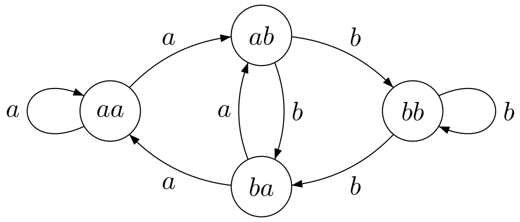


Figure 16. The standard $(1, 1)$-local automaton.

Complete automata. An automaton $\mathcal{A}$ on the alphabet A is called *complete* if any word on A is the label of some path in $\mathcal{A}$. As an example, the standard (m, n)-local automaton is complete.

The following result is from [3].

Theorem 6.4. *Any local automaton is contained in a complete local automaton.*

The proof relies on the following version of the masking lemma.

Proposition 6.5 (Masking lemma). *Let $\mathcal{A}$ and $\mathcal{B}$ be two automata and assume that $M(\mathcal{A})$ and $M(\mathcal{B})$ are elementary equivalent. If $\mathcal{B}$ is contained in an automaton $\mathcal{B}'$, then $\mathcal{A}$ is contained in some automaton $\mathcal{A}'$ which is conjugate to $\mathcal{B}'$.*

Proof. Let $\mathcal{A} = (Q, E)$, $\mathcal{B} = (R, F)$ and $\mathcal{B}' = (R', F')$. Let D be an $R \times Q$ nonnegative integral matrix and N be an alphabetic $Q \times R$ matrix such that $M(\mathcal{A}) = ND$ and $M(\mathcal{B}) = DN$. Set $Q' = Q \cup (F' \setminus F)$. Let D' be the $R' \times Q'$ nonnegative integral matrix defined for $r \in R'$ and $u \in Q'$ by

$$D'_{ru} = \begin{cases} D_{ru} & \text{if } r \in R, u \in Q \\ 1 & \text{if } u \in F' \setminus F \text{ and } u \text{ starts in } r \\ 0 & \text{otherwise} \end{cases}$$

Let N' be the alphabetic $Q' \times R'$ matrix defined for $a \in A$ for $u \in Q'$ and $s \in R'$ by

$$(N'_{us}, a) = \begin{cases} (N_{us}, a) & \text{if } u \in Q, s \in R \\ 1 & \text{if } u \in F' \setminus F \text{ and } u \text{ is labeled with } a \text{ and ends in } s, \\ 0 & \text{otherwise.} \end{cases}$$

Then $N'D'$ is the adjacency matrix of an automaton $\mathcal{A}'$. By definition, $\mathcal{A}'$ contains $\mathcal{A}$ and it is conjugate to $\mathcal{B}'$ by Proposition 3.10. □

We illustrate the proof of Proposition 6.5 by the following example.

Example 6.4. Consider the automata $\mathcal{A}$ and $\mathcal{B}$ given of Figure 17. The automaton $\mathcal{A}$ is the local automaton of Example 6.1. The automaton $\mathcal{B}$ is an in-split of $\mathcal{A}$. Indeed, we have $M(\mathcal{A}) = ND$, $M(\mathcal{B}) = DN$ with

$$N = \begin{bmatrix} 0 & a & b & 0 \\ 0 & 0 & 0 & b \\ a & 0 & 0 & 0 \end{bmatrix} \qquad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

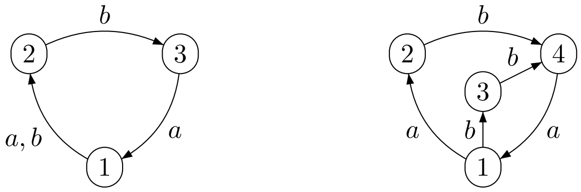


Figure 17. The automaton $\mathcal{B}$ on the right is an in-split of the local automaton $\mathcal{A}$ on the left.

We have represented on the right of Figure 18 the completion of $\mathcal{B}$ as a complete local automaton with the same number of states. On the left, the construction of the proof of Proposition 6.5 has been carried on to produce a local automaton containing $\mathcal{A}$. In terms of adjacency matrices, we have $M(\mathcal{A}') = N'D'$, $M(\mathcal{B}') = D'N'$ with

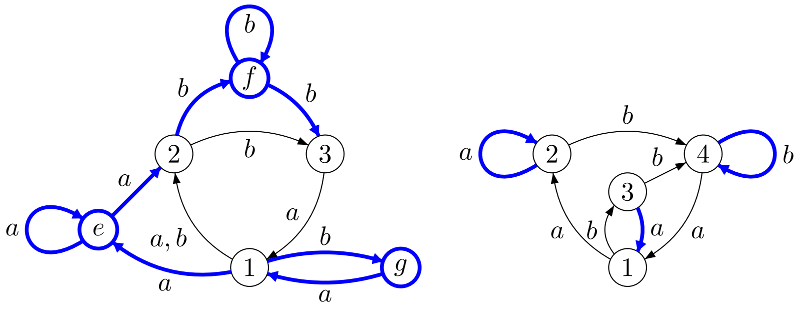


Figure 18. The automata $\mathcal{A}'$ and $\mathcal{B}'$. Additional edges are drawn thick.

$$N' = \begin{bmatrix} 0 & a & b & 0 \\ 0 & 0 & 0 & b \\ a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & 0 & b \\ a & 0 & 0 & 0 \end{bmatrix}, \qquad D' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Proof of Theorem 6.4. Since $\mathcal{A}$ is local, the map $\lambda_{\mathcal{A}}$ is a conjugacy from $X_{\mathcal{A}}$ to $L_{\mathcal{A}}$. Let (m,n) be the memory and anticipation of $\lambda_{\mathcal{A}}^{-1}$. There is a sequence $(\mathcal{A}_0,\ldots,\mathcal{A}_{m+n})$ of automata such that $\mathcal{A}_0 = \mathcal{A}$, each $\mathcal{A}_i$ is a split or a merge of $\mathcal{A}_{i-1}$ and $\mathcal{A}_{n+m}$ is contained in the standard $(n+m)$-local automaton. Applying iteratively Proposition 6.5, we obtain that $\mathcal{A}$ is contained in an automaton which is conjugate to the standard (m,n)-local automaton and which is thus complete. □

6.2 Automata with finite delay

An automaton is said to have *right delay* $d \geqslant 0$ if for any pair of paths

$$p \xrightarrow{a} q \xrightarrow{z} r, \quad p \xrightarrow{a} q' \xrightarrow{z} r'$$

with $a \in A$, if $|z| = d$, then $q = q'$. Thus a deterministic automaton has right delay 0. An automaton has *finite right delay* if it has right delay d for some (finite) integer d. Otherwise, it is said to have *infinite right delay*.

Example 6.5. The automaton represented on Figure 19 has right delay 1.

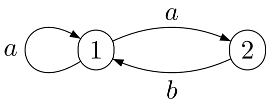


Figure 19. A automaton with right delay 1

Proposition 6.6. *An automaton has infinite right delay if and only if there exist paths $p \xrightarrow{v} q \xrightarrow{u} q$ and $p \xrightarrow{v} q' \xrightarrow{u} q'$ with $q \neq q'$ and $|u| > 0$.*

The following statement is Proposition 5.1.11 in [19].

Proposition 6.7. *An automaton has finite right delay if and only if it is conjugate to a deterministic automaton.*

In the same way the automaton is said to have *left delay* $d \geqslant 0$ if for any pair of paths $p \xrightarrow{z} q \xrightarrow{a} r$ and $p' \xrightarrow{z} q' \xrightarrow{a} r$ with $a \in A$, if $|z| = d$, then $q = q'$.

Corollary 6.8. *If two automata are conjugate, and if one has finite right (left) delay, then the other also has.*

Proposition 6.9. *An essential (m,n)-local automaton has right delay n and left delay m.*

Proof. Let $p \xrightarrow{a} q \xrightarrow{z} r$ and $p \xrightarrow{a} q' \xrightarrow{z} r'$ be two paths with $a \in A$ and $|z| = n$. Since $\mathcal{A}$ is essential there is a path $u \xrightarrow{y} p$ of length m in $\mathcal{A}$. Since $\mathcal{A}$ is (m,n)-local, we have $q = q'$. Thus $\mathcal{A}$ has right delay n. The proof for the left delay m is symmetrical. □

A shift space is said to have *almost finite type* if it can be recognized by a strongly connected automaton with both finite left and finite right delay.

An irreducible shift of finite type is also of almost finite type since a local automaton has finite right and left delay by Proposition 6.9.

Example 6.6. The even shift has almost finite type. Indeed, the automaton of Figure 5 on the right has right and left delay 0.

The following result is from [20].

Proposition 6.10. *An irreducible shift space is of almost finite type if and only if its Fischer automaton has finite left delay.*

Proof. The condition is obviously sufficient. Conversely, let X be a shift of almost finite type. Assume the Fischer automaton $\mathcal{A} = (Q, E)$ of X does not have finite left delay. Let, in view of Proposition 6.6 $u, v \in A^*$ and $p, q, q' \in Q$ with $q \neq q'$ be such that $q \cdot u = q$, $q' \cdot u = q'$ and $p = q \cdot v = q' \cdot v$. Since $\mathcal{A}$ is strongly connected, there is a word w such that $p \cdot w = q$.

Let $\mathcal{B} = (R, F)$ be an automaton with finite right and left delay which recognizes X. By Proposition 6.7, we may assume that $\mathcal{B}$ is deterministic. Let $\varphi : R \to Q$ be a reduction from $\mathcal{B}$ onto $\mathcal{A}$. Since R is finite, there is an $x \in u^+$ such that $r \cdot x = r \cdot x^2$ for all $r \in R$ (this means that the map $r \mapsto r \cdot x$ is idempotent; such a word exists since each element in the finite transition semigroup of the automaton $\mathcal{B}$ has a power which is an idempotent). Set

$$S = R \cdot x, \quad T = \varphi^{-1}(q) \cap S, \quad T' = \varphi^{-1}(q') \cap S$$

Since $q \neq q'$, we have $T \cap T' = \emptyset$. For any $t \in T$, we have $\varphi(t \cdot vw) = q$ and thus $t \cdot vwx \in T$. For $t, t' \in T$ with $t \neq t'$, we cannot have $t \cdot vwx = t' \cdot vwx$ since otherwise $\mathcal{B}$ would have infinite left delay. Thus the map $t \mapsto t \cdot vwx$ is a bijection of T.

Let $t' \in T'$. Since $\varphi(t' \cdot vw) = q$, we have $t' \cdot vwx \in T$. Since the action of vwx induces a permutation on T, there exists $t \in T$ such that $t \cdot vwx = t' \cdot vwx$. This contradicts the fact that $\mathcal{B}$ has finite left delay. □

Example 6.7. The deterministic automaton represented on Figure 20 has infinite left delay. Indeed, there are paths $\cdots 1 \xrightarrow{b} 1 \xrightarrow{a} 1$ and $\cdots 2 \xrightarrow{b} 2 \xrightarrow{a} 1$. Since this automaton cannot be reduced, $X = L_{\mathcal{A}}$ is not of almost finite type.

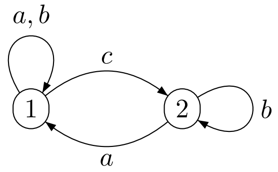


Figure 20. An automaton with infinite left delay

7 Syntactic invariants

We introduce in this section the syntactic graph of an automaton. It uses the Green relations in the transition semigroup of the automaton. We show that the syntactic graph is an invariant for symbolic conjugacy (Theorem 7.4). The proof uses bipartite automata.

The final subsection considers the characterization of sofic shifts with respect to the families of ordered semigroups known as pseudovarieties.

7.1 The syntactic graph

Let $\mathcal{A} = (Q, E)$ be a deterministic automaton on the alphabet A. Each word $w \in A^*$ defines a partial map denoted by $\varphi_{\mathcal{A}}(w)$ from Q to Q which maps $p \in Q$ to $q \in Q$ if $p \cdot w = q$. The transition semigroup of $\mathcal{A}$, already defined in Section 4.2, is the image of A^+ by the morphism $\varphi_{\mathcal{A}}$ (in this subsection, we will not use the order on the transition semigroup).

We give a short summary of *Green relations* in a semigroup (see [17] for example). Let S be a semigroup and let $S^1 = S \cup 1$ be the monoid obtained by adding an identity to S. Two elements s, t of S are $\mathcal{R}$-equivalent if $sS^1 = tS^1$. They are $\mathcal{L}$-equivalent if $S^1s = S^1t$. It is a classical result (see [17]) that $\mathcal{LR} = \mathcal{RL}$. Thus $\mathcal{LR} = \mathcal{RL}$ is an equivalence on the semigroup S called the $\mathcal{D}$-equivalence. A class of the $\mathcal{R}, \mathcal{L}$ or $\mathcal{D}$-equivalence is called an $\mathcal{R}, \mathcal{L}$ or $\mathcal{D}$-class. An *idempotent* of S is an element e such that $e^2 = e$. A $\mathcal{D}$-class is *regular* if it contains an idempotent. The equivalence $\mathcal{H}$ is defined as $\mathcal{H} = \mathcal{R} \cap \mathcal{L}$. It is classical result that the $\mathcal{H}$-class of an idempotent is a group. The $\mathcal{H}$-class of idempotents in the same $\mathcal{D}$-class are isomorphic groups. The *structure group* of a regular $\mathcal{D}$-class is any of the $\mathcal{H}$-classes of an idempotent of the $\mathcal{D}$-class.

When S is a semigroup of partial maps from a set Q into itself, each element of S has a rank which is the cardinality of its image. The elements of a $\mathcal{D}$-class all have the same rank, which is called the *rank* of the $\mathcal{D}$-class. There is at most one element of rank 0 which is the *zero* of the semigroup S and is denoted 0.

A *fixpoint* of a partial map s from Q into itself is an element q such that the image of q by s is q. The rank of an idempotent is equal to the number of its fixpoints. Indeed, in this case, every element in the image is a fixpoint.

The preorder $\leqslant_{\mathcal{J}}$ on S is defined by $s \leqslant_{\mathcal{J}} t$ if $S^1sS^1 \subset S^1tS^1$. Two elements $s, t \in S$ are $\mathcal{J}$-equivalent if $S^1sS^1 = S^1tS^1$. One has $\mathcal{D} \subset \mathcal{J}$ and it is a classical result that in a finite semigroup $\mathcal{D} = \mathcal{J}$. The preorder $\leqslant_{\mathcal{J}}$ induces a partial order on the $\mathcal{D}$-classes, still denoted $\leqslant_{\mathcal{J}}$.

We associate with $\mathcal{A}$ a labeled graph $G(\mathcal{A})$ called its *syntactic graph*. The vertices of $G(\mathcal{A})$ are the regular $\mathcal{D}$-classes of the transition semigroup of $\mathcal{A}$. Each vertex is labeled by the rank of the $\mathcal{D}$-class and its structure group. There is an edge from the vertex associated with a $\mathcal{D}$-class D to the vertex associated to a $\mathcal{D}$-class D' if and only if $D \geqslant_{\mathcal{J}} D'$.

Example 7.1. The automaton $\mathcal{A}$ of Figure 21 on the left is the Fischer automaton of the even shift (Example 4.3). The semigroup of transitions of $\mathcal{A}$ has 3 regular $\mathcal{D}$-classes of ranks 2 (containing $\varphi_{\mathcal{A}}(b)$), 1 (containing $\varphi_{\mathcal{A}}(a)$), and 0 (containing $\varphi_{\mathcal{A}}(aba)$). Its syntactic graph is represented on the right.

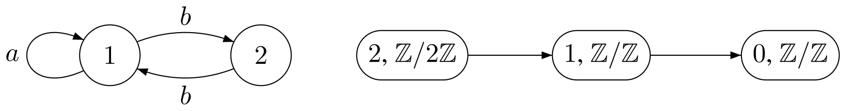


Figure 21. The syntactic graph of the even shift

The following result shows that one may reduce to the case of essential automata.

Proposition 7.1. *The syntactic graphs of an automaton and of its essential part are isomorphic.*

Proof. Let $\mathcal{A} = (Q, E)$ be a deterministic automaton on the alphabet A and let $\mathcal{A}' = (Q', E')$ be its essential part. Let $w \in A^+$ be such that $e = \varphi_{\mathcal{A}}(w)$ is an idempotent. Then any fixpoint of e is in Q' and thus $e' = \varphi_{\mathcal{A}'}(w)$ an idempotent of the same rank as e. This shows that $G(\mathcal{A})$ and $G(\mathcal{A}')$ are isomorphic. □

The following result shows that the syntactic graph characterizes irreducible shifts of finite type.

Proposition 7.2. *A sofic shift (resp. an irreducible sofic shift) is of finite type if and only if the syntactic graph of its Krieger automaton (resp. its Fischer automaton) has nodes of rank at most* 1.

In the proof, we use the following classical property of finite semigroups.

Proposition 7.3. *Let S be a finite semigroup and let J be an ideal of S. The following conditions are equivalent.*

(i) All idempotents of S are in J.

(ii) There exists an integer $n \geqslant 1$ such that $S^n \subset J$.

Proof. Assume that (i) holds. Let $n = \mathrm{Card}(S) + 1$ and let $s = s_1 s_2 \cdots s_n$ with $s_i \in S$. Then there exist i, j with $1 \leqslant i < j \leqslant n$ such that $s_1 s_2 \cdots s_i = s_1 s_2 \cdots s_i \cdots s_j$. Let $t, u \in S^1$ be defined by $t = s_1 \cdots s_i$ and $u = s_{i+1} \cdots s_j$. Since $tu = t$, we have $tu^k = t$ for all $k \geqslant 1$. Since S is finite, there is a $k \geqslant 1$ such that u^k is idempotent and thus $u^k \in J$. This implies that $t \in J$ and thus $s \in J$. Thus (ii) holds.
It is clear that (ii) implies (i). □

Proof of Proposition 7.2. Let X be a shift space (resp. an irreducible shift space), let $\mathcal{A}$ be its Krieger automaton (resp. its Fischer automaton) and let S be the transition semigroup of $\mathcal{A}$.

If X is of finite type, by Proposition 6.3, the automaton $\mathcal{A}$ is local. Any idempotent in S has rank 1 and thus the condition is satisfied.

Conversely, assume that the graph $G(\mathcal{A})$ has nodes of rank at most 1. Let J be the ideal of S formed of the elements of rank at most 1. Since all idempotents of S belong to J, by Proposition 7.3, the semigroup S satisfies $S^n = J$ for some $n \geqslant 1$. This shows that for any sufficiently long word x, the map $\varphi_{\mathcal{A}}(x)$ has rank at most 1. Thus for $p, q, r, s \in Q$, if $p \cdot x = r$ and $q \cdot x = s$ then $r = s$. This implies that $\mathcal{A}$ is $(n, 0)$-local. □

The following result is from [2].

Theorem 7.4. *Two symbolic conjugate automata have isomorphic syntactic graphs.*

We use the following intermediary result.

Proposition 7.5. *Let $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ be a bipartite automaton. The syntactic graphs of $\mathcal{A}, \mathcal{A}_1$ and $\mathcal{A}_2$ are isomorphic.*

Proof. Let $Q = Q_1 \cup Q_2$ and $A = A_1 \cup A_2$ be the partitions of the set of of states and of the alphabet of $\mathcal{A}$ corresponding to the decomposition $(\mathcal{A}_1, \mathcal{A}_2)$. Set $B_1 = A_1A_2$ and $B_2 = A_2A_1$. The semigroups $S_1 = \varphi_{\mathcal{A}_1}(B_1^+)$ and $S_2 = \varphi_{\mathcal{A}_2}(B_2^+)$ are included in the semigroup $S = \varphi_{\mathcal{A}}(A^+)$. Thus the Green relations of S are refinements of the corresponding Green relations in S_1 or in S_2. Any idempotent e of S belongs either to S_1 or to S_2. Indeed, if $e = 0$ then e is in $S_1 \cap S_2$. Otherwise, it has at least one fixpoint $p \in Q_1 \cup Q_2$. If $p \in Q_1$, then e is in $\varphi_{\mathcal{A}}(B_1^+)$ and thus $e \in S_1$. Similarly if $p \in Q_2$ then $e \in S_2$.

Let e be an idempotent in S_1 and let $e = \varphi_{\mathcal{A}}(u)$. Since $u \in B_1^+$, we have $u = au'$ with $a \in A_1$ and $u' \in B_2^* A_2$. Let $v = u'a$. Then $f = \varphi_{\mathcal{A}}(v)^2$ is idempotent. Indeed, we have

$$\varphi_{\mathcal{A}}(v^3) = \varphi_{\mathcal{A}}(u'au'au'a) = \varphi_{\mathcal{A}}(u'uua) = \varphi_{\mathcal{A}}(u'ua) = \varphi_{\mathcal{A}}(v^2)$$

Moreover e, f belong the same $\mathcal{D}$-class. Similarly, if $e \in S_2$, there is an idempotent in S_1 which is $\mathcal{D}$ equivalent to e. This shows that a regular $\mathcal{D}$-class of $\varphi_{\mathcal{A}}(A^+)$ contains idempotents in S_1 and in S_2.

Finally, two elements of S_1 which are $\mathcal{D}$-equivalent in S are also $\mathcal{D}$-equivalent in S_1. Indeed, let $s, t \in S_1$ be such that $s\mathcal{R}\mathcal{L}t$. Let $u, u', v, v' \in S$ be such that

$$suu' = s, \quad v'vt = t, \quad su = tv$$

in such a way that $s\mathcal{R}su$ and $vt\mathcal{L}t$. Then $su = vt$ implies that u, v are both in S_1. Similarly $suu' = s$ and $v'vt = t$ imply that $u'v' \in S_1$. Thus $s\mathcal{D}t$ in S_1. This shows that a regular $\mathcal{D}$ class D of S contains exactly one $\mathcal{D}$-class D_1 of S_1 (resp. D_2 of S_2). Moreover, an $\mathcal{H}$-class of D_1 is also an $\mathcal{H}$-class of D.

Thus the three syntactic graphs are isomorphic. □

Proof of Theorem 7.4. Let $\mathcal{A} = (Q, E)$ and $\mathcal{B} = (R, F)$ be two symbolic conjugate automata on the alphabets A and B, respectively. By the Decomposition Theorem (Theorem 5.9), we may assume that the symbolic conjugacy is a split or a merge. Assume that $\mathcal{A}'$ is an in-split of $\mathcal{A}$. By Proposition 7.1, we may assume that $\mathcal{A}$ and $\mathcal{A}'$ are essential. By Proposition 5.11, the adjacency matrices of $\mathcal{A}$ and $\mathcal{A}'$ are symbolic elementary equivalent.

By Proposition 5.5, there is a bipartite automaton $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2)$ such that $M(\mathcal{C}_1), M(\mathcal{C}_2)$ are similar to $M(\mathcal{A}), M(\mathcal{B})$ respectively. By Proposition 7.5, the syntactic graphs of $\mathcal{C}_1, \mathcal{C}_2$ are isomorphic. Since automata with similar adjacency matrices have obviously isomorphic syntactic graphs, the result follows. □

A refinement of the syntactic graph which is also invariant by flow equivalence has been introduced in [9]. The vertices of the graph are the *idempotent-bound* $\mathcal{D}$ classes, where an element s of a semigroup S is called idempotent-bound if there exist idempotents $e, f \in S$ such that $s = esf$. The elements of a regular $\mathcal{D}$-class are idempotent-bound.

Flow equivalent automata. Let $\mathcal{A}$ be an automaton on the alphabet A and let G be its underlying graph. An *expansion* of $\mathcal{A}$ is a pair (φ, ψ) of a graph expansion of G and a symbol expansion of $L_{\mathcal{A}}$ such that the diagram below is commutative. The inverse of an automaton expansion is called a contraction.

$$\begin{tikzcd}
X_{\mathcal{A}} \arrow[r, "\varphi"] \arrow[d, "\lambda_{\mathcal{A}}"] & X_{\mathcal{B}} \arrow[d, "\lambda_{\mathcal{B}}"] \\
L_{\mathcal{A}} \arrow[r, "\psi"] & L_{\mathcal{B}}
\end{tikzcd}$$

Example 7.2. Let $\mathcal{A}$ and $\mathcal{B}$ be the automata represented on Figure 22. The second automaton is an expansion of the first one.

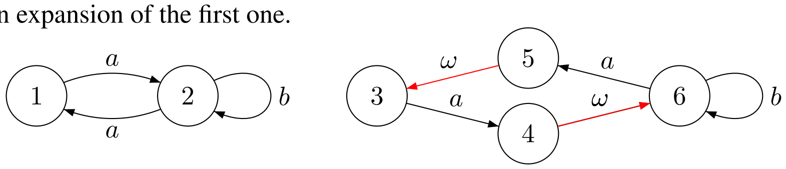


Figure 22. An automaton expansion

The *flow equivalence* of automata is the equivalence generated by symbolic conjugacies, expansions and contractions.

Theorem 7.4 has been generalized by Costa and Steinberg [12] to flow equivalence.

Theorem 7.6. *Two flow equivalent automata have isomorphic syntactic graphs.*

Example 7.3. The syntactic graphs of the automata $\mathcal{A}$, $\mathcal{B}$ of Example 5.2 are isomorphic to the syntactic graph of the Fischer automaton $\mathcal{C}$ of the even shift. Note that the automata $\mathcal{A}, \mathcal{B}$ are not flow equivalent to $\mathcal{C}$. Indeed, the edge shifts $X_{\mathcal{A}}$, $X_{\mathcal{B}}$ on the underlying graphs of the automata $\mathcal{A}$, $\mathcal{B}$ are flow equivalent to the full shift on 3 symbols while the edge shift $X_{\mathcal{C}}$ is flow equivalent to the full shift on 2 symbols. Thus the converse of Theorem 7.6 is false.

7.2 Pseudovarieties

In this subsection, we will see how one can formulate characterizations of some classes of sofic shifts by means of properties of their syntactic semigroup. In order to formulate these syntactic characterizations of sofic shifts, we introduce the notion of pseudovariety of ordered semigroups. For a systematic exposition, see the original articles [25], [27], or the surveys in [26] or [24].

A morphism of ordered semigroups φ from S into T is an order compatible semigroup morphism, that is such that $s \leqslant s'$ implies $\varphi(s) \leqslant \varphi(s')$. An ordered subsemigroup of S is a subsemigroup equipped with the restriction of the preorder.

A *pseudovariety* of finite ordered semigroups is a class of ordered semigroups closed under taking ordered subsemigroups, finite direct products and image under morphisms of ordered semigroups.

Let V be a pseudovariety of ordered semigroups. We say that a semigroup S is *locally* in V if all the submonoids of S are in V. The class of these semigroups is a pseudovariety of ordered semigroups.

The following result is due to Costa [10].

Theorem 7.7. *Let V be a pseudovariety of finite ordered semigroups containing the class of commutative ordered monoids such that every element is idempotent and greater than the identity. The class of shifts whose syntactic semigroup is locally in V is invariant under conjugacy.*

The following statements give examples of pseudovarieties satisfying the above condition.

Proposition 7.8. *An irreducible shift space is of finite type if and only if its syntactic semigroup is locally commutative.*

An *inverse semigroup* is a semigroup which can be represented as a semigroup of partial one-to-one maps from a finite set Q into itself. The family of inverse semigroups does not form a variety (it is not closed under homomorphic image. However, according to Ash's theorem [1], the variety generated by inverse semigroups is characterized by the property that the idempotents commute. Using this result, the following result is proved in [10].

Theorem 7.9. *An irreducible shift space is of almost finite type if and only if its syntactic semigroup is locally in the pseudovariety generated by inverse semigroups.*

The fact that shifts of almost finite type satisfy this condition was proved in [2]. The converse was conjectured in the same paper.

In [12] it is shown that this result implies that the class of shifts of almost finite type is invariant under flow equivalence. This is originally from [15].

References

[1] C. J. Ash. Finite semigroups with commuting idempotents. *J. Austral. Math. Soc. Ser. A*, 43(1):81–90, 1987. 936

[2] M.-P. Béal, F. Fiorenzi, and D. Perrin. The syntactic graph of a sofic shift is invariant under shift equivalence. *Internat. J. Algebra Comput.*, 16(3):443–460, 2006. 934, 936

[3] M.-P. Béal, S. Lombardy, and D. Perrin. Embeddings of local automata. In *Information Theory, ISIT 2008*, pages 2351–2355. IEEE, 2008. To appear in *Illinois J. Math.* 928

[4] J. Berstel, D. Perrin, and C. Reutenauer. *Codes and automata*. Cambridge University Press, 2009. 927

[5] R. Bowen and J. Franks. Homology for zero-dimensional nonwandering sets. *Ann. Math. (2)*, 106(1):73–92, 1977. 905

[6] M. Boyle. Flow equivalence of shifts of finite type via positive factorizations. *Pacific J. of Math.*, 204:273–317, 2002. 906

[7] M. Boyle. Open problems in symbolic dynamics. In *Geometric and probabilistic structures in dynamics*, volume 469 of *Contemp. Math.*, pages 69–118. Amer. Math. Soc., Providence, RI, 2008. 906

[8] M. Boyle and D. Huang. Poset block equivalence of integral matrices. *Trans. Amer. Math. Soc.*, 355(10):3861–3886 (electronic), 2003. 906

[9] A. Costa. Conjugacy invariants of subshifts: an approach from profinite semigroup theory. *Int. J. Algebra Comput.*, 16:629–655, 2006. 934

[10] A. Costa. Pseudovarieties defining classes of sofic subshifts closed under taking equivalent subshifts. *J. Pure Applied Alg.*, 209:517–530, 2007. 936

[11] A. Costa. *Semigroupos Profinitos e Dinâmica Simbólica*. PhD thesis, Universidade do Porto, 2007. 914

[12] A. Costa and B. Steinberg. Idempotent splitting categories of idempotents associated to subshifts are flow equivalence invariants. Technical report, 2010. 935, 936

[13] R. Fischer. Sofic systems and graphs. *Monatsh. Math.*, 80:179–186, 1975. 916

[14] J. Franks. Flow equivalence of subshifts of finite type. *Ergodic Theory Dynam. Systems*, 4(1):53–66, 1984. 905

[15] M. Fujiwara and M. Osikawa. Sofic systems and flow equivalence. *Math. Rep. Kyushu Univ.*, 16(1):17–27, 1987. 936

[16] T. Hamachi and M. Nasu. Topological conjugacy for 1-block factor maps of subshifts and sofic covers. In *Dynamical Systems*, volume 1342 of *Lecture Notes in Mathematics*, pages 251–260. Springer-Verlag, 1988. 924

[17] J. M. Howie. *An introduction to semigroup theory*. Academic Press [Harcourt Brace Jovanovich Publishers], London, 1976. L.M.S. Monographs, No. 7. 932

[18] W. Krieger. On sofic systems. I. *Israel J. Math.*, 48(4):305–330, 1984. 923

[19] D. A. Lind and B. H. Marcus. *An introduction to symbolic dynamics and coding*. Cambridge University Press, 1995. 896, 898, 900, 901, 902, 903, 904, 916, 917, 923, 930

[20] M. Nasu. An invariant for bounded-to-one factor maps between transitive sofic subshifts. *Ergodic Theory Dynam. Systems*, 5(1):89–105, 1985. 931

[21] M. Nasu. Topological conjugacy for sofic systems. *Ergodic Theory Dynam. Systems*, 6(2):265–280, 1986. 920, 923

[22] M. Nasu. Topological conjugacy for sofic systems and extensions of automorphisms of finite subsystems of topological markov shifts. In *Proceedings of Maryland special year in Dynamics 1986–87*, volume 1342 of *Lecture Notes in Mathematics*, pages 564–607. Springer-Verlag, 1988. 914

[23] B. Parry and D. Sullivan. A topological invariant of flows on 1-dimensional spaces. *Topology*, 14(4):297–299, 1975. 904

[24] D. Perrin and J.-E. Pin. *Infinite Words*. Elsevier, 2004. 935

[25] J.-E. Pin. Eilenberg's theorem for positive varieties of languages. *Izv. Vyssh. Uchebn. Zaved. Mat.*, (1):80–90, 1995. 935

[26] J.-E. Pin. Syntactic semigroups. In *Handbook of formal languages, Vol. 1*, pages 679–746. Springer, Berlin, 1997. 935

[27] J.-E. Pin, A. Pinguet, and P. Weil. Ordered categories and ordered semigroups. *Comm. Algebra*, 30(12):5651–5675, 2002. 935

Automatic structures (draft)

Sasha Rubin

email: sasha.rubin@gmail.com

May 18, 2015 22 h 15

2010 Mathematics Subject Classification: 03B70, 03B25, 03D05.

Key words: Synchronous automata, decidable theories.

Contents

1 Introduction

Decidability. Much elementary algebra and geometry can be expressed in the first-order (FO) language of the structure of *real arithmetic*

$$(\mathbb{R}, +, \times, <, +1, =, 0, 1).$$

Formulas in this FO language allow quantification over elements of $\mathbb{R}$, make use of the operations $+$, $\times$ and $+1$, the relations $<$ and $=$, the constants 0 and 1, and the Boolean connectives (negation, conjunction, etc). Tarski proved that there is an algorithm that decides the truth or falsity of every FO-sentence in this language (a sentence is a formula with no free variables). One says that the FO-theory (i.e. the set of true FO-sentences) of real arithmetic is decidable. Tarski's proof uses a technique called effective quantifier elimination. The approach looks familiar: the statement that a quadratic has two real roots

$$\exists x \exists y [x \neq y \wedge ax^2 + bx + c = 0 \wedge ay^2 + by + c = 0]$$

(written with abbreviations such as x^2 for $x \times x$) can be replaced by the simpler quantifier-free condition $a \neq 0 \wedge 4ac < b^2$.

This technique does more than just prove decidability. It gives some insight into the nature of the sets and relations definable by formulas of the language. Indeed, Tarski's procedure transforms a formula into an equivalent formula with no quantifiers and without changing its free variables, and it also provides a proof (from the axioms for a real closed field) that the two formulas are equivalent. A model-theoretic consequence is that every FO-definable set in real arithmetic is a finite union of intervals with algebraic endpoints. An algorithmic consequence is that one can effectively determine the number of real roots of a given polynomial. Tarski's paper [60] contains a nice discussion of his result and its uses. For an elementary proof see [47].

There are a number of techniques for proving theories decidable and understanding the definable relations. Notable approaches are providing a finite (or computably enumerable) axiomatisation of a theory[1]; and the composition method (associated with the names Feferman, Vaught and Shelah) that leverages a decomposition of the structure into pieces whose theories determine (and can be used to compute) the original theory.

Of course not all theories are decidable. A good heuristic is that if a long unsolved problem is expressible then the theory is probably undecidable. The canonical example is the FO-theory of the structure of integer arithmetic $(\mathbb{N}, +, \times, <, =, 0, 1)$ — its theory is very expressive and highly undecidable.[2] On the other hand the FO-theory $(\mathbb{N}, +, \mathrm{Pr})$ where $\mathrm{Pr} \subset \mathbb{N}$ is the set of primes is only known to be undecidable under some number theoretic assumptions related to the twin-primes conjecture [6].

Enter automata. Various systems of arithmetics furnish a hunting ground for questions of decidability. It is reported in [55] that Tarski had asked: is the (weak) monadic second

[1] Here a theory is any set of FO-sentences closed under logical deduction. If, in addition to being finitely axiomatisable, the theory is also complete (every sentence or its negation is in the set), then the theory is decidable. Showing completeness is the bread and butter of model theorists.

[2] Since a pairing function is definable one can code sequences of integers and so express whether or not a given FO-sentence of arithmetic is provable from a given computable set of FO-axioms; or if one prefers, whether or not a given Turing machine halts on a given input.

order theory of $(\mathbb{N}, +1)$ decidable? Here there are additional variables that range over (finite) *subsets* of the domain $\mathbb{N}$. Viewing finite subsets of $\mathbb{N}$ as finite binary strings Büchi [15], Elgot [24] and Trahtenbrot [64] (independently) showed that WMSO-formulas of $(\mathbb{N}, +1)$ define regular languages. In slogan form: formulas can be compiled into automata. This establishes that WS1S, the WMSO-theory of $(\mathbb{N}, +1)$, is decidable. Indeed, a sentence of the form $\exists X \phi(X)$ holds in $(\mathbb{N}, +1)$ if and only if the automaton for ϕ has non-empty language (itself a decidable property). Büchi [16] then introduced ω-string automata in order to settle that S1S, the MSO-theory of $(\mathbb{N}, +1)$, is decidable. There is a natural next step.

Definition 1.1. The structure $\mathfrak{T}_r$ has domain $\{0, 1, \cdots, r-1\}^*$ and binary relations suc_i $(i \in \{0, 1, \cdots, r-1\})$ consisting of the pairs (w, wi), where $w \in \{0, 1, \cdots, r-1\}^*$. The (W)MSO theory of $\mathfrak{T}_r$ is called $(W)SrS$.

Note that $\mathfrak{T}_1$ is isomorphic to $(\mathbb{N}, +1)$. Büchi [16] asked whether the MSO-theory of $\mathfrak{T}_2$ is decidable. Subsets of $\{0, 1\}^*$ are naturally viewed as $\{0, 1\}$-labeled trees. Doner [22] and Thatcher and Wright [61] used tree automata to show that WS2S is decidable. In a milestone paper Rabin [51] introduced automata operating on ω-trees to prove that S2S is decidable.

Actually more than just decidability was established. We have already seen, on the way to decidability, that definable relations are, modulo a coding, recognised by automata. The converse also holds. We illustrate for ω-tree automata. For a set $A \subseteq \{0, 1\}^*$ write $\chi(A)$ for its characteristic $\{0, 1\}$-labelled tree (Definition 2.1). For a tuple of trees $(t_1, \cdots, t_n)$ write $\otimes(\bar{t})$ for the convoluted $\{0, 1\}^n$-labelled tree (Definition 2.2).

Theorem 1.1 ([51]). *A set of tuples $(A_1, \cdots, A_n)$ of sets of strings is MSO definable in $(\{0, 1\}^\star, \mathrm{s}_0, \mathrm{s}_1)$ if and only if the set of $\{0, 1\}^n$-labelled trees $\otimes(\chi(A_1), \cdots, \chi(A_n))$ is recognised by an ω-tree automaton. Moreover, the translation between formulas and automata is effective.*

An accessible proof of this central theorem can be found in [62]. Similar theorems hold for (W)MSO of $\mathfrak{T}_r$. These theorems can be leveraged to deduce many other decidability and definability results. The method of interpretations is one such lever. Interpretations are familiar in mathematics — e.g., arithmetic on the rationals can be interpreted in arithmetic on the integers — and are studied systematically in model theory (see [30]). Let's illustrate the method with two examples.

Corollary 1.2. *Every FO-definable relation of $(\mathbb{N}, +)$ is, modulo a certain coding of natural numbers into binary strings, recognised by an automaton. The translation from formulas to automata is effective. Consequently, the FO-theory of $(\mathbb{N}, +)$ is decidable.*[3]

The following argument establishing the corollary can be found in [55] [15] [24]. There is a *finite-set interpretation* of the FO-theory of $(\mathbb{N}, +)$ in WS1S. The interpretation codes $n \in \mathbb{N}$ as a finite subset of $\mathbb{N}$ (e.g., thirteen is coded first in binary as 1101 and

[3] The FO-theory of $(\mathbb{N}, +)$ is called *Presburger arithmetic* after M. Presburger who showed it decidable using effective quantifier elimination [50].

then as the set $\{0, 2, 3\}$). This coding has the property that there is a WMSO formula $\phi_+(X_1, X_2, X_3)$ expressing that $X_1' + X_2' = X_3'$, where X_i' is the unique number coded by the set X_i. The formula implements the usual bit-carry procedure for addition: it guesses the existence of the carry set and uses the successor to scan the sets one place at a time verifying the addition. This allows one to translate every FO-formula $\phi(y_1, \cdots, y_n)$ of $(\mathbb{N}, +)$ into a WMSO-formula $\Phi(Y_1, \cdots, Y_n)$ of $(\mathbb{N}, +1)$ with the property that the two formulas define the same relation modulo the coding. This relation, viewed as one on strings, is recognised by a finite-automaton (cf. Theorem 1.1).

The second example concerns the MSO-theory of the rational order $(\mathbb{Q}, <)$.

Corollary 1.3. *Every MSO-definable relation of* $(\mathbb{Q}, <)$ *is, modulo a certain coding of rational numbers into* ω*-trees, recognised by an automaton. The translation from formulas to automata is effective. Consequently, the MSO-theory of* $(\mathbb{Q}, <)$ *is decidable.*

Rabin [51] provides a *monadic second-order interpretation* of the MSO-theory of $(\mathbb{Q}, <)$ in S2S. First consider the binary relation $\prec$ on $\mathfrak{T}_2$ where $x \prec y$ if $(x \sqcap y)0$ is a prefix of x or $(x \sqcap y)1$ is a prefix of y (here $x \sqcap y$ is the longest common prefix of x and y). Note that $(\{0, 1\}^*, \prec)$ is a countable dense linear order with no endpoints; so it is isomorphic to $(\mathbb{Q}, <)$. Since the relation $\prec$ is MSO-definable in $\mathfrak{T}_2$ one can translate every MSO-formula $\phi(X_1, \cdots, X_n)$ of $(\mathbb{Q}, <)$ into an MSO-formula $\phi'(X_1, \cdots, X_n)$ of $\mathfrak{T}_2$ with the property that the two formulas define the same relation modulo the isomorphism. Finally, Theorem 1.1 ensures the relation, viewed as a relation on ω-trees, is recognised by an automaton.

Thus MSO-interpretations allow the transfer of MSO decidability to MSO decidability while set-interpretations allow the transfer of MSO decidability to FO decidability. This paper focuses on the latter. A structure that is set-interpretable in $\mathfrak{T}_2$ is called *Rabin-automatic* (cf. Definition 3.2); for instance we have seen that the structure $(\mathbb{N}, +)$ is Rabin-automatic. The fact that formulas can be compiled into automata gives the *fundamental theorem of automatic structures* which says that a FO-definable relation in a Rabin-automatic structure is, modulo coding, recognised by an automaton (Theorem 3.4). Many results about automatic structures exploit this theorem.

Here is an outline of the types of problems discussed in this chapter.

(1) To what extent can we increase the expressive power of the logical language (FO) and still retain decidability of automatic structures? This is the problem of extending the fundamental theorem of automatic structures.
(2) Which operations on structures preserve being automatic?
(3) What do automatic structures look like? Meaningful answers can be used to show that a given structure is not automatic.
(4) An automatic structure may be set-interpretable in a number of ways. How are these ways different? This is the problem of studying non-equivalent automatic presentations.

Other accounts and omissions. The introduction above reflects one approach to motivating and defining automatic structures. The definition via interpretations rather than automata follows [9, 11] and [19]. Automatic structures were originally introduced in [31, 32] for proving decidability. They were independently introduced in [35] where the motivation stresses that automatic structures are part of computable model theory, a subject that looks at the effective content of mathematical statements. Here one asks for the automatic content of standard theorems from mathematics such as Kőnig's tree lemma, Ramsey's partition results or Cantor's results on countable linear orders, see [58, 38, 39, 45]. An equivalent way of identifying MSO and automata is that strings or trees are themselves considered to be structures (over a certain natural signature) and the regular languages correspond to sets of structures satisfying MSO *sentences*. This is usually discussed in the context of finite model theory, see for instance [23]. There is a sense in which automatic structures are complex — see see [37] [34] for the complexity of automatic structures measured by common rankings such as Scott rank, Cantor-Bendixson rank, and ordinal height. For model theoretic considerations (quantifier-elimination, VC-dimension) see [8]. Useful resources regarding the relationship between automatic structures and other finitary formalism (automatic groups, numeration systems, pushdown graphs, ...) as well as the computational complexity of query evaluation and model checking on automatic structures can be found in [11] [44] [3]. See [28] for a treatment of decision problems for MSO via automata as well as via composition, [52] for FO decision problems via standard techniques (quantifier elimination and completeness considerations), and the comprehensive book [13] for the classical decision problem.

A note on citations. By default I have cited what I think is the best, though not necessarily the first, reference for a given result. The interested reader may refer to [58] for some historical comments.

Acknowledgements. I am grateful to Vince Bárány and Olivier Finkel for their comments and corrections. I thank Achim Blumensath for serving as friendly referee; Wolfgang Thomas and Anil Nerode for clearing up some historical questions; and Jean-Eric Pin for his patient encouragement.

2 Logic, automata and interpretations

2.1 Logical languages

A *structure* $\mathfrak{A} = (\mathcal{A}, R_1, \cdots, R_N)$ consists of a set $\mathcal{A}$ called the *domain*, and relations[4] $R_i \subseteq \mathcal{A}^{r_i}$. The names of the R_i together with their arities $r_i \in \mathbb{N}$ form the *signature* of the structure. If $\mathfrak{B}$ is the restriction of $\mathfrak{A}$ to some set $\mathcal{B} \subseteq \mathcal{A}$ then write $\mathfrak{B} \subseteq \mathfrak{A}$ and say that $\mathfrak{B}$ is a *substructure* of $\mathfrak{A}$.

Example 2.1. The structure $\mathfrak{T}_r$ has domain $\{0, 1, \cdots, r-1\}^\star$ and for each i from $\{0, 1, \cdots, r-1\}$ a binary relation suc_i consisting of pairs (w, wi). Note that $\mathfrak{T}_1$ is isomorphic to $(\mathbb{N}, +1)$.

We write $(\mathfrak{A}, P)$ for the structure $\mathfrak{A}$ expanded by the predicate P. To talk about structures we need a logical language. Formulas of monadic second-order logic (MSO) are constructed using logical connectives ('and', 'or', 'not'), individual variables x, y, z (that are intended to range over elements of the domain), set variables $X, Y, Z, \cdots$ (that

[4] We deal with relational structures. This is no real handicap since we can replace an operation by its graph. For instance, addition is taken as those triples (x, y, z) such that $x + y = z$.

are intended to range over subsets of the domain), quantification over these variables, and the subset $X \subseteq Y$ and membership $x \in X$ relations, and finally they may use the names of the relations in a fixed signature such as $R(x, y)$.

Formulas of weak monadic second-order logic (WMSO) are defined as for MSO except that set variables are intended to range over finite subsets of the domain. Formulas of first-order logic (FO) are defined as for MSO but without set variables. A formula written $\phi(X_1, \cdots, X_n, x_1, \cdots, x_m)$ means that ϕ's free variables are included in the set $\{X_1, \cdots, X_n, x_1, \cdots, x_m\}$. A *sentence* is a formula without free variables. Here is a formula of two free individual variables x, y in the signature of the structure $(\mathbb{N}, +1)$:

$$(\forall Z)[x \in Z \wedge (\forall z)(z \in Z \implies z + 1 \in Z) \implies y \in Z].$$

We can see that it is satisfied by those pairs of natural numbers (x, y) such that $x \leqslant y$. One can similarly define the prefix relation $\prec_{\texttt{prefix}}$ in $\mathfrak{T}_2$. We often write (W)MSO formulas as $\phi(\overline{X})$ even if some of the free variables are individual variables.

Note that we are appealing to our natural sense of what it means for a sentence to be true of a structure. For a rigourous definition of truth and satisfaction in mathematical logic see, for instance, [12]. We will use the shorthand $\mathfrak{A} \models \Phi$ (read $\mathfrak{A}$ models Φ) to mean that the sentence Φ is true in $\mathfrak{A}$. Two formulas $\phi(\overline{X})$ and $\psi(\overline{Y})$ are *equivalent over* $\mathfrak{A}$ if

$$\mathfrak{A} \models \forall \overline{X} \forall \overline{Y} [\phi(\overline{X}) \iff \psi(\overline{Y})].$$

An MSO-formula $\phi(\overline{X}, \overline{x})$, in the signature of $\mathfrak{A}$, *defines* the relation

$$\phi^{\mathfrak{A}} := \{(A_1, \cdots, A_k, a_1, \cdots, a_n) \mid \mathfrak{A} \models \phi(\overline{A}, \overline{a}), A_i \subseteq \mathcal{A}, a_j \in \mathcal{A}\}.$$

A central problem in mathematical logic has been establishing the (non-)decidability of theories. Let $\mathcal{L}$ be one of MSO, WMSO, or FO. The *$\mathcal{L}$-theory* of a structure is the set of $\mathcal{L}$-sentences true in that structure. A set X of sentences is *decidable* if there is an algorithm that correctly decides, given a sentence ϕ in the language of $\mathcal{L}$, whether or not $\phi \in X$. We say that a structure has *decidable $\mathcal{L}$-theory* if its $\mathcal{L}$-theory is decidable.

2.2 Rabin's theorem

The Büchi/Elgot/Trahtenbrot revolution led to increasingly complex structures in which definable relations correspond with some type of automata. The cornerstone is Rabin's theorem: MSO definability in $(\{0,1\}^\star, \mathrm{s}_0, \mathrm{s}_1)$ coincides with recognisability by ω-tree automata. Tree automata operate on *(binary) A-labeled trees* $T : \{0,1\}^* \to A$. For a definition of Rabin tree-automaton see [62]. We code tuples of sets as ω-trees in two steps. In the first step code a set as a tree.

Definition 2.1 (characteristic ω-tree). For a set $Y \subseteq \{0,1\}^*$ define its *characteristic tree* χ_Y as the $\{0,1\}$-labeled ω-tree with a 1 in position $w \in \{0,1\}^*$ if and only if $w \in Y$.

Second, code a tuple of trees as a single tree by laying the tuples alongside each other.

Definition 2.2 (convoluting ω-trees). Let $\bar{t} = (t_1, \cdots, t_k)$ be a k-tuple of $\{0,1\}$-labelled ω-trees. The *convolution* $\otimes(\bar{t})$ is the $\{0,1\}^k$-labelled ω-tree such that for all positions $w \in \{0,1\}^*$ the ith component of $\otimes(\bar{t})(w)$ is equal to $t_i(w)$.

Theorem 2.1 (Rabin's theorem [62]). *For each MSO-formula $\phi(\overline{X})$ in the signature of $\mathfrak{T}_2$ there is an ω-tree automaton (and vice-versa) such that the language recognised by the automaton is*

$$\{\otimes(\chi_{X_1}, \cdots, \chi_{X_k}) \mid \mathfrak{T}_2 \models \phi(\overline{X})\}.$$

The translations are effective.

Proposition 2.2. *[62] The emptiness problem for Rabin-automata is decidable and consequently the MSO-theory of $\mathfrak{T}_2$ is decidable. Moreover, there is an effective procedure that given a non-empty automaton M produces (a finite presentation of) a regular ω-tree accepted by M (this fact is called Rabin's basis theorem).*

Remark. Similar results hold with $\mathfrak{T}_1$ replacing $\mathfrak{T}_2$ and are know as Büchi's theorem(s). These can be proven directly or as corollaries by coding $X \subseteq \mathbb{N}$ by the tree T with $T^{-1}(1) = \{0^n \mid n \in X\}$. Similar results also hold for WMSO. The standard reference is [62].

Rabin's theorem with additional set quantifiers We show that we can enrich MSO by certain set quantifiers (such as 'there are finitely many sets X such that ...') and still get decidability for $\mathfrak{T}_2$. We do this by showing that formulas with the additional quantifiers are actually equivalent to vanilla MSO formulas.

Lemma 2.3. *The property 'X is finite' is* MSO*-definable in $\mathfrak{T}_2$.*

Proof. The following simple argument is from [51]. The lexicographic (total) ordering $<_{\text{lex}}$ on $\{0,1\}^*$ is MSO-definable in $\mathfrak{T}_2$. Thus '$X \subseteq \{0,1\}^*$ is finite' is definable by the formula that says that every $B \subseteq X$ has both a maximum and minimum element with respect to $<_{\text{lex}}$. □

Remark. So every WMSO-definable relation R of $\mathfrak{T}_2$ is MSO-definable in $\mathfrak{T}_2$ (simply relativise the set quantifiers). In particular the following results also hold with WMSO replacing MSO.

For a cardinal κ let $\exists^{\geqslant\kappa}$ denote the quantifier 'there exists at least κ many sets X such that ...'. Write MSO($\exists^{\geqslant\kappa}$) for MSO enriched by the quantifier $\exists^{\geqslant\kappa}$.

Proposition 2.4. *[4] For every* MSO($\exists^{\geqslant\aleph_0}$) *formula $\phi(\overline{X})$ there is an* MSO *formula $\phi'(\overline{X})$ equivalent to $\phi(\overline{X})$ over $\mathfrak{T}_2$.*

Proof. The following are equivalent:

(1) There are only finitely many X satisfying $\phi(X, \overline{Y})$.

(2) There is a finite set Z such that every pair of different sets X_1, X_2 which both satisfy $\phi(X_i, \overline{Y})$ differ on Z.

The second condition can be expressed in MSO using lemma 2.3. □

Theorem 2.5. *[4] For every* MSO($\exists^{\geqslant\aleph_1}$) *formula $\phi(\overline{X})$ there is an* MSO *formula $\phi'(\overline{X})$ equivalent to $\phi(\overline{X})$ over $\mathfrak{T}_2$.*

Proof. The proof uses the composition method and basic ideas from descriptive set theory. We sketch a proof of the simpler case of $\mathfrak{T}_1$ following [42]. Say that two subsets of $\mathbb{N}$ have the *same end* if their symmetric difference is finite. There is a constant K (that depends only on ϕ) such that the following are equivalent for all $\overline{Y}$:

(1) There are uncountably many X satisfying $\phi(X, \overline{Y})$.
(2) There are K many sets X, each satisfying $\phi(X, \overline{Y})$, and that pairwise have different ends.

The second condition can be expressed in MSO. We argue correctness. The forward direction follows since each end class is countable. For the reverse let K be larger than the number of states of an ω-automaton for ϕ. The idea is that if there are too many sets with different ends we can find two that behave the same and so shuffle these to get uncountably many. For ease of writing assume that $\overline{Y}$ is a singleton and so write Y instead. Write ρ_i for an accepting run of the automaton on $\otimes(\chi_{X_i}, \chi_Y)$. There are two sets, say X_1 and X_2, and an infinite set $H \subset \mathbb{N}$ such that $\rho_1[n] = \rho_2[n]$ for all $n \in H$ (otherwise from some point on all pairs of runs disagree contradicting that the automaton has $< K$ states). Without loss we can assume, by passing to an infinite subset if required, that for all $n < m \in H$ with no element of H between them, both $\rho_1[n, m]$ and $\rho_2[n, m]$ mention final states and $x_1[n, m-1] \neq x_2[n, m-1]$. List H as $h_1 < h_2 < h_3 < \cdots$. By knitting segments of the runs we see that the automaton accepts every string of the form $\otimes(x_1[0, h_1 - 1]z_1z_2z_3 \cdots, \chi_Y)$ where $z_n \in \{x_1[h_n, h_{n+1} - 1], x_2[h_n, h_{n+1} - 1]\}$. This gives uncountably many distinct sets X satisfying $\phi(X, Y)$. □

Write MSO($\exists^{\mathrm{mod}}$) for MSO enriched by all quantifiers parameterised by $k, m \in \mathbb{N}$ of the form 'exists a set X, whose cardinality is congruent to k modulo m, such that ...'. The proof of the following theorem can be adapted from [36] or [42].

Theorem 2.6. *For every* MSO($\exists^{\mathrm{mod}}$) *formula* $\phi(\overline{X})$ *there is an* MSO *formula* $\phi'(\overline{X})$ *equivalent to* $\phi(\overline{X})$ *over* $\mathfrak{T}_2$.

Remark 2.7. Since $\mathfrak{T}_r$ (for $r < \omega$) is MSO-interpretable in $\mathfrak{T}_2$, the results above hold, for instance, with $\mathfrak{T}_1$ replacing $\mathfrak{T}_2$.

2.3 Interpretations

A good reference for interpretations is [30]. Let $\mathcal{L}$ denote FO or MSO or WMSO. Let $\mathcal{I} = (\delta, \Phi_1, \cdots, \Phi_N)$ be $\mathcal{L}$-formulas in the signature of a structure $\mathfrak{B}$ in which the free variables are individual variables. Suppose that δ has 1 free variable and Φ_i has r_i free variables. If $\Phi_i^{\mathfrak{B}}$ are relations over $\delta^{\mathfrak{B}}$ then we can define the structure $\mathcal{I}(\mathfrak{B}) := (\delta^{\mathfrak{B}}, \Phi_1^{\mathfrak{B}}, \cdots, \Phi_N^{\mathfrak{B}})$. This structure is said to be *$\mathcal{L}$-definable in* $\mathfrak{B}$. The tuple $\mathcal{I}$ is called an *$\mathcal{L}$-definition*.

Remark. We have overloaded the phrase '$\mathcal{L}$-definable': here one structure is definable in another; and the earlier meaning is that a relation is definable in a structure.

The following lemma says, in particular, that every relation FO-definable in $\mathcal{I}(\mathfrak{B})$ is FO-definable in $\mathfrak{B}$.

Lemma 2.8. *Fix an $\mathcal{L}$-definition $\mathcal{I}$ and an $\mathcal{L}$-formula $\Phi(x_1, \cdots, x_k)$ in the signature of $\mathcal{I}(\mathfrak{B})$. There is an $\mathcal{L}$-formula $\Phi_{\mathcal{I}}(x_1, \cdots, x_k)$ in the signature of $\mathfrak{B}$ such that for all elements b_i of $\delta^{\mathfrak{B}}$,*

$$\mathcal{I}(\mathfrak{B}) \models \Phi(b_1, \cdots, b_k) \textit{ if and only if } \mathfrak{B} \models \Phi_{\mathcal{I}}(b_1, \cdots, b_k).$$

Proof. The idea is to relativise all quantifiers to δ and replace and the ith atomic formula R_i in the signature of $\mathcal{I}(\mathfrak{B})$ by Φ_i. Formally, define $\Phi_{\mathcal{I}}$ inductively by $(R_i)_{\mathcal{I}} := \Phi_i$ and for the other cases:

$$\begin{array}{ll} (\Psi \wedge \Xi)_{\mathcal{I}} := \Psi_{\mathcal{I}} \wedge \Xi_{\mathcal{I}} & (\neg\Psi)_{\mathcal{I}} := \neg\Psi_{\mathcal{I}} \\ (\exists x_i \Psi)_{\mathcal{I}} := \exists x_i[\delta(x_i) \wedge \Psi_{\mathcal{I}}] & (\exists X_i \Psi)_{\mathcal{I}} := \exists X_i[(\forall x \in X_i \delta(x)) \wedge \Psi_{\mathcal{I}}] \\ (x \in X)_{\mathcal{I}} := (x \in X) & (X \subseteq Y)_{\mathcal{I}} := (X \subseteq Y). \end{array}$$

□

Definition 2.3. Let $\mathcal{I}$ be an $\mathcal{L}$-definition. If $\mathfrak{A}$ is isomorphic to $\mathcal{I}(\mathfrak{B})$, say via f, then say that $\mathfrak{A}$ *is $\mathcal{L}$-interpretable in $\mathfrak{B}$ via co-ordinate map f.*

Proposition 2.9. *Suppose that $\mathfrak{A}$ is $\mathcal{L}$-interpretable in a structure with decidable $\mathcal{L}$-theory. Then $\mathfrak{A}$ has decidable $\mathcal{L}$-theory.*

Proof. For a sentence ϕ of $\mathfrak{A}$, lemma 2.8 produces a sentence $\phi^{\mathcal{I}}$ preserving truth. Apply the given decision procedure to $\phi^{\mathcal{I}}$. □

For example for $r < \omega$ the structure $\mathfrak{T}_r$ is MSO-interpretable in $\mathfrak{T}_2$ and consequently the MSO-theory of $\mathfrak{T}_r$ is decidable.

There are other types of interpretations that could be introduced here, namely (finite)-set interpretations [25] [19]. For the sake of filling in undefined notions in the introduction we define a *(finite)-set interpretation* of $\mathfrak{A}$ in $\mathfrak{B}$ to be like a (W)MSO-interpretation except that the free variables are (finite) set variables. Thus elements of $\mathfrak{A}$ are coded by (finite) subsets of the domain of $\mathfrak{B}$.

3 Automatic structures

3.1 Rabin-automatic structures

MSO-definability can be thought of as FO-definability in a powerset structure. We use this idea to define the automatic structures. Let $\mathcal{P}(X)$ denote the set of subsets of X.

Definition 3.1. [19] The *power set* of a structure $\mathfrak{A} = (\mathcal{A}, R_1, \cdots, R_N)$ is the structure

$$\mathcal{P}[\mathfrak{A}] := (\mathcal{P}(\mathcal{A}), R_1', \cdots, R_N', \subseteq)$$

where

$$R_i' := \{(\{x_1\}, \cdots, \{x_{r_i}\}) \mid (x_1, \cdots, x_{r_i}) \in R_i\}.$$

For example, $\mathcal{P}[(\mathbb{N}, +1)]$ is the structure with domain $\mathcal{P}(\mathbb{N})$, the subset relation $\subseteq$, and the binary relation $\{(\{n\}, \{n+1\}) \mid n \in \mathbb{N}\}$. The following says that FO definability in the power set of a structure is the same as MSO definability in the structure.

Proposition 3.1. *For every FO-formula $\phi(\overline{x})$ there is an MSO-formula $\Psi(\overline{X})$ (and vice versa) such that for all structures $\mathfrak{A}$ and all $U_i \in \mathcal{P}(\mathcal{A})$*

$$\mathcal{P}[\mathfrak{A}] \models \phi(\overline{U}) \text{ if and only if } \mathfrak{A} \models \Psi(\overline{U}).$$

As a consequence we can transfer MSO-decidability of $\mathfrak{T}_r$ to FO-decidability of the power set structure.

Corollary 3.2. *The FO-theory of $\mathcal{P}[\mathfrak{T}_r]$ is decidable.*

Since FO-interpretations preserve decidability we make the following definition.

Definition 3.2. [9] A structure FO-interpretable in $\mathcal{P}[\mathfrak{T}_2]$ is called *Rabin-automatic* or *ω-tree automatic*. The collection of these structures is written ω`T-AutStr`.

Note that any structure that is FO bi-interpretable with $\mathcal{P}[\mathfrak{T}_2]$ could be used in this definition. For instance we can use the following structure [9]: the domain consists of all finite and infinite trees[5] and the atomic relations are

$$\preceq_{\texttt{ext}}, \equiv_{\text{dom}}, (\text{suc}_a^d)_{a \in \{0,1\}}^{d \in \{0,1\}}, \epsilon_0, \epsilon_1, \mathcal{F}$$

where $T \preceq_{\texttt{ext}} S$ if $\text{dom}(T) \subseteq \text{dom}(S)$ and $S(w) = T(w)$ for $w \in \text{dom}(T)$; $T \equiv_{\text{dom}} S$ if $\text{dom}(T) = \text{dom}(S)$; $\text{suc}_a^d(T) = S$ if the finite tree S is formed from the finite tree T by extending its leaves in direction d and labeling each new such node by a; ϵ_a is the tree with domain $\{\lambda\}$ labelled a; and $\mathcal{F}$ is the set of finite trees.

Corollary 3.3. *The FO-theory of every Rabin-automatic structure is decidable.*

The elements of the Rabin-automatic structure are naturally viewed as ω-trees.

Definition 3.3. For a Rabin-automatic structure $\mathfrak{A}$ (isomorphic via f to a FO-definable structure $\mathcal{I}(\mathcal{P}[\mathfrak{T}_2])$) and any relation $R \subseteq \mathcal{A}^k$, denote by $\text{code}(R)$ the set of ω-trees

$$\{\otimes(\chi_{f(a_1)}, \cdots, \chi_{f(a_k)}) \mid (a_1, \cdots, a_k) \in R\}.$$

The following is called the *fundamental theorem of automatic structures* and says that FO-definable relations in Rabin-automatic structures are, modulo coding, regular.

Theorem 3.4 (Fundamental theorem, cf. [35, 11]). *Let $\mathfrak{A}$ be Rabin-automatic.*

(1) For every first-order definable relation R in $\mathfrak{A}$ the set of trees $\text{code}(R)$ *is recognised by an ω-tree automaton.*

(2) The first-order theory of a Rabin-automatic structure is decidable.

[5] Here a tree is a function T from a prefix-closed subset of $\{0,1\}^*$ to $\{0,1\}$.

Proof. For the first item apply Proposition 3.1 to get that $f(R)$, being FO-definable in $\mathcal{P}[\mathfrak{T}_2]$, is MSO definable in $\mathfrak{T}_2$. Now apply Rabin's theorem. For the second item use the fact that it is decidable whether or not an ω-tree automaton accepts some tree. □

Since formulas are seen as automata, we may view a Rabin-automatic structure as being presented by automata.

Definition 3.4 (ω-tree automatic presentation [11]). Suppose that $f : \mathfrak{A} \simeq (B, S_1, \cdots, S_N)$ and

(1) the elements of B are $\{0, 1\}$-labeled trees;
(2) the set B is recognised by an ω-tree automaton, say M_B;
(3) the set $\{\otimes(\bar{t}) \mid \bar{t} \in S_i\}$ is recognised by an ω-tree automaton, say M_i, for $i \leqslant N$.

Then the data $\langle(M_B, M_1, \cdots, M_N), f\rangle$ is called a *ω-tree automatic presentation* of $\mathfrak{A}$.

The following characterisation is typically taken as a definition of a structure being Rabin-automatic.

Proposition 3.5 (Machine theoretic characterisation [11]). *A structure $\mathfrak{A}$ is FO-interpretable in $\mathcal{P}[\mathfrak{T}_2]$ if and only if $\mathfrak{A}$ has an ω-tree automatic presentation.*

Proof. The forward direction follows from the fundamental theorem 3.4. For the reverse direction convert regular sets and relations in the presentation into MSO formulas over $\mathfrak{T}_2$ (using Rabin's theorem) and then into FO-formulas using proposition 3.1. □

Example 3.1. The following structures are Rabin-automatic.

(1) The power set structure $\mathcal{P}[\mathfrak{T}_r]$ ($r \geqslant 1$) as well as its substructure $\mathcal{P}_f[\mathfrak{T}_r]$ whose domain consists of the finite subsets of $\mathfrak{T}_r$.
(2) The power set structure of the ordering of the rationals, namely $\mathcal{P}[(\mathbb{Q}, <)]$.
(3) Presburger arithmetic $(\mathbb{N}, +)$.
(4) Skolem arithmetic $(\mathbb{N}, \times)$.
(5) The structure $(\mathcal{P}(\mathbb{N}), \subseteq, =^*)$ where $X =^* Y$ means that X and Y have finite symmetric difference.
(6) Every ordinal $(\alpha, <)$ where $\alpha < \omega^{\omega^\omega}$.

Extension of the fundamental theorem We now rephrase the results on additional quantifiers on MSO into FO. We overload notation so that, for example, $\exists^\kappa$ denotes the quantifier 'there exists at least κ many individual elements x such that'. The following theorem is now immediate. Instances of it can be found in [11, 36, 42].

Theorem 3.6 (Extension of fundamental theorem). *Let $\mathfrak{A}$ be Rabin-automatic.*

(1) For every $\mathrm{FO}(\exists^{\geqslant\aleph_0}, \exists^{\geqslant\aleph_1}, \exists^{\mathrm{mod}})$*-definable relation R in $\mathfrak{A}$ the set* $\mathrm{code}(R)$ *is recognised by an ω-tree automaton.*
(2) The $\mathrm{FO}(\exists^{\geqslant\aleph_0}, \exists^{\geqslant\aleph_1}, \exists^{\mathrm{mod}})$ *theory of $\mathfrak{A}$ is decidable.*

How far can we push this? First we need a rigourous definition of quantifier. This is neatly provided by Lindström's definition of 'generalised quantifier', see [46]. We don't have a clear picture of those generalised quantifiers that can be added to FO and still get the properties as in Theorem 3.6. However here is a special case. Define the *cardinality quantifier parameterised by* C, for C a class of cardinals, as 'there exists exactly α many elements x such that ..., where $\alpha \in C$'. Examples include $\exists, \exists^{\mathrm{mod}}, \exists^{\geqslant \kappa}$, and 'there exist a prime number of elements such that ...'.

It turns out that the only cardinality quantifiers we can add to FO and still get the fundamental theorem are, essentially, the ones mentioned in Theorem 3.6.

Theorem 3.7 (cf. [58]). *Let* Q_C *be a cardinality quantifier parameterised by* C. *Suppose for every* $\mathfrak{A} \in \omega\texttt{T-AutStr}$ *and every* FO(Q)*-definable relation* R *in* $\mathfrak{A}$*, the set* $\mathrm{code}(R)$ *is recognised by a Rabin-automaton. Then every* FO(Q)*-definable relation is already* FO$(\exists^{\geqslant \aleph_0}, \exists^{\geqslant \aleph_1}, \exists^{\mathrm{mod}})$*-definable in* $\mathfrak{A}$.

Proof. We illustrate the proof for a set $C \subset \mathbb{N}$ of finite cardinals. Consider the Rabin-automatic presentation of $\mathfrak{A} := (\mathbb{N}, \leqslant)$ in which $n \in \mathbb{N}$ is coded by the set $\{0^n\} \subset \{0,1\}^*$. Since the set $C \subset \mathbb{N}$ is FO(Q_C)-definable in $\mathfrak{A}$, $\mathrm{code}(C)$ is is recognised by a tree-automaton. But the trees in $\mathrm{code}(C)$ are essentially unary strings and so the language $\mathrm{code}(C)$ is ultimately periodic. So C is already FO$(\exists^{\mathrm{mod}})$-definable in $\mathfrak{A}$. □

What about extensions of FO by set quantification? Unfortunately WMSO is too much to hope for. Since the configuration graph $\mathfrak{G}$ of a Turing machine (with the single-transition edge relation) is automatic, and reachability is expressible in WMSO, the halting problem reduces to the WMSO-theory of a certain $\mathfrak{G}$. It is a research programme to understand which quantifiers can be added to automatic structures and retain decidability, see [43, 57].

3.2 Other classes of automatic structures

We introduce subclasses of $\omega\texttt{T-AutStr}$ related to automata on finite strings/trees and infinite strings. Each has a machine theoretic charactersation as in Definition 3.4. A member of any of the four classes is said to be *automatic*.

Finite-string automatic structures Recall $\mathcal{P}_f(\mathfrak{A})$ is the substructure of $\mathcal{P}(\mathfrak{A})$ restricted to the finite subsets of $\mathcal{A}$.

Definition 3.5. A structure is called *finite-string automatic* if it is FO-interpretable in $\mathcal{P}_f[\mathfrak{T}_1]$. This collection of structures is written $\texttt{S-AutStr}$.

We have seen in Corollary 1.2 that $(\mathbb{N}, +)$ is *finite-string automatic*. Note that such structures have countable domain.

Example 3.2. The following structures are bi-interpretable with $\mathcal{P}_f[\mathfrak{T}_1]$:

(1) For $|\Sigma| \geqslant 2$, $(\Sigma^*, \{\sigma_a\}_{a \in \Sigma}, \prec_{\texttt{prefix}}, \mathrm{el})$ where σ_a holds on pairs (w, wa), el holds on pairs (u, v) such that $|u| = |v|$, and $\prec_{\texttt{prefix}}$ is the prefix relation.

(2) For $k \geqslant 2$, $(\mathbb{N}, +, |_k)$ where $|_k$ is the binary relation on $\mathbb{N}$ with $x|_k y$ if x is a power of k and x divides y.

The following definition turns a finite set Y into a finite string χ_Y as in $\chi_{\{1,3\}} = 0101$.

Definition 3.6 (characteristic finite-string). For a finite set $Y \subset \mathbb{N}$ define its *characteristic string* χ_Y as the $\{0,1\}$-labeled string of length $\max_{y \in Y} y + 1$ with a 1 in position n if and only if $n \in Y$.

Definition 3.7 (convoluting finite strings). Let $\overline{w} = (w_1, \cdots, w_k)$ be a k-tuple of $\{0,1\}$-labelled finite strings. Let $l := \max_i |w_i|$. The *convolution* $\otimes(\overline{w})$ is the $\{0,1,\Box\}^k$-labelled string of length l such that for all positions $n \leqslant l$ the ith component of $\otimes(\overline{w})[n]$ is equal to $w_i[n]$ if $n \leqslant |w_i|$ and the blank symbol $\Box$ otherwise.

Definition 3.8. Suppose f is an isomorphism witnessing $\mathfrak{A} \in \texttt{S-AutStr}$. For a relation $R \subseteq \mathcal{A}^k$ denote by $\mathrm{code}(R)$ the set of finite-strings $\{\otimes(\chi_{f(a_1)}, \cdots, \chi_{f(a_k)}) \mid \overline{a} \in R\}$.

Just as for Rabin-automatic structures, there is a fundamental theorem for finite-string automatic structures. We do not state it in full; simply replace Rabin-automatic by finite-string automatic and ω-tree automata by finite-string automata. However, we do slightly generalise the analogous definition of automatic presentation to cover an arbitrary alphabet Σ.

Definition 3.9 (finite-string automatic presentation [35]). Fix a finite alphabet Σ. Suppose that $f : \mathfrak{A} \simeq (\mathcal{B}, S_1, \cdots, S_N)$ and

(1) the elements of $\mathcal{B}$ are finite strings from Σ^*;
(2) the set $\mathcal{B}$ is recognised by a finite-string automaton, say $M_{\mathcal{B}}$;
(3) the set $\{\otimes(\overline{t}) \mid \overline{t} \in S_i\}$ is recognised by a finite-string automaton, say M_i, for $i \leqslant N$.

The data $\langle (M_{\mathcal{B}}, M_1, \cdots, M_N), f\rangle$ is called a *finite-string automatic presentation* of $\mathfrak{A}$.

Proposition 3.8 (Machine theoretic characterisation [11]). *A structure is FO-interpretable in $\mathcal{P}_f[\mathfrak{T}_1]$ if and only if it has a finite-string automatic presentation over an alphabet with $|\Sigma| \geqslant 2$.*

Proof. If $|\Sigma| > 2$ recode a string w by replacing individual symbols with binary blocks of size $\log_2 |\Sigma|$. □

Relationships amongst the classes of automatic structures There are two more standard classes of automatic structures, see [9]. A *Büchi-automatic structure* is one FO-interpretable in $\mathcal{P}[\mathfrak{T}_1]$. Collectively these are denoted $\omega\texttt{S-AutStr}$. An *ω-string* is a function from $\mathbb{N}$ to $\{0,1\}$ and automata operating on (convolutions of tuples of) these are called *ω-string automaton*. One can similarly characterise ω-string automatic structures as those with *ω-string automatic presentations*. For example, $([0,1), +, <) \in \omega\texttt{S-AutStr}$ where $+$ is taken modulo 1 (the usual binary coding works). A *finite-tree automatic structure* is one FO-interpretable in $\mathcal{P}_f[\mathfrak{T}_2]$. Collectively these are denoted $\texttt{T-AutStr}$.

A *finite tree* is a function from a finite prefix closed subset of $\{0,1\}^*$ to $\{0,1\}^*$ and automata operating on (convolutions of tuples of) these are called *finite-tree automata*. These automatic structures are those with *finite-tree automatic presentations*. For example $(\mathbb{N}, \times) \in$ `T-AutStr` (decompose n $n = \prod_i p_i^{e_i}$ where p_i is the ith prime and so code n as a tree with e_i written in binary on branch $0^i 1^*$). For the case of unranked finite-tree automatic structures see [7].

Proposition 3.9. `S-AutStr` *is a proper subset of* ω`S-AutStr` *and of* `T-AutStr`*, each of which is a proper subset of* ω`T-AutStr`.

Proof. The inclusions follow since finite-strings are special cases of finite-trees, etc. The structure of Skolem arithmetic $(\mathbb{N}, \times)$ separates `T-AutStr` from `S-AutStr` (see [9]). We will see (Theorem 6.9) that a structure separating ω`T-AutStr` from ω`S-AutStr` is $(\mathcal{P}(\{l, r\}^\star), \subset, V)$ where V is the unary relation consisting of those sets X such that the characteristic tree of X has the property that every infinite path is labelled with only finitely many 1s. □

Proposition 3.10. *If* $\mathfrak{A} \in \omega$`S-AutStr` *is countable then* $\mathfrak{A} \in$ `S-AutStr`. *If* $\mathfrak{A} \in \omega$`T-AutStr` *is countable then* $\mathfrak{A} \in$ `T-AutStr`.

Proof. The reason is that every ω-string automaton whose language is countable accepts only ultimately periodic strings with a uniform bound on the length of the periods. A similar conditions holds for ω-tree automata that accept countable languages. □

3.3 Operations on automatic structures

This section asks and gives some basic answers to the question:

> Which operations on automatic structures preserve automaticity?

See [10] for a survey of operations that preserve decidability.

Closure under interpretations Let $\circ$ stand for 'finite-string', 'ω-string', 'finite-tree', or 'ω-tree'. Since FO-definitions compose we have that:

Proposition 3.11. *[9] The* $\circ$*-automatic structures are closed under FO-interpretations.*

Example 3.3. $\circ$-automatic structures are closed under FO-definable expansions: if $\mathfrak{A}$ is $\circ$-automatic and ϕ is a FO-formula over the signature of $\mathfrak{A}$ then $(\mathfrak{A}, \phi^{\mathfrak{A}})$ is $\circ$-automatic.

There is a more general notion of interpretation called a *FO-interpretation of dimension* d. Here δ has $d \in \mathbb{N}$ free variables and each Φ_i has $d \times r_i$ free variables.

Proposition 3.12. *[11] The* $\circ$*-automatic structures are closed under FO-interpretations of dimension d.*

Proof. We illustrate the idea for `S-AutStr`. Suppose $\mathfrak{B} \in$ `S-AutStr`. It is enough to show that if $\mathfrak{A} = \mathcal{I}(\mathfrak{B})$ then we can find a finite-string automatic presentation of $\mathfrak{A}$. An element a of $\mathfrak{A}$ is a d-tuple of elements $(b_1, \cdots, b_d)$ of $\mathfrak{B}$ each of which is coded by a finite string $\mathrm{code}(b_i)$. Coding the element a by the string $\otimes(\mathrm{code}(b_1), \cdots, \mathrm{code}(b_d))$ we get a finite-string automatic presentation of $\mathfrak{A}$ over alphabet $\{0, 1, \Box\}^d$. By proposition 3.8 $\mathfrak{A}$ is in `S-AutStr`. □

Say $\mathfrak{A}$ and $\mathfrak{B}$ are FO-interpretable in $\mathfrak{U}$. Then the disjoint union of $\mathfrak{A}$ and $\mathfrak{B}$ is 2-dimensionally interpretable in $\mathfrak{U}$. Similarly for their direct product. Thus $\circ$-automatic structures are closed under disjoint union and direct product.

The *(weak) direct power* of $\mathfrak{A}$ is a structure with the same signature as $\mathfrak{A}$, its domain consists of (finite) sequences of $\mathcal{A}$, and the interpretation of a relation symbol R is the set of sequences σ such that $R^{\mathfrak{A}}(\sigma(n))$ holds for all n. For example the weak direct power of $(\mathbb{N}, +)$ is isomorphic to $(\mathbb{N}, \times)$; the isomorphism sends n to the finite sequence $(e_i)_i$ where $\prod p_i^{e_i}$ is the prime power decomposition of n. Since $(\mathbb{N}, \times)$ is neither in `S-AutStr` nor ω`S-AutStr`, these string classes are not closed under weak direct power (or direct power).

Proposition 3.13. *[9] Each of* `T-AutStr` *and* ω`T-AutStr` *is closed under under weak direct power. The class* ω`T-AutStr` *is closed under direct power.*

Proof. We illustrate the second statement. Let $\mathfrak{A}$ be a Rabin-automatic structure with relation symbol R. Let $\sigma = (a_n)_n$ be an element of the direct power of $\mathfrak{A}$. Code the sequence σ by the tree t_σ whose subtree at 0^n1 is the tree $\mathrm{code}(a_n)$. The interpretation of R in the direct power is recognised by a tree automaton: it processes t_σ by checking that the convolution of the subtrees rooted at 0^n1 is recognised by the automaton for $R^{\mathfrak{A}}$. □

Closure under quotients Let $\mathfrak{A} = (\mathcal{A}, R_1, \cdots, R_N)$ be a structure. An equivalence relation ϵ on the domain $\mathcal{A}$ is called a *congruence for* $\mathfrak{A}$ if each relation R_i satisfies the following property: for every pair of r_i-tuples $\overline{a}, \overline{b}$ of elements of $\mathcal{A}$, if $(a_j, b_j) \in \epsilon$ for $1 \leqslant j \leqslant r_i$ then $R_i(\overline{a})$ if and only if $R_i(\overline{b})$. The *quotient of* $\mathfrak{A}$ *by* ϵ, written $\mathfrak{A}/\epsilon$ is the structure whose domain is the set of equivalence classes of ϵ and whose ith relation is the image of R_i by the map sending $u \in \mathcal{A}$ to the equivalence class of u. We ask and give partial answers to the following question:

If $(\mathfrak{A}, \epsilon)$ is $\circ$-automatic, is $\mathfrak{A}/\epsilon$ $\circ$-automatic?

`S-AutStr`**:** Yes [11]. There is a regular well-ordering of the set of finite strings, for instance the length-lexicographic ordering $<_{\text{llex}}$. Use this order to define a regular set D of unique ϵ-representatives. Then restrict the presentation of $\mathfrak{A}$ to D to get a presentation of $\mathfrak{A}/\epsilon$.

`T-AutStr`**:** Yes [19]. Except in the finite string case, there is no regular well ordering of the set of all finite trees. However one can still convert a finite-tree automatic presentation of $(\mathfrak{A}, \epsilon)$ into one for $\mathfrak{A}/\epsilon$. The idea is to associate with each tree t a new tree $\hat{t}$ of the following form: the domain is the intersection of the prefix-closures of the domains of all trees that are ϵ-equivalent to t; a node is labelled σ if t had label σ in that position; a leaf x is additionally labelled by those states q from which the automaton for ϵ accepts the pair consisting of the subtree of t rooted at x and the tree with empty domain. Using transitivity and symmetry of ϵ, if $\hat{t} = \hat{s}$ then t is ϵ-equivalent to s. Moreover each equivalence class is associated with finitely many new trees, and so a representative may be chosen using any fixed regular linear ordering of the set of all finite trees.[6]

$\boldsymbol{\omega}$S-AutStr: It depends. Kuske and Lohrey [42] observed that there is no unique set of representatives of the equal almost-everywhere relation $\sim_{\text{ae}}$ that is regular. Thus we can't quotient using the trick that worked for S-AutStr. In fact, there is a structure in ωS-AutStr whose quotient is not ω-string automatic [29]. The proof actually shows that the structure has no Borel presentation, see Theorem 6.10. However, every regular ω-string equivalence relation with countable index has an ω-regular set of unique representatives. This follows from the following more general result [5]: If ϵ has countable index on A then there exist finitely many ω-strings $x_1, \cdots, x_c$ so that every x in A is $\approx$-equivalent to some y which is $\sim_{\text{ae}}$-equivalent to some x_i. Thus if $\mathfrak{A}/\epsilon$ is countable and $(\mathfrak{A}, \epsilon) \in \omega$S-AutStr then $\mathfrak{A}/\epsilon \in \omega$S-AutStr.

$\boldsymbol{\omega}$T-AutStr: A finer analysis shows that the quotient of the Rabin-automatic structure from the previous paragraph is not in ωT-AutStr [29]. Thus ωT-AutStr is not closed under quotients. However it is not yet known what happens in the case that congruence-relation has countable index.

Nonetheless, quotients still have decidable theory since to decide the truth in the quotient structure, for a given sentence replace $=$ by ϵ and apply the decision procedure for $(\mathfrak{A}, \epsilon)$.

Proposition 3.14. *[19] If $(\mathfrak{A}, \epsilon)$ is $\circ$-automatic then the quotient $\mathfrak{A}/\epsilon$ has decidable FO-theory.*

Countable elementary substructures In this section we show a way of producing, from an uncountable Büchi- or Rabin-automatic structure $\mathfrak{B}$, a countable substructure $\mathfrak{A}$ with the same theory. Thus although $\mathfrak{A}$ may not itself be automatic, it has decidable theory.

Example 3.4. This example is taken from [20] (pg. 106). Code reals in base 2 and so get an ω-string automatic presentation of $\mathfrak{A} := (\mathbb{R}, +, <)$. The substructure of $\mathfrak{A}$ consisting of those reals coded by ultimately periodic ω-strings is isomorphic to $\mathfrak{B} := (\mathbb{Q}, +, <)$. We will see that the FO-theory of these two structures are identical and so conclude that $(\mathbb{Q}, +, <)$ has decidable FO-theory. A recent breakthrough establishes that $(\mathbb{Q}, +)$ is not in S-AutStr [65]. For further work towards classifying the torsion-free abelian groups that are in S-AutStr, see [59]. It is not known whether or not $(\mathbb{Q}, +, <) \in \omega$T-AutStr.

[6] The construction given in [19] is slightly more general and allows one to effectively factor finite-subset interpretations in any tree.

Two structures with the same FO-theory are called *elementary equivalent*. Let $\mathfrak{A}, \mathfrak{B}$ have the same signature. Say that $\mathfrak{A}$ is an *elementary substructure* of $\mathfrak{B}$ if $\mathcal{A} \subseteq \mathcal{B}$ and for all formulas $\phi(\overline{x})$ and all $\overline{a}$ from $\mathcal{A}$,

$$\mathfrak{A} \models \phi(\overline{a}) \text{ if and only if } \mathfrak{B} \models \phi(\overline{a}) \quad (\dagger)$$

Then in particular: $\mathfrak{A}$ and $\mathfrak{B}$ are elementary equivalent (take ϕ to be a sentence) and $\mathfrak{A}$ is a substructure of $\mathfrak{B}$ (they agree on the atomic relations of $\mathfrak{A}$). There is a simple characterisation of being an elementary substructure.

Lemma 3.15 (Tarski-Vaught [30]). *Let $\mathfrak{A}$ be a substructure of $\mathfrak{B}$. Then $\mathfrak{A}$ is an elementary substructure of $\mathfrak{B}$ if and only if for every FO-formula $\phi(x, \overline{y})$ and all $\overline{a}$ from $\mathcal{A}$*

$$\mathfrak{B} \models \exists x \phi(x, \overline{a}) \implies \mathfrak{A} \models \exists x \phi(x, \overline{a}).$$

Say $f : \mathfrak{A} \simeq (\mathcal{B}, S_1, \cdots, S_N)$ is an ω-string automatic presentation of $\mathfrak{A}$. Write $\mathfrak{A}_{\mathrm{up}}$ for the substructure of $\mathfrak{A}$ isomorphic via f to the substructure whose domain consists of the ultimately periodic strings from $\mathcal{B}$. Similarly if $\mathfrak{A} \in \omega$`T-AutStr` define $\mathfrak{A}_{\mathrm{reg}}$ as the substructure of $\mathfrak{A}$ isomorphic via f to the substructure consisting of regular trees from $\mathcal{B}$.

Proposition 3.16. *[5, 29]*

(1) Let $\mathfrak{A} \in \omega$`S-AutStr`. The structure $\mathfrak{A}_{\mathrm{up}}$ is an elementary substructure of $\mathfrak{A}$.
(2) Let $\mathfrak{A} \in \omega$`T-AutStr`. The structure $\mathfrak{A}_{\mathrm{reg}}$ is an elementary substructure of $\mathfrak{A}$.

Proof. Use the fact that an automaton — possibly instantiated with ultimately periodic strings $\overline{a}$ — is non-empty only if it contains an ultimately periodic string. Similarly for the tree case with 'regular tree' replacing 'ultimately periodic string'. □

Similar reasoning shows that $\mathfrak{A}_{\mathrm{reg}}$ and $\mathfrak{A}_{\mathrm{up}}$ have decidable FO($\exists^{\geqslant \aleph_0}, \exists^{\mathrm{mod}}$)-theory.

4 Anatomy of structures

Not much is known about uncountable automatic structures. For instance, the only known technique for showing that a structure is not in ω`T-AutStr` is to show that its theory is sufficiently complex, for instance that the FO($\exists^{\geqslant \aleph_0}, \exists^{\geqslant \aleph_1}, \exists^{\mathrm{mod}}$)-theory is undecidable. In this section we ask and give partial answers to the question:

What do structures in `S-AutStr` or `T-AutStr` look like?

Here is a useful pumping observation:

Proposition 4.1. *[35] Suppose that the partial function $F : A^n \to A$ is finite-string/tree regular, and let p be the number of states of the automaton. If $\overline{x}$ is in the domain of F then the length/height of the string/tree $F(\overline{x})$ is at most p more than the length/height of the largest string/tree in $\overline{x}$.*

Proof. Otherwise, take a counterexample $\overline{x}$. After all of $\overline{x}$ has been read, and while still reading $F(\overline{x})$, some path in the run must have a repeated state. So the automaton also accepts infinitely many tuples of the form $(\overline{x}, \cdot)$ contradicting the functionality of F. □

Growth of generation

Definition 4.1. [35] Let $\mathfrak{A}$ be a structure with functions $f_1, \cdots, f_k$ of arities $r_1, \cdots, r_k$ respectively. Let $D \subset \mathcal{A}$ be a finite set. Define the *nth growth level*, written $G_n(D)$, inductively by $G_0(D) = D$ and $G_{n+1}(D)$ is the union of $G_n(D)$ and

$$\bigcup_{i \leqslant k} \{f_i(x_1, \cdots, x_{r_i}) \mid x_j \in G_n(D) \text{ for } 1 \leqslant j \leqslant r_i\}.$$

How fast does $|G_n(D)|$ grow as a function of n? For example, consider the free group with generating set $D = \{d_1, \cdots, d_m\}$. For $m \geqslant 2$ the set $G_n(D)$ includes all strings over D of length (in the generators) at most 2^n; so $|G_n(D)|$ is at least m^{2^n}.

Proposition 4.2. *[35] Let $\mathfrak{A} \in$ `T-AutStr` and $D \subset \mathcal{A}$ be a finite set. Then there is a linear function $t : \mathbb{N} \to \mathbb{N}$ so that for all $e \in G_n(D)$ the tree* code(e) *has height at most $t(n)$.*

Proof. Iterate proposition 4.1. □

Corollary 4.3. *If $\mathfrak{A} \in$ `T-AutStr` then $|G_n(D)| \leqslant 2^{2^{O(n)}}$. If $\mathfrak{A} \in$ `S-AutStr` then $|G_n(D)| \leqslant 2^{O(n)}$.*

Proof. Count the number of $\{0, 1\}$-labelled trees (strings) of height at most k. □

Thus the free group on more than one generator is not in `S-AutStr`.

Growth of projections We will be considering structures $(\mathfrak{A}, R)$ where R is a relation of arity > 1. For a tuple $\overline{u}$ of elements from $\mathcal{A}$ define $R(\cdot, \overline{u}) := \{a \in \mathcal{A} \mid (a, \overline{u}) \in R\}$.

Definition 4.2. For finite $E \subset \mathcal{A}$ the *shadow cast by $\overline{u}$ on E via R* is the set $R(\cdot, \overline{u}) \cap E$ and the *shadow count of E via R* is the number distinct shadows cast on E via R as $\overline{u}$ varies over tuples of elements of $\mathcal{A}$. We may suppress mention of R.

For example the random graph $(\mathcal{A}, R)$ has the property that for every pair of disjoint finite sets $E, F \subset \mathcal{A}$ there is a point $x \in \mathcal{A}$ that has an edge to every element in E and to no element of F. So for a given finite $E \subset \mathcal{A}$, the shadow count of E via R is the largest possible, namely $2^{|E|}$. The following propositions are due to Christian Delhommé [21] (the first one independently due to Frank Stephan) and limit the possible shadow counts in automatic structures.

Proposition 4.4. *Suppose $(\mathfrak{A}, R) \in$ `S-AutStr`. Then there is a constant k, that depends on the automata for domain $\mathcal{A}$ and R, and arbitrarily large finite subsets $E \subset \mathcal{A}$ such that the shadow count of E via R is at most $k|E|$.*

Proof. To simplify readability we suppose R is binary. Let $\mathcal{A}_n$ be the set of strings in $\mathcal{A}$ of length at most n. Let Q be the state set of the automaton for R. First, there is a constant $c := |Q|^{|Q|}$ such that for all n and all $x \in \mathcal{A}$ there is a $y \in \mathcal{A}_{n+c}$ such that x and y cast the same shadow on $\mathcal{A}_n$. Indeed, consider the sequence of functions $f_i : Q \to Q$ with $f_i(q)$ defined to be the state reached when the automaton for R starts in q and reads the word $\otimes(x[n+1, n+i], \lambda)$. If $|x| > n+c$ then there are two positions $k < l$ such that $f_k = f_l$. If we remove the segment $x[k, l-1]$ from x we get a shorter word x' that casts the same shadow on $\mathcal{A}_n$ as x does. Repeat until the word is short enough. Second, consider the sequence of sets $\mathcal{A}_{b+nc}$ where b is fixed so that $\mathcal{A}_b \neq \emptyset$. Write X_n for the cardinality of $\mathcal{A}_{b+nc}$ and S_n for the shadow count of $\mathcal{A}_{b+nc}$. We know that $S_n \leqslant X_{n+1}$. Suppose t were such that for almost all n, $S_n > tX_n$. Then $2^{b+(n+1)c} \geqslant X_{n+1} \geqslant S_n > t^n X_0$ for almost all n. So t is smaller than a constant that depends on b, c and X_0. So take k larger than this constant and conclude that $S_n \leqslant kX_n$ for infinitely many n, as required. □

The proof of the following proposition is similar.

Proposition 4.5. *Suppose* $(\mathfrak{A}, R) \in$ `T-AutStr`. *Then there is a constant k, that depends on the automata for* $\mathcal{A}$ *and R, and arbitrarily large finite subsets* $E \subset \mathcal{A}$ *such that the shadow count of E via R is at most* $|E|^k$.

These are used to show that certain structures are not automatic. An immediate application is that the random graph is not in `T-AutStr`.

Sum- and box-decompositions All definitions and results in this section are due to Delhommé [21].

Definition 4.3. Say that a structure $\mathfrak{B}$ is a *sum-decomposition* of a set of structures $\mathbf{C}$ (each having the same signature as $\mathfrak{B}$) if there is a finite partition of $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_n$ such that for each i the substructure $\mathfrak{B} \upharpoonright \mathcal{B}_i$ is isomorphic to some structure in $\mathbf{C}$.

Theorem 4.6. *Suppose* $(\mathfrak{A}, R) \in$ `S-AutStr`. *There is a finite set of structures* $\mathbf{C}$ *so that for every tuple of elements* $\overline{u}$ *from* $\mathcal{A}$, *the substructure* $\mathfrak{A} \upharpoonright R(\cdot, \overline{u})$ *is a sum-decomposition of* $\mathbf{C}$.

Proof. To simplify readability we suppose that $\mathfrak{A} = \langle \mathcal{A}, R, S\rangle$ and for each $T \in \{R, S\}$ fix a deterministic automaton $(Q_T, \iota_T, \Delta_T, F_T)$ recognising $\mathrm{code}(T)$. Naturally extend Δ_T to all words and so write $\Delta_T(q, w)$.

Given a tuple $\overline{u}$ write $|\overline{u}| := \max\{|u_i|\}$. Observe that we can partition the set $R(\cdot, \overline{u})$ into the finitely many sets: the singletons $\{c\}$ such that $R(c, \overline{u})$ and $|c| < |\overline{u}|$; as well as the sets

$$R^{a\{0,1\}^*}(\cdot, \overline{u}) := \{aw \in \mathcal{A} \mid (aw, \overline{u}) \in R, w \in \{0,1\}^*\}$$

where $|a| = |\overline{u}|$. There are finitely many isomorphism types amongst substructures of the form $\mathcal{A} \upharpoonright \{c\}$, for $c \in \mathcal{A}$. So, it is sufficient to show that as we vary the tuple $(a, \overline{u})$ subject to $|a| = |\overline{u}|$, there are finitely many isomorphism types amongst substructures of the form $\mathfrak{A} \upharpoonright R^{a\{0,1\}^*}(\cdot, \overline{u})$.

We do this by bounding the number of isomorphism types in terms of the number of states of the automata. To this end, define a function f as follows. Its domain consists of tuples $(a, \overline{u})$ satisfying $|a| = |\overline{u}|$; and f sends $(a, \overline{u})$ to the pair of states

$$\langle \Delta_R(\iota_R, \otimes(a, \overline{u})), \Delta_S(\iota_S, \otimes(a, \cdots, a))\rangle\,.$$

The range of f is bounded by $|Q_R| \times |Q_S|$; in particular, the range is finite.

To finish the proof, we argue that the isomorphism type of the substructure $\mathfrak{A} \restriction R^{a\{0,1\}^*}(\cdot, \overline{u})$ depends only on the value $f(a, \overline{u})$. This follows from the fact that if $f(a, \overline{u}) = f(a', \overline{u'})$, then the corresponding substructures are isomorphic via the mapping $aw \mapsto a'w$ ($w \in \{0,1\}^*$). For instance $(aw, \overline{u}) \in R$ if and only if the automaton for R starting in state $\Delta_R(\iota_R, \otimes(a, \overline{u}))$ and reading $\otimes(w, \lambda, \cdots, \lambda)$ reaches a final state if and only if starting in $\Delta_R(\iota_R, \otimes(a', \overline{u'}))$ and reading $\otimes(w, \lambda, \cdots, \lambda)$ it reaches a final state if and only if $(a'w, \overline{u'}) \in R$. This gives the bijective property. A similar argument shows that $(aw_1, \cdots, aw_s) \in S$ if and only if $(a'w_1, \cdots, a'w_s) \in S$. □

Corollary 4.7. *The ordinal* $(\omega^\omega, <)$ *is not in* S-AutStr.

Proof. Suppose for a contradiction that $(\omega^\omega, <)$ has an automatic presentation and let $\mathbf{C}$ be the finite set of structures guaranteed by the theorem using $<$ for R. Consider the following fact (proved by induction): if the domain of a well-order, isomorphic to some ordinal of the form ω^n for $n \in \mathbb{N}$, is partitioned into finitely many pieces $\{B_i\}_i$, then there is some i so that the substructure on domain B_i is isomorphic to ω^n. This means that $\mathbf{C}$ must contain (isomorphic copies of) $(\omega^n, <)$ for every $n \in \mathbb{N}$, contradicting the finiteness of $\mathbf{C}$. □

We now state the analogous results for T-AutStr.

Definition 4.4. Let $(\mathfrak{B}_i)_{i \in I}$ be a non-empty finite sequence of structures over the same signature. Their *synchronous product* is a structure over the same signature defined as follows. Its domain is $\prod_{i \in I} \mathcal{B}_i$. Write π_j for the projection $\prod_{i \in I} \mathcal{B}_i \to \mathcal{B}_j$. The interpretation of an r-ary relation symbol R consists of those tuples $(x_1, \cdots, x_r)$ such that for all $k \in I$,

$$\mathfrak{B}_k \models (\pi_k(x_1), \cdots, \pi_k(x_r)) \in R.$$

Theorem 4.8. *Suppose* $(\mathfrak{A}, R) \in$ T-AutStr. *There is a finite set of structures* $\mathbf{C}$ *so that for every tuple of elements* $\overline{u}$ *from* A*, the substructure* $\mathfrak{A} \restriction R^{\mathfrak{A}}(\cdot, \overline{u})$ *is a sum-decomposition of the set of all synchronous products of sequences from* $\mathbf{C}$.

Corollary 4.9. *The ordinal* $(\omega^{\omega^\omega}, <)$ *is not in* T-AutStr.

5 Equivalent presentations

In this section we focus on finite-string automatic presentations of a fixed structure. The main result in this section is due to Bárány [1]. We illustrate with base k presentations of $(\mathbb{N}, +)$ where $2 \leqslant k \in \mathbb{N}$. Write μ_k for the map sending $n \in \mathbb{N}$ to the base-k representation of n. The translation between bases p and q is the map $\mu_q \circ \mu_p^{-1}$. It sends a string in base-p to that string in base-q that represents the same natural number.

Call two bases p and q *multiplicatively dependent* if for some positive integers k, l

$$p^k = q^l.$$

Proposition 5.1. *If p and q are multiplicatively dependent, then every relation $R \subseteq \mathbb{N}^r$ is regular when coded in base p if and only if it is regular when coded in base q.*

To see this we may use semi-synchronous rational relations: these can be thought of as being recognised by a multi-tape automaton where each read-head advances at a different, but still constant, speed. In the following definition the ith head moves m_i symbols at a time.

Definition 5.1. Fix a finite alphabet Σ and a vector of positive integers $\underline{m} = (m_1, \cdots, m_r)$. Let $\square$ be a symbol not in Σ and write $\Sigma_\square$ for $\Sigma \cup \{\square\}$. For each component m_i introduce the alphabet $(\Sigma_\square)^{m_i}$. The *$\underline{m}$-convolution of a tuple* $(w_1, \cdots, w_r) \in (\Sigma^\star)^r$ is formed as follows. First, consider the intermediate string $(w_1\square^{a_1}, \cdots, w_r\square^{a_r})$ where the a_i are minimal such that there is some $k \in \mathbb{N}$ so that for all i, $|w_i| + a_i = km_i$. Second, partition each component $w_i\square^{a_i}$ into k many blocks of size m_i, and view each block as an element of $(\Sigma_\square)^{m_i}$. Thus the string $\otimes_{\underline{m}}(w_1, \cdots, w_r)$ is formed over alphabet $(\Sigma_\square)^{m_1} \times \cdots \times (\Sigma_\square)^{m_r}$. The *$\underline{m}$-convolution of a relation* $R \subseteq (\Sigma^\star)^r$ is the set $\otimes_{\underline{m}} R$ defined as

$$\{\otimes_{\underline{m}} \overline{w} \mid \overline{w} \in R\}.$$

A relation R is *$\underline{m}$-synchronous rational* if there is a finite automaton recognising $\otimes_{\underline{m}} R$. Call R *semi-synchronous* if it is $\underline{m}$-synchronous rational for some $\underline{m}$.

For example, if $\underline{m} = (1, \cdots, 1)$ then $\otimes_{\underline{m}}$ is the same as $\otimes$. For another example, the base-changing translation from base p to base q assuming $p^k = q^l$ is (k, l)-synchronous. Proposition 5.1 now follows from the fact that the image of a regular relation under a semi-synchronous transduction is regular. The converse of Proposition 5.1 is also true and follows from the Cobham-Semenov theorem, see [14]. For instance, if p and q are multiplicatively independent then the set of powers of p is regular in base-p but not regular in base-q. This discussion is the inspiration for the following generalisation.

For a given finite-string automatic presentation $\mu : \mathfrak{A} \simeq (\mathcal{B}, S_1, \cdots, S_N)$ write μ_{Reg} for the collection of relations

$$\{\mu^{-1}(R) \mid R \subseteq \mathcal{B}^k \text{ is a regular relation}, k \in \mathbb{N}\}.$$

Let $\nu : \mathfrak{A} \simeq (\mathcal{C}, R_1, \cdots, R_N)$ be another finite-string automatic presentation of $\mathfrak{A}$.

Definition 5.2. [1] The presentations μ and ν of $\mathfrak{A}$ are *equivalent* if $\mu_{\text{Reg}} = \nu_{\text{Reg}}$.

For instance if p and q are multiplicatively dependent then Proposition 5.1 says that the presentations μ_p and μ_q are equivalent.

Theorem 5.2. *[1] The presentations μ and ν are equivalent if and only if the map $\nu\mu^{-1} : \mathcal{B} \to \mathcal{C}$, namely*

$$\{(\mu(x), \nu(x)) \in \mathcal{B} \times \mathcal{C} \mid x \in \mathcal{A}\},$$

is semi-synchronous.

Proof. The interesting case is the forward direction. Let f denote the translation $\nu\mu^{-1} : \mathcal{B} \to \mathcal{C}$. Here is an outline: starting with $x \in \mathcal{B}$, we apply f to get $f(x) \in \mathcal{C}$, then pad to get $f'(x) \in \mathcal{C}'$, then cut into blocks to get $f''(x) \in \mathcal{C}''$. Write π for the padding $\mathcal{C} \to \mathcal{C}'$ and β for the blocking $\mathcal{C}' \to \mathcal{C}''$. Then f can be decomposed into semi-synchronous maps

$$\mathcal{B} \overset{f''}{\to} \mathcal{C}'' \overset{\beta^{-1}}{\to} \mathcal{C}' \overset{\pi^{-1}}{\to} \mathcal{C}.$$

We need some definitions. For a set X of strings, write $\mathbb{L}_X$ for the regular relation of pairs $(x, y) \in X \times X$ such that $|x| \geqslant |y|$. The *growth* of a function g between regular sets is the function $G : n \mapsto \max_{|a| \leqslant n} |g(a)|$. A bijection g is *length preserving* if $|g(x)| = |x|$. It is *length-monotonic* if $|x_1| \leqslant |x_2|$ implies $|g(x_1)| \leqslant |g(x_2)|$. It has *$\delta$-delay* if $|x_2| > |x_1| + \delta$ implies $|g(x_2)| > |g(x_1)|$.

Claim 1. There is a constant δ such that f has δ-delay.

Since $f^{-1}(\mathbb{L}_{\mathcal{C}}) := \{(a, b) \mid |f(a)| \geqslant |f(b)|\} \subseteq \mathcal{B} \times \mathcal{B}$ is regular (by assumption of equivalence) and locally finite (every a is related to finitely many bs) there is a δ (by a pumping argument as in proposition 4.1) such that $(a, b) \in f^{-1}(\mathbb{L}_{\mathcal{C}})$ implies $|b| \leqslant |a| + \delta$.

The next claim says that the strings shorter than x are not translated into strings that are more than a constant longer than the string x is translated into.

Claim 2. There is a constant K with $F(|x|) - |f(x)| \leqslant K$, where F is the growth of f.

Since $f(\mathbb{L}_{\mathcal{B}}) := \{(f(a), f(b)) \mid |a| \leqslant |b|\} \subseteq \mathcal{C} \times \mathcal{C}$ is regular and locally finite, $|a| \leqslant |x|$ implies $|f(a)| \leqslant |f(x)| + K$. Thus $F(|x|) := \max_{|a| \leqslant |x|} |f(a)|$ is at most $|f(x)| + K$.

Let $\natural$ be a new symbol. Define $f' : x \mapsto f(x)\natural^{F(|x|)-|f(x)|}$, for $x \in \mathcal{B}$. Thus f' pads $f(x)$ by $\natural$ symbols so that the length of the translation by f' of x is no smaller than the length of the translation of any shorter word than x. Since the property $\{x \mid F(|x|) - |f(x)| = i\}$ is regular for fixed i the graph of f' is regular (by the previous claim). Write $\mathcal{C}'$ for $f'(\mathcal{C})$. Write F' for the growth of f'.

Claim 3. The translation $f' : \mathcal{B} \to \mathcal{C}'$ is length-monotonic and has δ-delay.

Use the fact that $|f'(x)| = F(|x|) = F'(|x|)$.

Claim 4. There exists $p, s \in \mathbb{N}$ such that the sequence $F'(n + p) - F'(n) = s$ for almost all n.

Let $l_0 < l_1 < \cdots$ be the sequence of integers l for which there is a string in $\mathcal{B}$ of length l. Let $u_n \in \mathcal{C}'$ denote the length-lexicographically smallest element amongst $\{f'(x) \mid |x| = l_n, x \in \mathcal{B}\}$. The set L of all such u_n is regular. Note that $|u_n| = F'(l_n)$ and so $|u_i| \leqslant |u_{i+1}|$ (length-monotonic) and $|u_i| < |u_{i+\delta}|$ (δ-delay). Thus L has at most δ many strings of any given length. So partition L into regular sets L_k for $k \leqslant \delta$: $x \in L_k$ if there are exactly k strings of length $|x|$ in L. The length-preserving projection of these onto 0^* results in unary presentations of L_k. These are ultimately periodic.

For simplicity assume the previous claim holds for all n. Now for $x \in \mathcal{B}$ of length n write $f'(x)$ as $v_1 v_2 \cdots v_n$ where $|v_i| = s$ (if $|f'(x)|$ is not a multiple of s, append a (new) blank symbol until it is). For a word w of length s write $\widehat{w}$ for a new alphabet symbol.

Define $f'' : x \mapsto \widehat{v_1} \cdots \widehat{v_n}$. Write $\mathcal{C}''$ for $f''(\mathcal{B})$. Clearly the translation $f'' : \mathcal{B} \to \mathcal{C}''$ is length-preserving.

Claim 5. Since f'' is length-preserving and preserves all regular relations, the graph of f'' is regular.

The idea is that we can use lengths of elements of $\mathcal{B}$ (and $\mathcal{C}''$) as pointers to simultaneously identify the symbols in x and $f''(x)$. For simplicity, suppose that for every $n \in \mathbb{N}$ there is an element in $\mathcal{B}$ of length n (in general the gap between lengths is bounded). For a symbol σ define the regular relation $S_\sigma(p, b) \subset \mathcal{B} \times \mathcal{B}$ saying that σ occurs in b at position $|p|$. Write $R_\sigma \subset \mathcal{C}'' \times \mathcal{C}''$ for the image of S_σ under f''. It is also regular. Then $f''(x) = y$ if and only if $|x| = |y|$ and for all $p \in \mathcal{B}$ and $q \in \mathcal{C}''$ with $|p| = |q|$ and each symbol σ we have $S_\sigma(p, x) \iff R_\sigma(q, y)$. This latter condition is regular.

Finally, write $\pi : \mathcal{C} \to \mathcal{C}'$ for the map sending $f(x) \mapsto f'(x)$, and $\beta : \mathcal{C}' \to \mathcal{C}''$ for the map sending $f'(x) \mapsto f''(x)$. Since π^{-1} is a projection its graph is regular. Finally, β^{-1} is semi-synchronous sending blocks of size 1 to blocks of size s. □

6 Relatives of automatic structures

6.1 Expansions by predicates and automatic with advice

Elgot and Rabin [25] use automata theoretic arguments to show that certain expansion of $\mathfrak{T}_1$ by unary predicates have decidable MSO. For instance they showed that $(\mathbb{N}, +1, \mathrm{Fact})$ with $\mathrm{Fact} := \{n! \mid n \in \mathbb{N}\}$ has decidable MSO-theory. For a predicate $P \subseteq \mathbb{N}$, the *P-membership problem* is to decide, given a Büchi-automaton M, whether or not M accepts χ_P. Recall that we write $(\mathfrak{A}, P)$ for the structure $\mathfrak{A}$ expanded by the predicate P.

Lemma 6.1. *[25] For every predicate $P \subseteq \mathbb{N}$, the structure $(\mathfrak{T}_1, P)$ has decidable MSO-theory if and only if the P-membership problem is decidable.*

Proof. Let Φ be a sentence of $(\mathfrak{T}_1, P)$. Let X be a variable not used in Φ. Build a formula $\Psi(X)$ from Φ in which every occurence of P has been replaced by the variable X. By construction $(\mathfrak{T}_1, P) \models \Phi$ if and only if $\mathfrak{T}_1 \models \Psi(P)$. The latter condition is equivalent to the problem of whether the automaton corresponding to Ψ accepts P or not. □

We now briefly discuss how to find explicit predicates whose P-membership problem is decidable.

The contraction method for $P \subseteq \mathbb{N}$. Almost trivially, ultimately periodic P have decidable P-membership problem. For more general predicates, like the factorials Fact, we use the *contraction method* of [25] and its generalisation by Carton and Thomas [17] that we now explain.

Call P *residually ultimately-constant* if there is an infinite sequence $x_0 < x_1 < \cdots$ of numbers such that for every semigroup morphism $h : \{0, 1\}^* \to S$ with S finite, the sequence $(h(u_i))_{i \geqslant 0}$ is ultimately constant, where $u_i = \chi_P[x_i, x_{i+1})$. Call P *effectively residually ultimately-constant* if the function $i \mapsto x_i$ is computable and given $h(0), h(1)$ and S one can compute an integer l, *a lag*, such that for all $m \geqslant l$, $h(u_l) = h(u_m)$. For example, it can be shown that the set of factorials Fact is effectively residually ultimately-constant.

Proposition 6.2. *[17] If P is effectively residually ultimately-constant then the P-membership problem is decidable.*

Proof. There is a standard effective way to associate with a given automaton M a morphism and finite semigroup $h : \{0,1\}^* \to (S_M, \star)$ with the following property[7]: if a word $\alpha_0\alpha_1\alpha_2 \cdots$ $(\alpha_i \in \{0,1\}^*)$ is accepted by M then every word β that can be decomposed into $\beta_0\beta_1\beta_2 \cdots$ $(\beta_i \in \{0,1\}^*)$ with $h(\alpha_i) = h(\beta_i)$ (for all i) is also accepted by M. Then χ_P is accepted by M if and only if the ultimately periodic string uv^ω is accepted by M, where $u = \chi_P[0, x_l)$ and $v = \chi_P[x_l, x_{l+1})$. The latter property is decidable since we can compute a lag l and the elements $h(u), h(v)$. □

It turns out that P being effectively residually ultimately-constant is also a necessary condition for $(\mathfrak{T}_1, P)$ having decidable MSO. See [54, 53] for a proof of this and other characterisations of $(\mathfrak{T}_1, P)$ having decidable MSO. See [17, 2] for explicit effectively residually ultimately-constant predicates, including the morphic predicates.

Why restrict to expansions by unary predicates only? The reason is that expansions by non-trivial binary relations result in undecidability.

Theorem 6.3. *[25] Let $g : \mathbb{N} \to \mathbb{N}$ be a function such that $x < y$ implies $1+g(x) < g(y)$. The expansion of $\mathfrak{T}_1$ by the relation $G_g := \{(n, g(n)) \mid n \in \mathbb{N}\}$ has undecidable* WMSO-*theory.*

In fact, one shows that one can quantify over finite relations which in turn gives the power of recursion to define addition and then multiplication. An example is taking $g(n) := 2n$.

What about expansions of $\mathfrak{T}_2$? For predicate $P \subseteq \{0,1\}^*$, define the P-membership problem as above but with Rabin automata instead. Then identical arguments show that $(\mathfrak{T}_2, P)$ has decidable MSO if and only if the P-membership problem is decidable. The pushdown/Caucal hierarchy is a well studied collection of trees (and graphs) with decidable MSO [18] [63]. We illustrate an approach for decidability due to Fratani [27] (that can yield decidability of all trees in the hierarchy). To a semigroup $(\mathcal{M}, +)$ with finitely many generators $g_1, \cdots, g_k$, associate the Cayley structure $(\mathcal{M}, S_1, \cdots, S_k)$ where $S_i(m) = m + g_i$. For instance, the semigroup of words under concatenation $(\{0,1\}^*, \cdot)$ viewed as a structure is $\mathfrak{T}_2$.

Theorem 6.4. *[27] Take $\mu : \{0,1\}^* \to \mathcal{M}$ a surjective semigroup morphism and a set $R \subset \mathcal{M}$. If the MSO-theory of $(\mathcal{M}, S_1, \cdots, S_k, R)$ is decidable then the MSO-theory of $(\mathfrak{T}_2, \mu^{-1}(R))$ is decidable.*

[7] The idea appeared in Büchi's complementation proof: define $h : \{0,1\}^* \to 2^{Q \times Q \times \{=, \neq\}}$ so that $(q, q', \oplus) \in h(u)$ if and only if there is a path in M from state q to state q' labelled u such that if C are the states occurring on this path then $C \cap F \oplus \emptyset$. Note that if $h(u_i) = h(v_i)$ for $i = 1, 2$ then $h(u_1u_2) = h(v_1v_2)$. Thus define the associative operation $\star$ on S by $s_1 \star s_2 = h(u_1u_2)$ where u_i is any element such that $h(u_i) = s_i$.

Proof. Use the relationship between Rabin automata and parity games (see [62]) to show that a given tree automaton (with state set Q, initial state ι, transition set Δ) accepts $\mu^{-1}(R)$ if and only if the first player has a memoryless winning strategy in the parity game defined as follows: the arena is $\mathcal{M} \times (Q \cup \Delta)$; the priority of (m, q) and $(m, (q, \sigma, q_0, q_1))$ is the priority of q, the starting node is $(\mu(\lambda), \iota)$, and for every transition $\delta = (q, \sigma, q_0, q_1) \in \Delta$ and $m \in \mathcal{M}$ such that $m \in R \iff \sigma = 1$ the first player's moves are of the form (m, q) to (m, δ) and the second player's moves are of the form (m, δ) to $(m + \mu(i), q_i)$ for $i \in \{0, 1\}$. Having a memoryless winning strategy is expressible in MSO over $(\mathcal{M}, S_1, \cdots, S_k, R)$. □

Example 6.1. Consider the semigroup morphism $\mu : \{0,1\}^* \to \mathbb{N}$ that sends u to the number of 1s in u (the operation on $\mathbb{N}$ is addition). We have seen that $(\mathfrak{T}_1, \mathrm{Fact})$ has decidable MSO and so conclude that $(\mathfrak{T}_2, \mu^{-1}(\mathrm{Fact}))$ does too.

Automatic with advice If $(\mathfrak{T}_2, P)$ has decidable MSO-theory then every structure FO-interpretable in $\mathcal{P}[(\mathfrak{T}_2, P)]$ has decidable FO-theory. This justifies the following definition.

Definition 6.1. [19] A structure is *Rabin-automatic with advice* $P \subseteq \{0,1\}^*$ if it is FO-interpretable in $\mathcal{P}[(\mathfrak{T}_2, P)]$. A structure is *Büchi-automatic with advice* $P \subseteq \mathbb{N}$ if it is FO-interpretable in $\mathcal{P}[(\mathfrak{T}_1, P)]$.

A machine theoretic characterisation holds. A *Rabin-automaton with advice* $P \subseteq \{0,1\}^*$ is one that, while in position $u \in \{0,1\}^*$, can decide on its next state using the additional information of whether or not $u \in P$.[8] In other words, the advice P is simply read as part of the input. Thus a structure has a presentation by Rabin-automata with advice P if and only if it is FO-interpretable in $\mathcal{P}[(\mathfrak{T}_2, P)]$. An analogous statement holds for Büchi-automata with advice $P \subseteq \mathbb{N}$.

The theory of Rabin-automatic structures with advice is yet to be developed. These generalise Rabin-automatic structures which themselves still hold some mystery, e.g. are the countable quotients already finite-tree automatic? It is known that the extension of the fundamental theorem 3.6 holds: if $\mathfrak{A}$ is Rabin-automatic with advice P then the code of every FO($\exists^{\geqslant \aleph_0}, \exists^{\geqslant \aleph_1}, \exists^{\mathrm{mod}}$)-definable relation is recognised by a Rabin-automaton with advice P, see [4].

Structures in which elements are coded by *finite* strings/trees have received some consideration.

Definition 6.2. [19] A structure is *finite-tree automatic with advice* $P \subseteq \{0,1\}^*$ if it is FO-interpretable in $\mathcal{P}_f[(\mathfrak{T}_2, P)]$. A structure is *finite-string automatic with advice* $P \subseteq \mathbb{N}$ if it is FO-interpretable in $\mathcal{P}_f[(\mathfrak{T}_1, P)]$.

Again, if $(\mathfrak{T}_i, P)$ has decidable WMSO then $\mathcal{P}_f[(\mathfrak{T}_i, P)]$ has decidable FO-theory. Note that a machine model for, say finite-tree automatic with advice P, would have to

[8] The word 'advice' is meant to connote that we can ask for a bit of information based on the current state and the current symbol being read. The other term found in the literature is 'oracle' which I choose not to use because in computability theory it means that the machine can ask if the whole content written on a tape is in the oracle language.

have an infinitary acceptance condtions (such as the Rabin acceptance condition) since the automaton has to process P which is typically infinite [19].

Example 6.2. The structure $(\mathbb{Q}, +)$, although not finite-string automatic [65], is finite-string automatic with advice.[9] To simplify exposition we give a presentation $([0, 1) \cap \mathbb{Q}, +)$ by finite strings over the alphabet $\{0, 1, \#\}$ where the automata have access to the advice string

$$10\#11\#100\#101\#110\#111\#1000\# \cdots$$

which is a version of the Champernowne-Smarandache string and known to have decidable MSO [2]. To every rational in $[0, 1)$ there is a unique *finite* sequence of integers $a_1 \cdots a_n$ such that $0 \leqslant a_i < i$ and $\sum_{i=2}^{n} \frac{a_i}{i!}$ and n minimal. The presentation codes this rational as $f(a_2)\#f(a_3)\#f(a_4) \cdots \#f(a_n)$ where f sends a_i to the binary string of length $\lceil \log_2 i \rceil + 1$ representing a_i. Addition $a + b$ is performed least significant digit first (right to left) based on the fact that

$$\frac{a_i + b_i + c}{i!} = \frac{1}{(i-1)!} + \frac{a_i + b_i + c - i}{i!}$$

where $c \in \{0, 1\}$ is the carry in. In other words, if $a_i + b_i + c \geqslant i$ then write $a_i + b_i + c - i$ in the ith segment and carry a 1 into the $(i-1)$st segment; and if $a_i + b_i + c < i$ then write this under the ith segment and carry a 0 into the $(i-1)$st segment. These comparisons and additions can be performed since the advice tape is storing i in the same segment as a_i and b_i. Of course since the automaton reads the input and advice from left to right it should non-deterministically guess the carry bits and verify the addition.

Structures that are finite-tree automatic with advice were first studied in [19].

Theorem 6.5. *[19] If $(\mathfrak{A}, \equiv)$ is finite-tree automatic with advice P and $\equiv$ is a congruence on $\mathfrak{A}$ then $\mathfrak{A}/_{\equiv}$ is also finite-tree automatic with advice P.*

Theorem 6.6. *[19] If $\mathcal{P}_f[\mathfrak{S}]$ is finite-tree automatic with advice P then $\mathfrak{S}$ is WMSO-interpretable in $(\mathfrak{T}_2, P)$.*

Consequently it can be shown that the following structures are not finite-tree automatic with any advice: the free monoid on two or more generators; the random graph; the structure $\mathcal{P}_f[(\mathbb{N}, +)]$. For instance, the last item follows from the fact that $(\mathbb{N}, +)$ is not WMSO-interpretable in any tree. It is not known what happens in these examples and theorems if we replace WMSO by WMSO and $\mathcal{P}_f$ by $\mathcal{P}$.

6.2 Descriptive set theory and Borel presentations

A standard reference for classical descriptive set theory is [33]. See [48] for a short survey about Borel presentable structures. There is a natural topology, called the Cantor topology, on $\{0, 1\}^\omega$, namely the one whose basic open sets are of the form $\{\alpha \mid \tau \preceq_{\mathtt{prefix}} \alpha\}$ for

[9] This was communicated independently by Frank Stephan and Joe Miller and reported in [49].

$\tau \in \{0,1\}^*$. A subset $X \subseteq \{0,1\}^\omega$ is called *Borel (over* $\{0,1\}$*)* if it is in the smallest class of subsets of $\{0,1\}^\omega$ containing the basic open sets and closed under countable unions and complementation.

Example 6.3. (1) A set X is a countable union of basic open sets if and only if there exists $W \subseteq \{0,1\}^*$ such that

$$\alpha \in X \iff (\exists i)\alpha[i] \in W.$$

These are the *open* sets.

(2) Complements of open sets are called *closed*. Every singeleton is closed and thus every countable subset of $\{0,1\}^\omega$ is Borel.

(3) A set X is a countable intersection of open sets if and only if there exists $W \subseteq \{0,1\}^*$ such that

$$\alpha \in X \iff (\forall j)(\exists i > j)\alpha[i] \in W$$

(4) A set X is a countable union of closed sets if and only if there exists $W \subseteq \{0,1\}^*$ such that

$$\alpha \in X \iff (\exists j)(\forall i > j)\alpha[i] \in W$$

(5) A language X recognised by a *deterministic* Muller automaton is a boolean combination of sets of the previous two forms. Thus every ω-regular language is Borel.

In a similar way we can form Borel subsets of A^ω where A is a finite set, not just $\{0,1\}$. Thus we can define Borel relations: we call $S \subseteq (\{0,1\}^\omega)^r$ Borel if $\otimes S$ is a Borel subset of $(\{0,1\}^r)^\omega$.

Lemma 6.7. *Borel relations are closed under Boolean combinations and instantiation, i.e. if R is Borel over X and $x \in X^\omega$ is fixed, then*

$$\{(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_r) \mid (x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_r) \in R\}$$

is Borel for every i.

Definition 6.3 (Borel presentation). A *Borel structure* $(\mathcal{B}, S_1, \cdots, S_N)$ is one for which $\mathcal{B} \subseteq \{0,1\}^\omega$ is Borel and each of the relations $S_1, \cdots, S_N$ are Borel. Any structure isomorphic to it is called *Borel presentable*.

Example 6.4. The following structures are Borel presentable.

(1) Every structure with a countable domain.
(2) The field $(\mathbb{C}, +, \times)$.
(3) The order $(\mathcal{P}(\{0,1\}^\star), \subseteq)$.
(4) The power set structure of $(\mathbb{N}, +, \times)$.
(5) Every Büchi-automatic structure.

Although Borel presentable structures do not neccessarily have decidable first-order theories, methods of descriptive set theory can be used to answer questions about automatic presentations. To illustrate we will make use of expansions and extensions of the Borel presentable structure $(\mathcal{P}(\{0,1\}^\star), \subseteq)$.

Lemma 6.8 ([29]). *Let C be a countable set. If Φ is an isomorphism between two Borel presentations of $(\mathcal{P}(C), \subseteq)$, say (D, L) and (D', L'), then the graph of Φ is Borel.*

Proof. Every Borel presentation of $(\mathcal{P}(C), \subseteq)$ can be expanded to include the countable unary predicate Sing $:= \{\{u\} \mid u \in C\}$. So consider expansions (D, L, S) and (D', L', S'). Note that $\Phi(X) = Y$ if and only if for every U in the countable set S, $(U, X) \in L \iff (\phi(U), Y) \in L'$. Thus the graph of Φ is a countable intersection of Borel relations (by Lemma 6.7), and so is itself Borel. □

We will use the following facts.

Fact 1. The graph of a function is Borel if and only if the preimage of every Borel set is Borel.

Fact 2. Suppose a function $F : \{0, 1\}^\omega \to \{0, 1\}^\omega$ satisifies that for all X, Y the strings X and Y are eventually equal if and only if $F(X) = F(Y)$. Then the graph of F is not Borel.

The following theorem separates ωT-AutStr from ωS-AutStr.

Theorem 6.9 ([29]). *There is a structure in ωT-AutStr that has no Borel presentation; in particular this structure is not in ωS-AutStr.*

Proof. Consider the structure $\mathfrak{V} = (\mathcal{P}(\{0, 1\}^\star), \subseteq, V)$ where V is the unary relation consisting of those sets X such that the characteristic tree of X has the property that every infinite path is labelled with only finitely many 1s. The structure $\mathfrak{V}$ has a natural presentation placing it in ωT-AutStr. A bijection between $\{0, 1\}^\star$ and $\mathbb{N}$ allows us to identify $\mathcal{P}(\{0, 1\}^\star)$ with $\{0, 1\}^\omega$. Write $\mathfrak{V}' = (\{0, 1\}^\omega, \subseteq', V')$ for the corresponding structure. Now if $\mathfrak{V}'$ were Borel presentable, say via isomorphism Ψ, then by Lemma 6.8 the graph of Ψ would be Borel and so would V' be Borel. But this contradicts the fact that V' is Π^1_1-complete. □

The following theorem says that ωS-AutStr is not closed under quotients.

Theorem 6.10 ([29]). *There is a structure $(\mathfrak{A}, \approx)$ in ωS-AutStr whose quotient $\mathfrak{A}/_\approx$ has no Borel presentation, and is thus not in ωS-AutStr.*

Proof. Let $\mathfrak{B}_0$ and $\mathfrak{B}_1$ be structures with disjoint domain, each isomorphic to $(\mathcal{P}(\mathbb{N}), \subseteq)$. Define $\mathfrak{A}$ to be the structure with domain $B_0 \cup B_1$ and relations $\leqslant$, U, f where $\leqslant$ is $\subseteq$ restricted to each B_i, U holds on the elements of B_0 and $f : B_0 \to B_1$ is the identity. Let $=^*$ be an equivalence relation on the $B_0 \cup B_1$ which is the identity on B_0 and 'the symmetric difference is finite' on B_1. Clearly $(\mathfrak{A}, =^*) \in \omega$S-AutStr. We now sketch the proof that $\mathfrak{A}/_{=^*}$ is not Borel presentable. Assume that $\mathfrak{A}/_{=^*}$ is isomorphic to a Borel structure $(B', \leqslant', U', f')$ via Φ and let Φ_0 be the restriction of Φ to $[B_0]_{=^*}$ (which is B_0). By Lemma 6.8 Φ_0 is Borel. Then by Fact 1 the composition $F := f' \circ \Phi_0$ is Borel, contradicting Fact 2. □

A more sophisticated set-theoretical argument shows that the structure $\mathfrak{A}/_\approx$ from the previous theorem is not in ωT-AutStr [29].

Methods from (descriptive) set theory have also been used to study the isomorphism problem for automatic structures. Specifically, writing $\mathcal{M} = (M_B, M_1, \cdots, M_N)$ for a tuple of automata presenting a structure, the *isomorphism problem* for a collection $\mathbf{C}$ is the set of pairs $(\mathcal{M}, \mathcal{M}')$ such that both $\mathcal{M}$ and $\mathcal{M}'$ present the same structure from $\mathbf{C}$. See [57, 40, 41] in which various natural isomorphism problems can be placed within the arithmetic and analytic hierarchy (see [56] for definitions). For instance, the complexity of the isomorphism problem for $\texttt{S-AutStr}$ is Σ^1_1-complete.

We end with a stunning result:

Theorem 6.11 ([26]). *The isomorphism problem for the collection*

$$\{\mathfrak{A}/_{\approx} \mid (\mathfrak{A}, \approx) \in \omega\texttt{T-AutStr}\}$$

is not determined by the axiomatic system ZFC.

Proof. Here is the barest of sketches. Write FIN for the ideal of finite subsets of $\mathbb{N}$, and ANC for the ideal of subsets of $\{l, r\}^\star$ with no infinite antichain with respect to the prefix-order. Both $(\mathcal{P}(\mathbb{N}), \cap, \cup, \neg, \mathrm{FIN})$ and $(\mathcal{P}(\{l, r\}^\star), \cap, \cup, \neg, \mathrm{ANC})$ are Rabin-automatic. Write $\mathfrak{B}_1$ for the Boolean algebra $\mathcal{P}(\mathbb{N})/\mathrm{FIN}$ and $\mathfrak{B}_2$ for $\mathcal{P}(\{l, r\}^\star)/\mathrm{ANC}$. Using results from descriptive set theory it can be proved that $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are isomorphic under ZFC+CH, but not isomorphic under ZFC+OCA (open colouring axiom). □

7 Questions

This chapter focused on what I consider foundational problems of automatic structures. Here are some directions and questions raised along the way.

(1) Which generalised quantifiers can be added to FO to still get definability/decidability for all $\circ$-automatic structures as in Theorem 3.6? What if one restricts to a specific automatic structure?

(2) Suppose $(\mathfrak{A}, \epsilon)$ is Rabin-automatic. In general, the quotient $\mathfrak{A}/\epsilon$ is not Rabin-automatic. If we further suppose that $\mathfrak{A}/\epsilon$ is countable, is $\mathfrak{A}/\epsilon$ then Rabin-automatic?

(3) Can automata be used to explain why the countable random graph, which is not finite-tree automatic with any advice, has a decidable FO-theory? Is it the regular quotient of a Rabin-automatic structure with advice? The same question can be asked of other important structures such as real arithmetic $(\mathbb{R}, +, \times, <)$ which is decidable in double exponential time.

(4) Find techniques to show that certain structures are not Rabin-automatic or that certain uncountable structures are not Büchi-automatic.

(5) If the power structure $\mathcal{P}[\mathfrak{S}]$ is Rabin-automatic with advice Q then is $\mathfrak{S}$ already MSO-interpretable in $(\mathfrak{T}_2, Q)$?

References

[1] V. Bárány. Invariants of automatic presentations and semi-synchronous transductions. In *Proceedings of the 23rd Annual Symposioum on Theoretical Aspects of Computer Science, STACS 2006*, pages 289–300, 2006. 958, 959

[2] V. Bárány. *Automatic Presentations of Infinite Structures*. Phd thesis, RWTH Aachen, 2007. 962, 964

[3] V. Bárány, E. Grädel, and S. Rubin. Automata-based presentations of infinite structures. In *Finite and Algorithmic Model Theory*, London Mathematical Society Lecture Note Series (No. 379). Cambridge University Press, 2011. 943

[4] V. Bárány, L. Kaiser, and A. Rabinovitch. Cardinality quantifiers in MLO over trees. In *Computer Science Logic*, pages 117–131, 2009. 945, 963

[5] V. Bárány, L. Kaiser, and S. Rubin. Cardinality and counting quantifiers on omega-automatic structures. In *STACS '08: Proceedings of the 25th Annual Symposium on Theoretical Aspects of Computer Science*, 2008. 954, 955

[6] P. T. Bateman, C. G. J. Jr., and A. R. Woods. Decidability and undecidability of theories with a predicate for the primes. *J. Symb. Log.*, 58(2):672–687, 1993. 940

[7] M. Benedikt, L. Libkin, and F. Neven. Logical definability and query languages over ranked and unranked trees. *ACM Transactions on Computational Logic*, 8(2), 2007. 952

[8] M. Benedikt, L. Libkin, T. Schwentick, and L. Segoufin. Definable relations and first-order query languages over strings. *J. ACM*, 50(5), 2003. 943

[9] A. Blumensath. *Automatic Structures*. Diploma thesis, RWTH Aachen, 1999. 942, 948, 951, 952, 953

[10] A. Blumensath, T. Colcombet, and C. Löding. Logical theories and compatible operations. In J. Flum, E. Grädel, and T. Wilke, editors, *Logic and Automata: History and Perspectives*, Texts in Logic and Games, pages 73–106. Amsterdam University Press, 2007. 952

[11] A. Blumensath and E. Grädel. Automatic structures. In *15th Symposium on Logic in Computer Science* (LICS), pages 51–62, 2000. 942, 943, 948, 949, 951, 952, 953

[12] G. Boolos, Burgess, and Jeffrey. *Computability and Logic*. Cambridge University Press, 2007. 944

[13] E. Börger, E. Grädel, and Y. Gurevich. *The Classical Decision Problem*. Perspectives in Mathematical Logic. Springer, 1997. 943

[14] V. Bruyère, G. Hansel, C. M. Christian, and R. Villemaire. Logic and p-recognizable sets of integers. *Bulletin of the Belgian Mathematical Society. Simon Stevin.*, 1(2):191–238, 1994. Journées Montoises (Mons, 1992). 959

[15] J. Büchi. Weak second-order arithmetic and finite automata. *Z. Math. Logik Grundlagen Math.*, 6:66–92, 1960. 941

[16] J. Büchi. On a decision method in restricted second order arithmetic. *Logic, Methodology and Philosophy of Science (Proc.* 1960 *Internat. Congr)*, pages 1–11, 1962. 941

[17] O. Carton and W. Thomas. The monadic theory of morphic infinite words and generalizations. *Inf. Comput.*, 176(1):51–65, 2002. 961, 962

[18] D. Caucal. On infinite terms having a decidable monadic theory. In *Mathematical foundations of computer science 2002*, volume 2420 of *Lecture Notes in Comput. Sci.*, pages 165–176, Berlin, 2002. Springer. 962

[19] T. Colcombet and C. Löding. Transforming structures by set interpretations. *Logical Methods in Computer Science*, 3(2), 2007. 942, 947, 953, 954, 963, 964

[20] M. Dauchet. Rewriting and tree automata. In *Term Rewriting*, pages 95–113, 1993. 954

[21] C. Delhommé. Automaticité des ordinaux et des graphes homogènes. *Comptes Rendus Mathematique*, 339(1):5–10, 2004. 956, 957

[22] J. Doner. Tree acceptors and some of their applications. *Journal of Computer and System Sciences*, 4:406–451, 1970. 941

[23] H. Ebbinghaus and J. Flum. *Finite Model Theory*. Perspectives in Mathematical Logic. Springer, 1995. 943

[24] C. Elgot. Decision problems of finite automata design and related arithmetics. *Transactions of the American Mathematical Society*, 98:21–51, 1961. 941

[25] C. Elgot and M. Rabin. Decidability of extensions of theory of successor. *Journal of Symbolic Logic*, 31(2):169–181, 1966. 947, 961, 962

[26] O. Finkel and S. Todorčević. The isomorphism relation between tree-automatic structures. *Central European Journal of Mathematics*, 8:299–313, 2010. 967

[27] S. Fratani. *Automates piles de piles ... de piles*. PhD thesis, Bordeaux, 2005. 962

[28] Y. Gurevich. Monadic second-order theories. In J. Barwise and S. Feferman, editors, *Model-Theoretic Logics*, pages 479–506. Springer-Verlag, Perspective in Mathematical Logic, 1985. 943

[29] G. Hjorth, B. Khoussainov, A. Montalbán, and A. Nies. From automatic structures to Borel structures. In *23rd Symposium on Logic in Computer Science (LICS)*, 2008. 954, 955, 966

[30] W. A. Hodges. *Model Theory*. Cambridge University Press, 1993. 941, 946, 955

[31] B. Hodgson. *Théories décidables par automate fini*. Phd thesis, University of Montréal, 1976. 942

[32] B. Hodgson. Decidabilite par automate fini. *Annales de Sciences Math/'ematiques du Qu/'ebec*, 7:39–57, 1983. 942

[33] A. Kechris. *Classical Descriptive Set Theory*. Springer-Verlag, 1995. 964

[34] B. Khoussainov and M. Minnes. Model theoretic complexity of automatic structures. *Annals of Pure and Applied Logic*, To appear, 2008. 943

[35] B. Khoussainov and A. Nerode. Automatic presentations of structures. *Lecture Notes in Computer Science*, 960:367–392, 1995. 942, 948, 951, 955, 956

[36] B. Khoussainov, S. Rubin, and F. Stephan. Definability and regularity in automatic structures. In *STACS 2004*, volume 2996 of *LNCS*, pages 440–451. Springer, Berlin, 2004. 946, 949

[37] B. Khoussainov, S. Rubin, and F. Stephan. Automatic linear orders and trees. *ACM Transactions on Computational Logic (TOCL)*, 6(4):675–700, 2005. 943

[38] D. Kuske. Is Cantor's theorem automatic ? *Proceedings of the 10th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR)*, 2850:332–345, 2003. 943

[39] D. Kuske. Is Ramsey's theorem omega-automatic? *27th International Symposium on Theoretical Aspects of Computer Science (STACS 2010)*, 5:537–548, 2010. 943

[40] D. Kuske, J. Liu, and M. Lohrey. The isomorphism problem for *omega*-automatic trees. In *CSL*, pages 396–410, 2010. 967

[41] D. Kuske, J. Liu, and M. Lohrey. The isomorphism problem on classes of automatic structures. In *LICS*, pages 160–169, 2010. 967

[42] D. Kuske and M. Lohrey. First-order and counting theories of omega-automatic structures. Technical Report Fakultätsbericht Nr. 2005/07, Universität Stuttgart, Fakultät Informatik, Elektrotechnik und Informationstechnik, 2005. 946, 949, 954

[43] D. Kuske and M. Lohrey. Hamiltonicity of automatic graphs. *FIP TCS 2008*, 2008. 950

[44] D. Kuske and M. Lohrey. Automatic structures of bounded degree revisited. In *CSL*, pages 364–378, 2009. 943

[45] D. Kuske and M. Lohrey. Some natural decision problems in automatic graphs. *J. Symb. Log.*, 75(2):678–710, 2010. 943

[46] P. Lindström. First order predicate logic with generalized quantifiers. *Theoria*, 32:186–195, 1966. 950

[47] C. Michaux and A. Ozturk. Quantifier elimination following muchnik. Technical Report 10, Universit de Mons-Hainaut, 2002. 940

[48] A. Montalbán and A. Nies. Borel structures: a brief survey. Submitted. 964

[49] A. Nies. Describing groups. *Bulletin of Symbolic Logic*, 13(3):305–339, 2007. 964

[50] M. Presburger. Über die vollständigkeit eines gewissen systems der arithmetik ganzer zahlen, in welchem die addition als einzige operation hervortritt. *Comptes Rendus du I congrès de Mathmaticiens des Pays Slaves. Warszawa: 92101*, 1929. 941

[51] M. Rabin. Decidability of second-order theories and automata on infinite trees. *Transactions of the American Mathematical Society*, 141:1–35, 1969. 941, 942, 945

[52] M. O. Rabin. Decidable theories. In J. Barwise, editor, *HANDBOOK OF MATHEMATICAL LOGIC*, volume 90 of *Studies in Logic and the Foundations of Mathematics*, pages 595 – 629. Elsevier, 1977. 943

[53] A. Rabinovich. On decidability of monadic logic of order over the naturals extended by monadic predicates, 2007. 962

[54] A. M. Rabinovich and W. Thomas. Decidable theories of the ordering of natural numbers with unary predicates. In *CSL*, 2006. 962

[55] R. M. Robinson. Restricted set-theoretical definitions in arithmetic. *Proc. Amer. Math. Soc.*, 9:238–242, 1958. 940, 941

[56] H. Rogers. *Theory of Recursive Functions and Effective Computability*. McGraw-Hill, 1967. 967

[57] S. Rubin. *Automatic Structures*. Phd thesis, University of Auckland, 2004. 950, 967

[58] S. Rubin. Automata presenting structures: a survey of the finite string case. *Bull. Symbolic Logic*, 14(2):169–209, 2008. 943, 950

[59] L. Strüngmann and G. Braun. Finite automata presentable abelian groups. *Journal of Algebra and Computation*, to appear. 954

[60] A. Tarski. A decision method for elementary algebra and geometry, 1951. 940

[61] J. W. Thatcher and J. B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Mathematical Systems Theory. An International Journal on Mathematical Computing Theory*, 2:57–81, 1968. 941

[62] W. Thomas. *Automata on infinite objects*, pages 133–191. Elsevier, 1990. 941, 944, 945, 963

[63] W. Thomas. Constructing infinite graphs with a decidable MSO-theory. In *Mathematical foundations of computer science 2003*, volume 2747 of *Lecture Notes in Comput. Sci.*, pages 113–124, Berlin, 2003. Springer. 962

[64] B. Trahtenbrot. Finite automata and the logic of one-place predicates. Russian. *Siberian Mathematical Journal*, 3:103–131, 1962. English translation: American Mathematical Society Translations, Series 2, 59 (1966), 23-55. 941

[65] T. Tsankov. The additive group of the rationals does not have an automatic presentation. *Journal of Symbolic Logic*, to appear. 954, 964

Automata and finite model theory

Wouter Gelade[1],*and Thomas Schwentick*[2]

[1]Hasselt University and transnational University of Limburg
Agoralaan gebouw D, 3590 Diepenbeek, Belgium
email: wouter.gelade@uhasselt.be

[2]TU Dortmund University
Otto-Hahn-Str. 16, 44227 Dortmund, Germany
email: thomas.schwentick@udo.edu

2010 Mathematics Subject Classification: 68Q45 or what is appropriate

Key words: Finite automata, Logic, Finite model theory

Contents

1 Introduction

Model Theory studies the “interplay between the syntactic structure of an axiom system on the one hand and ... properties of its models on the other hand” [7]. The dominating

*Research Assistant of the Fund for Scientific Research - Flanders (Belgium) logic in Model Theory is first-order predicate logic. In the right kind of structures it can basically express arbitrarily complicated predicates. As an example, in arithmetic (i.e., in the structure $(\mathbb{N}, +, \times)$) it can express all computable predicates (with the consequence of Gödel's incompleteness theorems) [8]. This strength makes it the ideal language for mathematics.

Model Theory has many beautiful results that offer powerful proof tools like compactness, Löwenheim-Skolem and the like.

Finite Model Theory investigates the properties of the *finite* models of formulas or sets of formulas. Even though it is a subfield of Model Theory it has a completely different flavor. In first place, (most of) the beautiful results mentioned above fail to hold on finite structures and thus the powerful tools are not available [12]. Furthermore, on finite structures the expressive power of first-order logic is rather limited. As an example, it can only express *local properties* (see [16] for a precise account).

This is why in Finite Model Theory extensions of first-order logic play a much more important role than in general Model Theory. These extensions typically allow to manipulate second-order objects, i.e., relations, in some specific way. For instance, second-order logic allows to quantify over relations (as opposed to elements in first-order logic), fixed-point logics allow to define relations in an inductive fashion and transitive-closure logic allows to access the transitive closure of a given or defined relation. For precise definitions we refer to Section 2.

The investigation of Finite Model Theory was initiated and pursued mainly by researchers with a computer science background or with computer science applications in mind. For example, its sub-field *Descriptive Complexity Theory* tries to capture computational complexity in a resource-independent way by means of measuring the "syntactic effort" needed to express a given property by a formula. Many characterizations have been obtained, references can be found in [7, 14, 16].

However, other Computer Science applications of logic on finite structures are interested in decidable logics, that is, logics, where it can be algorithmically checked whether a given formula φ is satisfiable. Unfortunately, the problem whether a given first-order formula has a *finite* model is undecidable, this is actually (one of?) the first results of Finite Model Theory [29]. Thus, on general finite structures, first-order logic has several undesirable features, it is expressively weak and its satisfiability problem is undecidable. Extensions of first-order logic can help for the former but, naturally, do not help for the latter. Furthermore, first-order logic allows for much fewer proof tools on finite structures than on arbitrary (finite and infinite) structures.

There is one notable and extremely important exception to this general picture, the Finite Model Theory of strings and trees. Here, the situation is more favorable in several respects. First of all, on trees and strings there is even a second-order extension of first-order logic with a decidable satisfiability problem: monadic second-order (MSO) logic. Furthermore, this logic and even some important sub-logics can express useful properties on these structures and, for use in verification, on unravellings of transition systems. The key result here is Büchi's theorem that MSO logic on strings can express exactly those properties than can be tested by a finite automaton, that is, all regular properties of strings. This result is presented as Theorem **??** in this chapter, together with a full proof. generalizations of the theorem for trees and infinite strings and infinite trees occur in other chapters of this handbook.

In the first main part of the chapter, constituted by Section 3 we describe the Finite Model Theory of strings and trees in more detail, highlighting the role of automata.

Fortunately, the use of the close connection between logics and automata is by no means limited to the Finite Model Theory of strings and trees themselves. Just to the contrary, researchers have found many different ways to utilize it for the Model Theory of general finite structures. In the second part of the chapter, in Section 4, we survey results that have been obtained by this approach. Specifically, we give examples where the logic-automata connection is generalized to structures that are "tree-like", where it is "transported" by means of logical reductions, or where the presence of trees (e.g., proof trees) or strings can be exploited in other ways.

We got inspiration from [23, 28, 19].

2 Definitions

For the definitions and notations, we largely follow Thomas [28].

Logical structures Logical formulas are interpreted over logical structures. To introduce the latter, let a (relational) *vocabulary* σ be a set of relation names, together with an arity $k \in \mathbb{N}$ for every relation name. A (relational) *structure* over σ consists of a set of universe elements U and, for every relation name R in σ, a k-ary relation R over U, where k is the arity of R. In this chapter, we will be interested in structures for three specific types of objects: strings, trees, and graphs.

For a string $w = a_1 \cdots a_n$ over alphabet A, the corresponding string structure is

$$\underline{w} = (\{1, \ldots, n\}, <, +1, (P_a)_{a \in A}),$$

where, for $i, j \in \{1, \ldots, n\}$, we have $<(i, j)$ when $i < j$; $+1(i, j)$ when $i = j + 1$; and $P_a(i)$ when $a_i = a$. As usual, we write $i < j$ and $i = j + 1$ instead of $<(i, j)$ and $+1(i, j)$.

For a (binary) tree t over alphabet A, the corresponding tree structure is

$$\underline{t} = (\text{Nodes}(t), <, S0, S1, (P_a)_{a \in A}),$$

where, for nodes $u, v \in \text{Nodes}(t)$, we have $<(u, v)$ if u is a descendant of v; $S0(u, v)$ when v is the left child of u; $S1(u, v)$ when v is the right child of u; and $P_a(u)$ when the label of u is a. As before, we write $u < v$, $S0(u) = v$ and $S1(u) = v$.

For a graph G with vertex set V and nodes labeled by elements of the alphabet A, the corresponding graph structure is

$$\underline{G} = (V, E, (P_a)_{a \in A}),$$

where, for vertices $u, v \in V$, we have $E(u, v)$ if there is an edge from u to v; and $P_a(u)$ if the label of u is a.

First-order (FO) logic A formula of first-order logic over vocabulary σ makes use of variables $x, y, \ldots$ and is built up from basic formulas of the form $x = y$ and $R(x_1, \ldots, x_k)$, where R is a relation symbol in σ of arity k; which can then be combined using negation ($\neg$), disjunction ($\vee$), and existential quantification ($\exists$). The operators of conjunction ($\wedge$), implication ($\rightarrow$), bi-directional implication ($\leftrightarrow$), and universal quantification ($\forall$) can be written in terms of the other operators and are used as abbreviations. Variables which are not bound by a quantifier are *free*. A formula which does not contain any free variables is a *sentence*.

Let $\mathfrak{A}$ be a structure, ϕ a first-order formula with free variables $x_1, \ldots, x_n$ over the same vocabulary as $\mathfrak{A}$, and $p_1, \ldots, p_n$ universe elements. Then, $(\mathfrak{A}, p_1, \ldots, p_n) \models \phi(x_1, \ldots, x_n)$ if ϕ is satisfied when interpreting each x_i by p_i, the equalities $x = y$ by equality of the universe elements, and the relation symbols $R(x_1, \ldots, x_k)$ by their corresponding relation in $\mathfrak{A}$. For instance, for the word $w = aba$ and formula $\phi = P_a(x) \wedge \exists y\, y > x$, we have $(\underline{w}, 1) \models \phi$, but $(\underline{w}, 2) \not\models \phi$ and $(\underline{w}, 3) \not\models \phi$.

When considering string structures, we are often interested in the fragment of FO which only makes use of the linear ordering ($<$) but not the successor relation ($+1$), which we denote by FO($<$); and, conversely, those formulas which use the successor relation but not the linear ordering, denoted by FO(+1). Moreover, we use the abbreviations first(x) and last(x) for the formulas $\forall y\, (y = x \vee y > x)$ and $\forall y\, (y < x \vee y = x)$, respectively, which state that x is the first or last position in the string. Similarly, for tree structures, FO($<$) denotes those formulas which only use the descendant relation, whereas formulas in FO(+1) are only allowed to use the two successor relations.

Monadic second-order (MSO) logic A formula of monadic-second order logic is a FO formula which can additionally make use of set variables $X, Y, \ldots$ Thereto, it allows additional basic formulas of the form $X(x)$, stating that element x is in X, and existential quantification ($\exists X$) of set variables. Universal quantification ($\forall X$) can be written in terms of the other operators and is thus used as an abbreviation.

For instance, the MSO formula

$$\exists X \forall x[\text{first}(x) \rightarrow X(x) \wedge \text{last}(x) \rightarrow \neg X(x) \wedge \forall y\, y = x + 1 \rightarrow (X(x) \leftrightarrow \neg X(y))]\,,$$

when evaluated over a string structure, asserts that there exists a set of positions X which contains all odd positions, but does not contain the last one. In other words, it asserts that the string has an even length.

3 Finite model theory of strings and trees

In the first part of this chapter we investigate finite model theory on strings and trees and its relation to automata. We start with strings and, in particular, the Büchi-Elgot-Trakhtenbrot Theorem, which states the equivalence of equivalence of monadic second-order logic and finite automata; and highlight the consequences of this theorem. Then, we look at several simpler logics such as first-order logic, which is expressively equivalent to the star-free languages [17], and several fragments of first order logic such as FO^2 and the Σ_n-hierarchy.

3.1 Strings

In this section, we study the finite model theory of strings. We first consider several extensions of first-order logic: mainly monadic second-order (MSO) logic, but also unary Transitive Closure (TC) logic and the μ-calculus. These logics have in common that they are all unary: MSO is monadic, TC is unary, and the μ-calculus computes fixed points for sets of elements, not sets of tuples of elements. This property makes them stay within, and in fact coincide with, the regular languages. This is immediately the reason for their, and in particular MSO's, prominent position when considering the connection between automata and Finite Model Theory.

Nonetheless, first-order logic also remains an important logic which ...

Monadic second-order logic (MSO) We start with the Büchi-Elgot-Trakhtenbrot Theorem which states that a string language is definable in MSO if and only if it is regular. This is proved by showing that one can translate any finite automaton to an MSO formula, and vice versa. In the former translation, an MSO formula can *simulate* an automaton by guessing, using existential set quantification, a run of the automaton; and verifying that the guessed run is valid. To see that the converse translation can also be done, consider the following simple example: $\phi = \neg\exists x \exists y\, y = x + 1 \wedge P_a(x) \wedge P_b(x)$, which states that a string does not ab as a substring. For each of the three basic expressions, $y = x + 1$, $P_a(x)$, and $P_b(x)$, we construct an automaton which accept words over A together with an assignment for x and y consistent with the expression. The assignment is indicated by replacing a normal symbol a by the symbol (a, x) or (a, y). For instance, the automaton for $y = x + 1$ accepts words for which y and x are assigned subsequent positions, such as $a(b, x)(c, y)ab$; and the automaton for $P_a(x)$ makes sure that x is assigned to an a-labeled position, such as in $(a, x)b(c, y)ab$. Then, the intersection of these three automata, obtained through a product construction, is an automaton for $y = x + 1 \wedge P_a(x) \wedge P_b(x)$ and only accepts strings in which subsequent a and b positions are labeled x and y. Then, to deal with the quantification $\exists x \exists y$, we need to replace the symbols (a, x) and (b, y) by a and b in strings accepted by the current automaton, which can be done as the regular languages are closed under projection. Finally, to take the complement of this language, it suffices to complement the automaton, using a subset construction. This approach of inductively translating an MSO formula to an automaton by using the closure properties of the regular languages, can be applied more generally and allows to prove the following theorem.

Theorem 3.1 (Büchi [4], Elgot [9], Trakhtenbrot [30]). *A string language is definable in MSO if and only if it is regular.*

Proof. To show that any regular langauge is definable in MSO, let $\mathcal{A}$ be an automaton and w a word. It is accepted by $\mathcal{A}$ if there exists a run of $\mathcal{A}$ on w, that is, a sequence of states $q_0, \ldots, q_n$ which starts in an initial state, can be followed by $\mathcal{A}$ when reading w, and ends in a final state. Put differently, $\mathcal{A}$ labels every position of w with a colour (its state) and checks whether local conditions about subsequent positions of this colouring (are transitions applied correctly?) are satisfied. This can easily be done using MSO. Indeed, a colouring can be guessed using set quantification, and the local conditions of this colouring can be verified using first-order logic. More formally, let $\mathcal{A} = (Q, I, E, T)$, with $Q = \{q_1, \dots, q_k\}$, be an automaton. The corresponding formula ϕ will guess a run of $\mathcal{A}$ in the form of k sets $X_1, \dots, X_k$. Here, a position is in X_i if the run is in state q_i after reading this position. Then, ϕ verifies whether the run is valid. Note that, as in a word of length n a run consists of $n+1$ states, we do not guess the initial state of the run. However, we verify in ϕ whether a proper initial state p for this run exists.

$$\begin{aligned}\phi = \exists X_1 \cdots \exists X_k (\forall x \bigvee_{i \in [1,k]} X_i(x)) \wedge (\forall x \bigwedge_{1 \leqslant i < j \leqslant k} \neg X_i(x) \vee \neg X_j(x)) \wedge \\ \forall x (\text{first}(x) \to \bigvee_{p \in I, (p,a,q_i) \in E} R_a(x) \wedge X_i(x)) \wedge \\ \forall x (\text{last}(x) \to \bigvee_{q_i \in T} X_i(x)) \wedge \\ \forall x \forall y (y = x + 1 \to \bigvee_{(q_i,a,q_j) \in E} X_i(x) \wedge P_a(y) \wedge X_j(y)) \,.\end{aligned}$$

The same principle can be applied to more complicated structures such as trees and infinite strings and trees. The automata models for these structures all label positions with states (or colours), which can be guessed and verified using MSO. Hence, for all these classes, if a language is regular it is definable in MSO.

We next show that any MSO-definable language is regular. Already on an intuitive level there is a strong correspondence between MSO formulas and automata. MSO has a complementation, disjunction, and existential (set) quantification operator. The former two find their counterparts in automata as the regular languages are closed under complementation and union. The third, existential quantification, represents in some sense the capability of MSO to guess, which can be simulated by the non-determinism of automata.

We make this connection more formal by constructing, for an MSO formula ϕ, an equivalent automaton $\mathcal{A}_\phi$ by induction on the structure of ϕ. Here, one needs to be careful in stating what an *equivalent* automaton is as subformulas of ϕ may contain free variables, which need to be taken into account. In particular, $\mathcal{A}_\phi$ needs to consider strings together with an assignment for the free variables of ϕ and accept these strings for which ϕ is satisfied under the given variable assignment.

Hence, for a subformula ϕ with free set variables $X_1, \cdots, X_k$ and free first-order variables $x_1, \dots, x_\ell$, the automaton $\mathcal{A}_\phi$ accepts strings over $A \times \mathcal{P}(\{X_1, \dots, X_k, x_1, \dots, x_\ell\})$, where $\mathcal{P}$ is the power set operator. Hence, next to an A-symbol, a position lists those set variables in which it is contained and those first-order variables to which it is assigned. Therefore, every first-order variable must occur at exactly one position in the string. Let us denote all strings over $A \times \mathcal{P}(\{X_1, \dots, X_k, x_1, \dots, x_\ell\})$ which satisfy this property by $L_{k,\ell}$. Note that $\mathcal{A}_\phi$ only accepts strings in $L_{k,\ell}$.

Then, $\mathcal{A}_\phi$ accepts a string in $L_{k,\ell}$ if the string over A, together with the assignments for the variables, satisfies ϕ. Note that when ϕ has no free variables, $\mathcal{A}_\phi$ accepts normal strings in A^* and, in particular, those strings which are defined by ϕ.

As said, we construct $\mathcal{A}_\phi$ by induction on the structure of ϕ. The base case is immediate, but also the inductive cases are fairly straightforward. Indeed, for $\neg\phi$ the desired automaton accepts $L_{k,\ell} \setminus L(\mathcal{A}_\phi)$ which is regular as $L_{k,\ell}$ is regular and regular languages are closed under complementation. Similarly, for the disjunction of two expressions, we can take the union of their corresponding automata, while making sure that they have the same type. Finally, existential quantification of a variable X_i or x_i amounts to projecting away X_i or x_i in strings accepted by $\mathcal{A}$; which can be done as regular languages are closed under projection.

Moreover, the above approach can be used to extend this result to trees and infinite strings and trees, although these proofs are more difficult as proving the necessary closure properties becomes more complicated. Proofs of this theorem for infinite strings [5], finite trees [6, 25], and infinite trees [21] are given in Chapters **??**, **??**, and **??**, respectively. □

This theorem, and in particular the ability to translate any MSO formula over strings to an automaton, has many consequences. First, it makes MSO "decidable" as many problems can be reduced to corresponding (decidable) problems on automata, such as testing emptiness.

Corollary 3.2. *It is decidable whether an MSO formula over strings is satisfiable.*

Second, as the expressive power of the regular languages is very well understood, it makes the expressive power of MSO over strings equally clear through the following obvious, but important, corollary.

Corollary 3.3. *Every non-regular string language is not definable in MSO.*

Third, after compiling an MSO formula into an automaton, model checking of the formula can be done very efficiently.

Corollary 3.4. *Given an MSO sentence ϕ and a (string) structure $\underline{w}$, deciding whether $\underline{w} \models \phi$ can be done in time $\mathcal{O}(|\underline{w}| \cdot f(|\phi|))$, where f is a computable function.*

Unfortunately, the function f in the above statement depends on the size of the automaton obtained after the translation, which can, in the worst case, be very big. Thereto, a function in n is *elementary* if it is in $\mathcal{O}(2^{\cdot^{\cdot^{\cdot^{2^n}}}})$, for some k, where $2^{\cdot^{\cdot^{\cdot^{2}}}}$ stands for a tower of k exponents.

Theorem 3.5 (Stockmeyer, Meyer [22]).

- *For any n, there exists an MSO formula of size $\mathcal{O}(n)$ such that any equivalent automaton is not of size elementary in n.*
- *Deciding whether an MSO formula is satisfiable has a non-elementary worst case complexity.*

Finally, we note that the Büchi-Elgot-Trakhtenbrot Theorem and the translation from automata to logic in the proof, give us a normal form for MSO on strings.

Corollary 3.6. *Any MSO formula on strings is equivalent to an MSO formula of the form $\exists X_1 \cdots \exists X_k \phi$, where ϕ is in FO^2.*

We next consider unary Transitive Closure (TC) logic. This is an extension of first-order logic with a transitive closure operator $TC_{xy}\phi$, which takes a formula ϕ with free variables x, y, and possibly additional ones $\overline{z}$, and is defined as follows:

$$(\mathfrak{A}, u_1, u_2, \overline{v}) \models TC_{xy}\phi(x, y, \overline{z}) \text{ if}$$
$$(\mathfrak{A}, u_1, u_2, \overline{v}) \models \forall X[(X(x') \wedge \forall x \forall y(X(x) \wedge \phi) \rightarrow X(y)) \rightarrow X(y')](x', y', \overline{z})\,.$$

This fragment is called unary TC logic because the transitive closure is computed over single values x and y instead of tuples of values. From the definition, it is clear that TC is a fragment of MSO. On strings, however, the converse is also true as TC can simulate, for instance, regular expressions. Here, the TC operator is used to simulate the star operator of the regular expression.

Theorem 3.7 (see [7]). *A string language is definable in MSO if and only if it is definable in TC.*

First-order logic (FO) This section started with the logic MSO because of the Büchi-Elgot-Trakhtenbrot Theorem, and the fundamental connection it establishes between finite model theory and automata. First-order logic, however, is also an important logic in the setting of string languages. As a restriction of MSO it falls within the regular languages, yet it can still express many interesting properties. For instance, the formula $\forall x(P_a(x) \rightarrow \exists y\, y > x \wedge P_b(y))$ expresses that every a is eventually followed by a b; and $\exists x \forall y(P_a(x) \wedge y < x \rightarrow P_a(y) \wedge y > x \rightarrow P_b(y))$ expresses that the string consists of a sequence of a's followed by a sequence of b's. Nesting such formulas allows to express increasingly complex properties.

However, two important properties of first-order logic restrict its expressive power: (1) it is unable to count and (2) it is local. To illustrate the first property, consider the language defined by the expression $(aa)^*$, which counts whether the number of a's in the string is 0 (modulo 2). By Theorem 3.1 this language is expressible in MSO but it is not in FO, as it can not express such counting properties.

Hence, FO(<) can not express the modulo counting which is introduced in regular expression through the star operator. Leaving out this star operator, however, allows to characterize the expressive power of FO(<) exactly. Recall that the star-free languages are those languages definable by regular expression using concatenation, union, and complement (but not star).

Theorem 3.8 (McNaughton, Papert [17]). *A string language is definable in FO(<) if and only if it is star-free.*

The second restriction of first-order logic is its locality which, roughly speaking, says that a first-order formula can only relate elements which are "close to each other". To formalize this, consider the *Gaifman graph* of a structure $\mathfrak{A}$ whose nodes are the elements of A and which has an edge between elements u and v if they occur together in a relation, that is, if there exists a relation R and tuple $\overline{x}$ containing u and v such that $R(\overline{x})$ is true in $\mathfrak{A}$. Then, the *distance* from u to v is the distance from u to v in the Gaifman graph. For instance, in a string structure with the linear order $<$, the distance between any two different positions is 1; whereas in a string structure with only the successor relation, the distance between two positions is their distance in the string. Given $\ell \in \mathbb{N}$, we can write an FO formula $d_{\leqslant \ell}(x, y)$ expressing that x and y are at distance at most ℓ.

Next, we consider formulas which can only look at positions up to a distance ℓ from a position x. Let $\phi(x)$ be a first-order formula with one free variable x. Then, $\phi_{\leqslant \ell}(x)$ is obtained from $\phi(x)$ by replacing every subformula of the form $\exists y\, \psi$ by $\exists y(d_{\leqslant \ell}(x, y) \wedge \psi)$. That is, $\phi_{\leqslant \ell}(x)$ quantifies only over positions at distance at most ℓ from x, and thus only inspects the local neighbourhood of x. A sentence is *basic local* if, for some ℓ, it has the form

$$\exists x_1 \cdots \exists x_n \bigwedge_{1 \leqslant i < j \leqslant n} \neg d_{\leqslant 2\ell}(x_i, x_j) \wedge \phi_{\leqslant \ell}(x)$$

A sentence is *local* if it is a boolean combination of basic local sentences.

Theorem 3.9 (Gaifmain [11]). *Every FO sentence is logically equivalent to a local FO sentence.*

Hence, FO can only express local properties. For FO(<) this is not an issue as the distance between any two elements is at most 1; for FO(+1), where elements can be at an arbitrary distance, the situation is different. For instance, the first-order definable string language in which every a is eventually followed by a b is not definable in FO(+1) as the a and b which must be inspected together can have, when the alphabet contains at least three symbols, an arbitrary number of different symbols in between and thus be arbitrarily far apart.

Using this property of FO, one can exactly capture the expressivity of FO(+1). Thereto, for $\ell, r \in \mathbb{N}$, strings w_1 and w_2 are *ℓ, r-equivalent* if the prefix and suffix of length $\ell - 1$ of w_1 and w_2 are the same; and every string v of length ℓ occurs either k times in both w_1 and w_2, where $k \leqslant r$, or v occurs more than r times in both w_1 and w_2. Hence, when only looking locally up to distance ℓ and counting up to a threshold r, w_1 and w_2 are the same. A language is *ℓ, r-locally threshold testable* if, for every two strings which are ℓ, r-equivalent, either both or neither are in the language. A language is *locally threshold testable* if it is ℓ, r-locally threshold testable for some ℓ and r.

Theorem 3.10 (Thomas [27]). *A string language is definable in FO(+1) if and only if it is locally threshold testable.*

These two theorems concerning the expressiveness of first-order logic are proved and discussed in more detail in Chapter **??**. We finally note that, though FO is a fragment of MSO, also here the complexity of translating a formula into an automaton or deciding satisfiability is non-elementary [22].

Different in flavour than FO and MSO are temporal logics, such as Linear Temporal Logic (LTL), which consider a word as a sequence in time. This allows to express properties such as "in the future there is an a and, from this point in time, there are only b's", or "until there is a b, there is always an a". I turns out that such statements can exactly express all first-order definable properties on strings.

Theorem 3.11 (Kamp [15]). *A string language is definable in FO if and only if it is definable in LTL.*

Restrictions of first-order logic As mentioned in the introduction, first-order logic over general (infinite) structures is very powerful and, therefore, undecidable. In search of decidable logics, various restrictions of FO have been studied in classical model theory [3]. When considering strings, this is not an issue: even MSO is decidable in this setting (Corollary 3.2). Here, however, we have a different problem: complexity. Indeed, basic problems, such as testing satisfiability of an FO formula, have a non-elementary complexity [22], the problem being that a translation to an automaton can require non-elementary time. Hence, as has been done in classical model theory, we look at two orthogonal restrictions of first-order logic: (1) restricting the number of quantifier alternations, and (2) restricting the number of different variables in the formula.

Let us first consider the first restriction, the number of quantifier alternations. The number of quantifier alternations is strongly related to the number of nested negations in the formula. Indeed, consider a formula in *prenex normal form*, that is, a formula which consists of a sequence of quantifiers followed by a quantifier-free formula, such as $\phi = \exists x_1 \forall x_2 \forall x_3 \exists x_4\, x_1 < x_2 \wedge x_3 < x_4$. It is equivalent to the formula $\exists x_1 \neg \exists x_2 \exists x_3 \neg \exists x_4\, x_1 < x_2 \wedge x_3 < x_4$, or, in other words, the two quantifier-alternations are implicitly two nested negations. In the translation from a logical formula to an automaton, every negation corresponds to the complementation of the automaton, which yields an exponential blow-up, and hence n nested negations can lead to an automaton of non-elementary size in n (Theorem 3.5). Hence, reducing the number of quantifier alternations can drastically reduce the complexity of basic problems concerning first-order formulas.

This leads us to consider the following subclasses of FO($<$). For any $n \in \mathbb{N}$, let Σ_n and Π_n be the class of all formulas in prenex normal form which start with an existential and universal quantifier, respectively, and have $n-1$ quantifier alternations. For example, the formula ϕ above is in Σ_3. The class of boolean combinations of formulas in Σ_n is denoted by $\mathcal{B}\Sigma_n$.

Restricting the quantifier alternations also restricts the expressive power of the logic and hence it is important to understand the exact expressive power of the different classes. The following theorem, which is discussed in more detail in Chapter **??**, reveals the close connection to the Σ_n-hierarchy and a natural hierarchy within the star-free languages.

For a class of languages $\mathcal{L}$ over an alphabet A, the *polynomial closure* of $\mathcal{L}$ is the set of languages that are finite unions of languages of the form $L_0 a_1 L_1 \cdots a_n L_n$, where each $a_i \in A$ and each $L_i \in \mathcal{L}$. The 0th level of the hierarchy consists of the languages $\emptyset$ and A^*; and, for every $n \geqslant 0$, level $n + 1/2$ is the polynomial closure of level n, and level $n+1$ is the boolean closure of level $n + 1/2$.

Theorem 3.12 (Thomas [27]). *For any n, a string language is definable in*

- $\mathcal{B}\Sigma_n$ *if and only if it is of level* n*;*
- Σ_{n+1} *if and only if it is of level* $n + 1/2$*;*
- Π_{n+1} *if and only if its complement is of level* $n + 1/2$.

The second restriction mentioned above is the number of variables used in the formula. On general structures, this leads to an infinite hierarchy of classes as, for instance, n variables are needed to express that a graph consists of at least n vertices. On strings the situation is different: any first-order property can be expressed using only three variables.

Theorem 3.13 (Kamp [15]). *Any FO sentence on strings is logically equivalent to an*

FO^3 sentence.

A consequence of this theorem is that FO^3 inherits the negative properties of FO. For instance, satisfiability of FO^3 still has a non-elementary worst case complexity [22]. This naturally leads us to consider the fragment of FO with only two variables, FO^2. The complexity of this fragment is again significantly better:

Theorem 3.14 (Etessami, Vardi, Wilke [10]). *The satisfiability problem for FO^2 is NEXPTIME-complete.*

Moreover, it also defines a natural class of languages. We have seen that FO corresponds to LTL; FO^2, in contrast, is expressively equivalent to unary-LTL [10]. That is, LTL without the Until operator. Several other characterizations of this class exist and are described in more detail in Chapters **??**.

Although the two restrictions we have now studied (quantifier alternation and number of variables) seem at first sight to be completely orthogonal, they are in fact very closely related.

Theorem 3.15 (Therien, Wilke [26]). *A string language is definable in FO^2 if and only if it is definable in both Σ_2 and Π_2.*

3.2 Trees

In this section, we study the finite model theory of trees. Thereto, we reconsider most of the logics of the previous section. Some results are almost identical in the tree setting as they are on string, such as, for instance, MSO which captures exactly the regular tree languages. Others are quite different. For instance, unary Transitive Closure logic does not capture all of MSO on trees. For others still, the situation is not clear. FO($<$) is a prominent example of a logic on trees which, thus far, resists classification.

Monadic second-order (MSO) logic As for strings, MSO captures exactly the regular tree languages.

Theorem 3.16 (Doner [6], Thatcher, Wright [25]). *A tree language is definable in MSO if and only if it is regular.*

A complete proof is given in Chapter **??**. The same result holds for infinite trees [21], which is discussed in more detail in Chapter **??**. This theorem has similar consequences as its counterpart on strings (Corollaries 3.2-3.6). Let us here only mention that the proof of Theorem 3.16 gives an easy normal form for MSO on trees as well.

Corollary 3.17. *Any MSO formula on trees is equivalent to an MSO formula of the form $\exists X_1 \cdots \exists X_k \phi$, where ϕ is in FO^2.*

We next consider unary Transitive Closure logic. In contrast to the situation on strings, TC is strictly included in MSO on trees [24]. This problem, which was open for quite some time, was solved by finding an appropriate automata model which captures the expressive power of TC on strings. As automata are in general very well suited to prove that they can not express a language - think, for instance, about the pumping lemma - this then allows to prove that some regular tree language can not be expressed by these automata, thus separating TC and MSO on trees. To give an idea of this proof, we present the automata model and the separating language.

Definite nested tree walking automata.

Lemma 3.18. *A tree language is definable in TC if and only if it is accepted by a nested tree walking automaton.*

Present the separating language.

Lemma 3.19. *The language L_{sep} can not be accepted by a nested tree walking automaton.*

As L_{sep} is regular ...

Theorem 3.20 (ten Cate, Segoufin [24]). *Any tree language definable in TC is definable in MSO. There exist tree languages definable in MSO which are not definable in TC.*

First-order (FO) logic The logic FO($<$) on trees is not well understood yet. In particular, there is no algorithm which, given a regular tree language, decides whether it is definable in FO($<$). Moreover, there exist definitions of star-free tree expressions which coincide with the first-order definable tree languages [2, 13], while another captures all of the regular tree languages [20].

What is clear, is that first-order logic can behave quite differently on trees than on strings. Consider, for instance, the language of all full binary trees in which all leaves are at an even depth. As this language seems to require modulo counting, which first-order logic cannot do, one expects that it is not definable in FO. However, using a simple but clever trick allows to express it, as is explained in more detail in Chapter **??**. On the other hand, there exist regular tree languages which do not require counting, that is, are aperiodic, but are not expressible in FO($<$) [13].

The logic FO(+1) on trees is better understood as it corresponds to the locally treshold testable tree languages and as it can be decided whether a regular tree language is definable in FO(+1) [1]. In Chapter **??**, these results are discussed in more detail.

Restrictions of first-order logic Mention FO^2 and XPath, maybe one word about Σ_n hierarchy.

Modal logics Mention μ-calculus and other modal logics. LTL, CTL, CTL*, ...

3.3 Applications

Typechecking XML transformations XML is a standard for describing and exchanging data and XML documents are typically required to obey to a schema, which describes the form the document must have. In formal language theoretic terms, an XML document is a tree t and the schema is a regular tree language L, which requires that $t \in L$. Moreover, there exist XML transformations f, which take an input tree t and transform it into an output tree $f(t)$. When the input trees must adhere to schema L and the output trees to schema K, it is useful to know whether, for every tree $t \in L$, also $f(t) \in K$, that is, whether the transformation only produces valid trees. This problem is called the *typechecking problem*. This is studied in [18], where K and L are modeled as tree automata and f as a k-pebble transducer, a formalism which allows to simulate several existing transformation languages for XML documents.

To show that this problem is decidable, it is shown that the language $f^{-1}(K) = \{t \mid f(t) \in K\}$ is definable by a so-called k-pebble automaton. The answer to the typechecking problem is positive if $f^{-1}(K) \subseteq L$. To decide this, it is shown that the k-pebble automaton can be translated to a tree automaton (inclusion of tree automata is decidable), that is, that k-pebble automata define only regular languages. This last, crucial step is accomplished by using the equivalence of MSO and the regular tree languages (Theorem 3.16). The flexibility of MSO allows to express the behaviour of a k-pebble automaton, and by Theorem 3.16 the obtained MSO formula can be translated to a tree automaton.

MSO interpretations? MSO-interpretations on words is non-deterministic 2-way transducers [Engelfriet, Hoogeboom]. Reference transducers section.

Decomposition Method for MSO?

Verification?

XPath XPath is a popular query language for XML documents, which has strong connections to logic. It allows to navigate through the tree over the axes of the tree. This "navigational core" of XPath, also sometimes called Core XPath, has exactly the expressive power of FO^2 over trees. Moreoever, the latest XPath standard, XPath 2.0, is explicitly modeled to be expressively equivalent to full first-order logic. These connections to well-studied logics, make clear which queries are, and which are not, expressible in XPath.

4 Automata and finite model theory of arbitrary structures

In the first part of the chapter, we discussed the Finite Model Theory of strings and trees in which automata play an essential role through the tight connection established by the Theorems of Büchi [4], Elgot [6], and, independently, Trakhtenbrot [30] for strings and Doner [6] and Thatcher and Wright [25] for trees.

We have seen that this connection has many kinds of nice consequences for Finite Model Theory on strings and trees, e.g., it allows to use automata as an algorithmic model to evaluate formulas, for testing satisfiability of formulas and to understand the expressive power of logics. It is a natural desire to exploit the connections between logic and automata beyond strings and trees, i.e., for more general classes of finite structures, thus following the maxim “to a hammer, the whole world looks like a nail”.

Indeed, many ways have been found to translate a setting with general finite structures into a setting with strings and trees (nails!) in order to apply the logic-automata connection. We will present examples for different approaches in this second part of the chapter.

The first kind of approaches generalizes the logic-automata connection to classes of structures that are not “too much” different from trees. As examples we will see Courcelle’s famous theorem about monadic second-order logic on classes of structures of bounded tree-width and recent investigations of logics on strings and trees that are extended by data values.

Furthermore, there is the classical model-theoretical approach of logical interpretations, (also known as logical reductions). Here, either a reduction from strings or trees to general structures is exploited to prove inexpressibility or lower bounds or a reduction from general structures to trees is exploited to obtain an algorithm witnessing decidability or a good complexity. We will present examples for both kinds of reductions.

Yet another possibility is to make use of string- or tree-structured “secondary objects” and to apply automata on them. As an example, we will consider the decidability of guarded fixed-point logic which was shown by tree automata on tableaux proofs.

4.1 Tree-like structures

Classes of structures of bounded tree width

Theorem 4.1. *Every MSO-definable property can be evaluated in linear time on structures of bounded tree-width.*

What about FO model checking on structures with bounded local tree-width (Grohe)? Does this still use automata?

Unravellings This might be already captured in a different chapter. Then it should be only mentioned here.

Data words and data trees Decidability results and extended automata models.

- Alternating 1-register automata
- Data automata

4.2 Logical reductions

- Reduction from $\{a^n b^n \mid n \geqslant 1\}$ to Hamiltonian graphs showing that the latter are not MSO-definable. (Ebbinghaus-Flum book)

- An example for MSO-interpretations in trees (reference to infinite case, Caucal-hierarchy)
- FO-queries over strings (paper with Michael, Leonid, Luc) ?
- Pointer to automatic structures

4.3 Other ways of exploiting the logic-automata connection

Decidability of guarded fixed-point logic which was shown by tree automata on tableaux proofs.

5 Miscellaneous

Things that could be said.

- Logic and automata can be mixed: e.g., a language L can be turned into a logical quantifier Q_L such that, on a linearly ordered structure S, a formula $Q_L x\ \varphi(x)$ would hold, if the (0-1-) word induced by φ on the linear order is a string of L. See [**?**].
- Logical characterization of the context-free languages.
- FO2 and turtle automata?

References

[1] M. Benedikt and L. Segoufin. Regular tree languages definable in fo and in fo$_{od}$. *ACM Trans. Comput. Log.*, 11(1), 2009. 982

[2] M. Bojanczyk. Forest expressions. In J. Duparc and T. A. Henzinger, editors, *CSL*, volume 4646 of *Lecture Notes in Computer Science*, pages 146–160. Springer, 2007. 982

[3] E. Börger, E. Grädel, and Y. Gurevich. *The Classical Decision Problem*. Perspectives in Mathematical Logic. Springer, 1997. 980

[4] J. R. Büchi. Weak second-order arithmetic and finite automata. *Z. Math. Logik und Grundl. Math.*, 6:66–92, 1960. 975, 983

[5] J. R. Büchi. On a decision method in restricted second order arithmetic. In *Logic, Methodology and Philosophy of Science (Proc. 1960 Internat. Congr .)*, pages 1–11. Stanford Univ. Press, Stanford, Calif., 1962. 977

[6] J. Doner. Tree acceptors and some of their applications. *J. Comput. System Sci.*, 4:406–451, 1970. 977, 981, 983

[7] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer-Verlag, 1995. 971, 972, 978

[8] H.-D. Ebbinghaus, J. Flum, and W. Thomas. *Mathematical Logic*. Undergraduate Texts in Mathematics. Springer-Verlag, second edition, 1994. 972

[9] C. C. Elgot. Decision problems of finite automata design and related arithmetics. *Trans. Amer. Math. Soc.*, 98:21–51, 1961. 975

[10] K. Etessami, M. Y. Vardi, and T. Wilke. First-order logic with two variables and unary temporal logic. *Information and Computation*, 179(2):279–295, 2002. 981

[11] H. Gaifman. On local and non-local properties. In *Proceedings of the Herbrand Symposium, Logic Colloquium '81*, pages 105–135, 1982. 979

[12] Y. Gurevich. Toward logic tailored for computational complexity. In M. Richter et al., editors, *Computation and Proof Theory*, volume 1104 of *Lecture Notes in Mathematics*, pages 175–216. 1984. 972

[13] U. Heuter. First-order properties of trees, star-free expressions, and aperiodicity. *ITA*, 25:125–146, 1991. 982

[14] N. Immerman. *Descriptive Complexity*. Springer, 1999. 972

[15] H. Kamp. *Tense Logic and the Theory of Linear Order*. PhD thesis, University of California, Los Angeles, 1968. 979, 980

[16] L. Libkin. *Elements of Finite Model Theory*. Springer, 2004. 972

[17] R. McNaughton and S. Papert. *Counter-free automata*. The M.I.T. Press, Cambridge, Mass.-London, 1971. With an appendix by William Henneman, M.I.T. Research Monograph, No. 65. 974, 978

[18] T. Milo, D. Suciu, and V. Vianu. Typechecking for xml transformers. *J. Comput. Syst. Sci.*, 66(1):66–97, 2003. 983

[19] D. Perrin and J.-E. Pin. *Infinite Words*, volume 141 of *Pure and Applied Mathematics*. Elsevier, 2004. ISBN 0-12-532111-2. 973

[20] A. Potthoff and W. Thomas. Regular tree languages without unary symbols are star-free. In Z. Ésik, editor, *FCT*, volume 710 of *Lecture Notes in Computer Science*, pages 396–405. Springer, 1993. 982

[21] M. O. Rabin. Decidability of second-order theories and automata on infinite trees. *Trans. Amer. Math. Soc.*, 141:1–35, 1969. 977, 981

[22] L. J. Stockmeyer and A. R. Meyer. Word problems requiring exponential time: Preliminary report. pages 1–9, 1973. 977, 979, 980, 981

[23] H. Straubing. *Finite automata, formal logic, and circuit complexity*. Birkhäuser Boston Inc., Boston, MA, 1994. 973

[24] B. ten Cate and L. Segoufin. Xpath, transitive closure logic, and nested tree walking automata. In M. Lenzerini and D. Lembo, editors, *PODS 2008*, pages 251–260. ACM, 2008. 981, 982

[25] J. W. Thatcher and J. B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Math. Systems Theory*, 2:57–81, 1968. 977, 981, 983

[26] D. Thérien and T. Wilke. Over words, two variables are as powerful as one quantifier alternation. In *Thirtieth Annual ACM Symposium on the Theory of Computing*, pages 234–240, 1998. 981

[27] W. Thomas. Classifying regular events in symbolic logic. *J. Comput. System Sci.*, 25(3):360–376, 1982. 979, 980

[28] W. Thomas. Languages, automata, and logic. In *Handbook of formal languages, Vol. 3*, pages 389–455. Springer, Berlin, 1997. 973

[29] B. A. Trakhtenbrot. The impossibilty of an algorithm for the decision problem for finite models. *Doklady Akademii Nauk SSR*, 70:569–572, 1950. 972

[30] B. A. Trakhtenbrot. Finite automata and monadic second order logic (russian). *Siberian Math. J*, 3:103–131, 1962. (English translation in *Amer. Math. Soc. Transl.* **59**, 1966, 23–55). 975, 983

Finite automata, image manipulation and automatic real functions *

Juhani Karhumäki and Jarkko Kari

Department of Mathematics, University of Turku, Finland
email: karhumak, jkari@utu.fi

2010 Mathematics Subject Classification: 68Q45, 68U10

Key words: weighted finite automaton, weighted finite transducer, automatic real function, fractal image generation, image transformation

Contents

1 Introduction

Finite automata are simple devices capable of describing and classifying finitary objects. In the simplest form they are used to specify the family of regular languages. Different characterizations of this family highlight its mathematical importance, and some of those like those via syntactic notions, emphasize the finitary nature of the family.

However, in contrast to above, finite automata also have several features which make them very powerful devices in modelling complicated phenomena. A crucial such fact is *recursion*. Indeed, due to the very basic requirement, finiteness, any finite automaton returns repeatedly to a same local situation during its evolution (computation). Hence, it is

*Work supported by the Academy of Finland projects 121419, 126211 and 131558

intuitively quite believable that these devices can be used to define complicated recursive structures, such as fractal type phenomena, see e.g., [3] or [20]. This is one point we want to make clear in this chapter.

Another phenomenon which makes finite automata very powerful in some context is their extendability. Indeed, one can consider finite automata on *infinite words*, which immediately brings nonconstructive elements to the theory. We illustrate that in a moment. Also one can add the notion of *multiplicity*, that is a mechanism to count how many times, or on which weight, something is computed. The latter approach was taken by M.P. Schützenberger [23] already at the very early stage of the theory, and its power was nicely witnessed by S. Eilenberg in [15] by showing, e.g., that this leads to very simple and natural questions of finite automata which are algorithmically undecidable. These both extensions of the theory of finite automata are employed in this presentation.

Interaction is another tool which can be used to make finite automata extremely powerful. Indeed, the theory of cellular automata, which was introduced even before that of finite automata [24], is just a theory of homogenous finite automata interacting according to a single unified protocol. This extension yields universal computing power. Consequently, it allows to specify many phenomena, which are extremely complicated in classical sense of natural sciences, in a very simple automata-theoretic setting. This is a view which is strongly sold by S. Wolfram in his monograph [25]. Our presentation here supports this view.

As illustrations of above lines we recall a few more concrete examples. *Cantor's Dust* is an example of an anomaly which played an interesting role in the history of mathematics. It is defined as a subset of the unit interval $I_0 = [0, 1]$ as follows. Divide I_0 into three equally long parts, delete the middle one, and call the remaining set by I_1. Repeat the process for segments in I_1 to obtain I_2. Do the same for I_2, and further on infinitely many times. The remaining set is called Cantor's Dust I_∞. As an easy calculation shows, the total length of the intervals which have been removed is equal to that of the length of the original interval. So, what does remain to I_∞? A simple answer is seen via infinite words. That is, let us associate each number ζ in I_0 to its unique infinite ternary representation w_ζ in $W = 0^\omega \cup \{0, 1, 2\}^\omega \setminus \{0, 1, 2\}^* 0^\omega$. It follows straightforwardly from the construction that

$$I_\infty = \{w \in W \mid 1 \text{ does not occur in } w\} .$$

In order to have this formally precise we always delete half open intervals, including the right end point. Then I_∞ is nondenumerable, and we are unavoidably touched with nonconstructivity, although the set of all words is defined by a trivial finite automaton on infinite words.

As another illustration, also essentially connected to our models of automata, we recall a few results on automata with multiplicities. We assume here that multiplicities are taken from the semiring of nonnegative integers, that is we consider ordinary finite nondeterministic automata, and count how many times a word is accepted. In Eilenberg's terms we consider $\mathbb{N} - \Sigma$- automata. Now, a basic problem of inclusion, that is asking whether a language accepted by a finite automaton is included in that accepted by another one, turns from decidable into undecidable when ordinary finite automata are replaced by finite automata with multiplicities. The former is a folklore result on automata while the latter follows from the Post correspondence problem via an embedding

$$\varphi : \ \Sigma^* \hookrightarrow \mathcal{M}_{2\times 2}(\mathbb{N}),$$

that is, via an injective morphism from the word monoid Σ^* into the multiplicative monoid of 2×2 matrices over $\mathbb{N}$. Obviously the above undecidability result can be translated to that on matrices over $\mathbb{N}$. Accordingly, one can also see that the former problem for matrices is a problem asking whether for a finite set of matrices over $\mathbb{N}$ there exist two sequences of these matrices with equal products. This problem remains undecidable even if upper triangular matrices of dimension 3 are considered, see [16]. We stated this not only as an example of the power of finite automata, but also since upper triangular matrices are used in our model of automata for image manipulation.

The goal of this chapter is to show how the theory of finite automata can be used for image manipulation, and in particular that at least for some aspects of that theory, finite automata are very useful. The idea of using finite automata to describe real world images grew from different considerations. Berstel and Morcrette [4] were among the first to use finite automata for this purpose. In the same spirit Lindenmayer and Prusinkiewicz [20] used the theory of L-systems to generate fractal type images. Around the same time K. Culik initiated his systematic research using finite automata to describe real world images, see [6] and [7]. A goal here was to provide finite automata as an alternate to Iterated Function Systems used by Barnsley [3] for image generation and manipulation.

In early 1990's the theory became established and the term weighted finite automata (WFA for short) was fixed for these devices. In this model ordinary finite automata are extended in two directions: the weights, that is real numbers, are used to model the multiplicity of computations, and automata operate on infinite inputs. Research on this field emerged simultaneously to theoretical foundations [9, 12], as well as to practical implementations [9]. Comprehensive surveys of this research can be found in [11], [18] and [1].

As a more specific goal of this chapter we do not intend to give complete survey of this topic – for that we refer to surveys just mentioned – but instead try to focus on two specific aspects. Namely, we show how several well known properties, in particular closure properties, of finite automata are very powerful and useful in this theory. Secondly, we concentrate on models, where WFA are used to compute real functions $f : [0, 1) \to \mathbb{R}$ (or more generally $f : [0, 1)^n \to \mathbb{R}$). We call such functions *automatic real functions* (to distinguish them from automatic functions of [2]).

The first point is reflected in the fact that when a digital picture is compressed to WFA-representation, many modifications of this picture can be done as operations on this compressed WFA-representation. Of course, closure properties, including the seldom used complete product where also inputs become pairs, play a central role here.

In the second point we analyze the power of this model in both directions: "what can be done?" and "what cannot be done?". It turns out that relatively few, in a classical sense, nicely behaving functions can be computed, but on the other hand some monster-type functions can be computed efficiently. It also turns out that some natural operations of functions, like integration, can be computed uniformly based on the WFA-representation of the function.

The structure of this chapter is as follows.

In Section 2 we recall definitions and fix the terminology. We define our model of

WFA, as well as explain how digital images are represented by WFA. The use of WFA for image generation, computation and, in general, manipulation is described. A special feature of our chapter is that we use WFA to compute real functions $f : [0,1) \to \mathbb{R}$. The argument $\hat{x}$ of f, a real number, is given to the WFA $\mathcal{A}$ as the infinite binary word x representing $\hat{x}$. The value of the function is the multiplicity given by $\mathcal{A}$ on input x. Consequently, convergency considerations come into the play. We also analyze what the continuity of a function means in their WFA-representation. We call these functions computable by WFA as *automatic real functions*.

Sections 3 and 4 search for borderlines for the computing power of WFA. On the one hand, in Section 4 we give examples of what can be computed by these models. Typically the functions are of fractal type and noncontinuous, but we also show how all polynomials are among WFA-computable functions. In fact, there is a single automaton capable of computing all polynomials of at most a given degree. On the other hand, we show in Section 3 that the only smooth functions which can be computed by these models are polynomials. It also turns out that automatic real functions are not closed under the usual composition of functions. We return to these questions in Section 5. At the end of Section 3 we analyze decidability issues of weighted finite automata.

Section 5 is central to our presentation. We apply known (and slightly modified) closure properties of finite automata and analyze how they are reflected to functions defined by corresponding WFA. The core here is that many transformations of images, or functions, can be carried out on the level of their WFA-representation. Also rather seldom used complete direct product of automata, where not only states are paired, but also input symbols are paired, turns out to be useful: it allows to transform 1-dimensional images into 2-dimensional images. As a consequence of these closure questions, for example, gradually – either linearly or quadratically – darkening pictures are easy to create. Another feature achievable in our model is that automatic real functions are closed under convolution, and hence integration of a function becomes possible. Finally, the theory of finite transducers is modified for WFA as a further tool of image manipulation.

In Section 6 we further analyze the power of WFA as a method to compute functions. Namely, we introduce a simple automaton which computes an everywhere continuous function which, however, does not possess a derivative at any point. The construction provides 1-parameter family of 4-state automata, all of those defining, by out earlier criteria, a continuous function. Also all are of fractal type, and almost all appear to be the monster like function described above – only a few special values of the parameters give a function which has a derivative in exactly those points having a finite binary representation. An interesting point here is that the computation of the values of these functions is not more demanding than the computation of the values of a cubic polynomial – as automatic real functions. Indeed, this can be done in low polynomial time.

2 Definitions and notation

In this section we define the basic notions and fix the terminology of our presentation. This requires, e.g., to interrelate quite different areas of mathematical research. On the one hand we are dealing with very concrete objects like natural (and artificial) images, or their

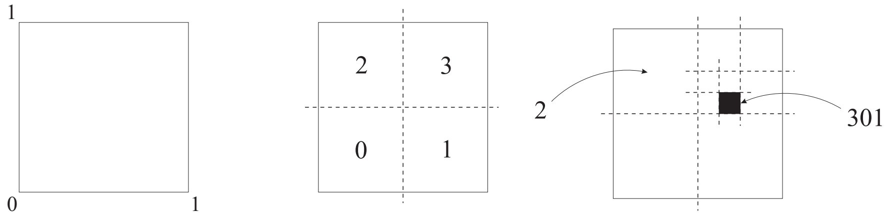


Figure 1. The unit square, its division into four subsquares and the addresses of the pixels.

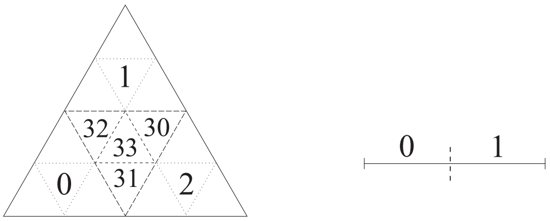


Figure 2. Alternative division patterns

one-dimensional counterparts, functions, and how they can be represented in terms of words. On more theoretical side, we are interrelating automata theory to that of classical analysis, that is the theory of functions – a connection which is rather rarely analyzed. Crucial notions here are matrices and their products, allowing the use of powerful theory of linear algebra.

We denote by A a finite alphabet. In our considerations it is typically binary, say $A_2 = \{0,1\}$, or quartic, that is $A_4 = A_2 \times A_2$, which we rename as: $(0,0) \leftrightarrow 0$, $(0,1) \leftrightarrow 1$, $(1,0) \leftrightarrow 2$ and $(1,1) \leftrightarrow 3$. An *image* F is a formal power series

$$F : A^n \to S\,,$$

where S is a semiring. Semirings considered here are those of number semirings, such as $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{Q}$ or $\mathbb{Q}_+$. The power series F, for $A = A_4$, can be interpreted as an n-resolution picture, and hence referred to as an *n-image*, as described below.

Consider the unit square on Euclidean space, and its division into four equal subsquares shown in Fig. 1. The subsquares are addressed by elements of A_4. We repeat the process n times. Then any element of A^n can be viewed as an *address* of a pixel in this division as shown in Fig. 1. As a conclusion the formal power series F is a multiresolution representation of a picture, where for each $w \in A_4^n$, $F(w)$ tells the value of the pixel addressed by w in the $2^n \times 2^n$ -resolution image.

The above representation is very flexible. There is no reason to consider only the (unit) square as the shape of the picture. This can equally well be a rectangular or, for example, a triangle or a line as illustrated in Fig. 2.

The division pattern, and hence also the size of the address alphabet seems to be irrelevant (up to technicalities) for any presentation. The choice of the values of F allows to consider in the same framework

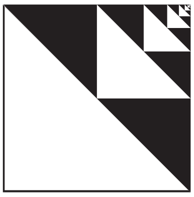
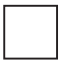
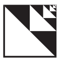
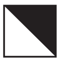


Figure 3. Sierpinski's triangle

- *black and white* pictures, when S is the Boolean set $\{0, 1\}$,
- *greytone* pictures, when S is $\mathbb{R}$ or its subset,
- *colour* pictures, when S is $\mathbb{R}^3$ or its subset.

In the last case the three components of the image tell, in a natural way, the *intensity* of the three main colour components.

We continue with an example. Consider the picture of a (modified) Sierpinski's triangle depicted in Fig. 3. The above division argument yields subpictures

and the first automaton depicted in Fig. 4 telling how these subpictures are related to the addresses. After the second step of divisions we obtain the second automaton depicted in Fig. 4. From now on the process gets stabilized, no new states are introduced in divisions. The automaton becomes completed when the original picture is chosen to initial state of the automaton, as marked by ↑. Now, we obtain an approximation, an $2^n \times 2^n$ -resolution image, of Sierpinski's triangle by associating to each address $w \in A_4^n$ the subimage given by that state of the automaton where w leads from the initial state. Interestingly, in this example all resolutions are computed by the same automaton – a property which, of course, is not true in general.

In the above example the value of a pixel was determined by the state reached by the address of that pixel in an automaton. An alternative way of defining this is to associate outputs to each transition of the automaton. When doing this we have to calculate which ”amount of darkness” goes with each transition. Easy calculations show that this leads from the second automaton of Fig. 4 to that of third automaton of Fig. 4, where the numbers after semicolons tell these portions.

In this automaton the darkness of a pixel is obtained as the product of weights of its address. Relative darknesses of pixels of any level are in correct portion. However, the image is getting paler and paler when the resolution is increased – indeed, the weight of any infinite computation tends to zero. In order to avoid this we *scale* the automaton as follows. Since in each division step we introduce four times more pixels, we multiply all the weights by four. This yields out final (weighted) automaton approximating the Sierpinski's triangle.

Actually the second and fourth automata of Figure 4 compute the same $2^n \times 2^n$ -resolution approximations of our image. The difference being that in the second automaton the value of a pixel is determined by the end state, while in the fourth it is obtained as the product of all weights of the computation. A few first approximations are shown in Fig. 5.

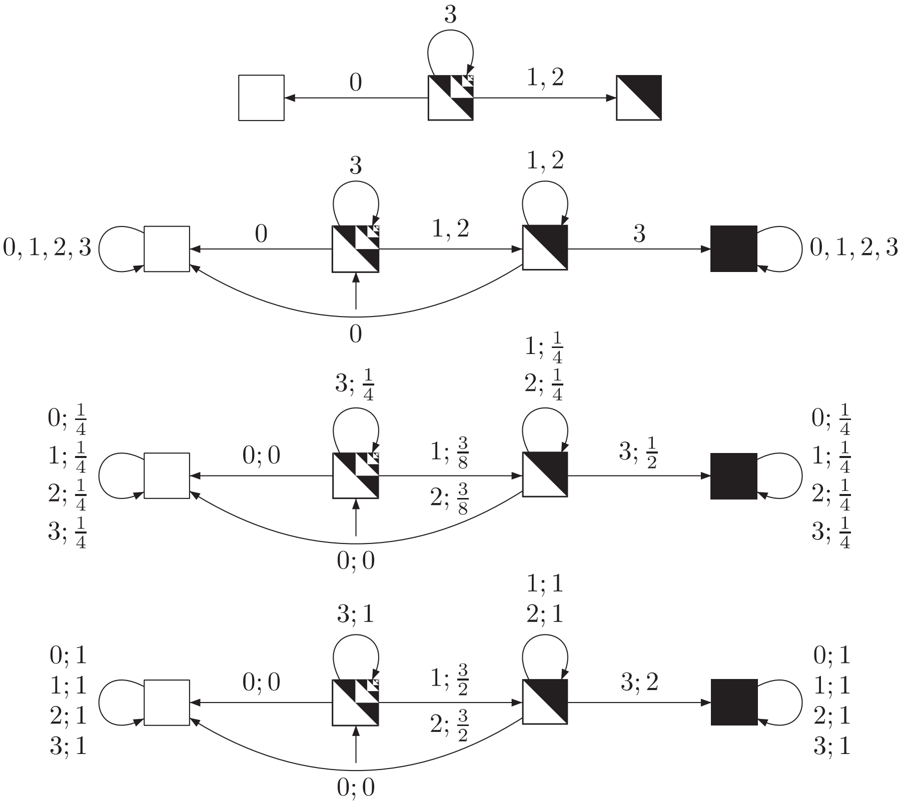


Figure 4. The four steps in the construction of an automaton for Sierpinski's triangle: after the first and second divisions, after adding weights and the final one after normalization.

Before going into more formal definitions we want to make a few remarks. In above we were very detailed not only to make it clear what we mean by automata-theoretic representation of images, but to emphasize the diversified nature of this approach. Indeed from above it should be clear that the dimension of the picture, that it whether it is one, two or three, makes only notational differences and difficulties.

The above approach also suits to describe approximation of Cantor's Dust. Indeed, it is computed by the two state automaton depicted in Fig. 6. Its behavior is not surprising: on all inputs containing the symbol 1, the weight is zero, while on all other infinite words it tends to infinity.

There is still one more interesting and theoretically important feature in the final automaton. It is *average preserving*, in the sense that in each division step the average of the darkness of a pixel is preserved. Formally this means that, for all $w \in A^n$, the following holds:

$$F(w) = \frac{F(w0) + F(w1) + F(w2) + F(w3)}{4}\,.$$

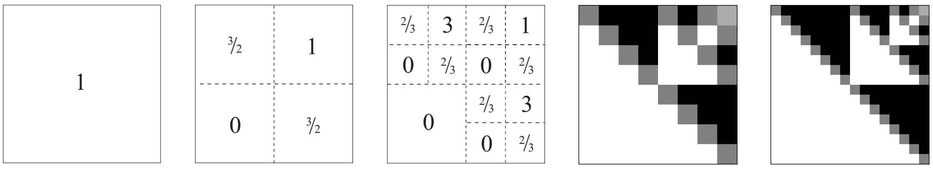


Figure 5. Approximation of Fig. 3 by automata of Fig. 4.

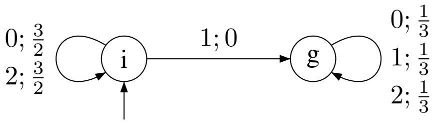


Figure 6. An automaton for Cantor's Dust.

We continue by giving our formal definitions. A *weighted finite automaton* (WFA for short) is a 5-tuple $\mathcal{A} = (Q, A, W, I, T)$, where

- Q is a finite set of states;
- A is a finite input alphabet;
- $W : Q \times A \times Q \to \mathbb{R}$ is a weight function;
- $I : Q \to \mathbb{R}$ is the initial distribution and;
- $T : Q \to \mathbb{R}$ is the final distribution.

The weight function W decomposes in a natural way to functions $W_a : Q \times Q \to \mathbb{R}$ for $a \in A$, and hence can be viewed as a set of matrices over $\mathbb{R}$ indexed by elements of A. A WFA is a natural extension of the notion of *nondeterministic* finite automaton (NFA for short). Indeed, by setting

$$\begin{aligned} E &= \{(p, a, q) \mid W_a(p, q) \neq 0\}\,, \\ i &= \{q \in Q \mid I(q) \neq 0\}\,, \text{ and} \\ t &= \{q \in Q \mid T(q) \neq 0\} \end{aligned}$$

we obtain from $\mathcal{A}$ an NFA (Q, A, E, i, t), so-called *underlying automaton* of WFA $\mathcal{A}$. It is worth mentioning already here that it does not make any difference whether W is total or partial, as long as the zero is allowed as a weight.

The theory of weighted finite automata (or $\mathbb{R} - A$ -automata in terms of Eilenberg) is a well studied research area, see, e.g., [15, 22, 5], usually referred to as the theory of *rational formal power series*. The two aspects which make this chapter justified are the following two features. We use the theory to a particular application, namely to deal with images and real functions, and we consider these automata on infinite inputs.

Let $\mathcal{A} = (Q, A, W, I, T)$ be a weighted finite automaton. It defines the functions

$$F_{\mathcal{A}} : A^* \to \mathbb{R}, \quad F_{\mathcal{A}}(w) = I \cdot W_w \cdot T$$

and

$$f_{\mathcal{A}} : A^\omega \to \mathbb{R}, \quad f_{\mathcal{A}}(w) = \lim_{n \to \infty} I \cdot W_{\text{pref}_n(w)} \cdot T\,,$$

where pref_n denotes the prefix of length n and the matrix W_w for $w = a_1, \ldots a_t$ is defined by $W_w = W_{a_1} \cdots W_{a_t}$. As we noted (due to zero weights) $F_{\mathcal{A}}$ can be assumed to be always defined while $f_{\mathcal{A}}$ is only a partial function – if the limit on the right does not exist then the value of $f_{\mathcal{A}}$ is undefined. We call $F_{\mathcal{A}}$ and $f_{\mathcal{A}}$ as *word and ω-word functions* defined by $\mathcal{A}$, respectively.

The question whether or not $f_{\mathcal{A}}$ is a total function seems to be highly nontrivial, in general. This is one reason why we want to introduce a special class of WFA such that this problem does not occur. The other reason is that this special class is large enough to illustrate the power of automata theory in image manipulation. Most of the examples presented so far are, at least essentially, within this class.

A WFA $\mathcal{A} = (Q, A, W, I, T)$ is called a *level* automaton if the following conditions are satisfied:

(i) the only loops in the underlying automaton are $p \xrightarrow{a} p$;
(ii) $W(p, a, q) \geqslant 0$ for all $a \in A$ and $p, q \in Q$;
(iii) for every $p \in Q$:
  - $W(p, a, p) < 1$ for all $a \in A$, if there exists $q \neq p$ and $b \in A$ such that $W(p, b, q) \neq 0$;
  - $W(p, a, p) = 1$ for all $a \in A$, otherwise;

(iv) $I \in \mathbb{R}_+^t$ and $T \in \mathbb{R}_+^t$;
(v) the underlying automaton of $\mathcal{A}$ is reduced.

In terms of classical theory, word functions computed by level automata are $\mathbb{R}_+$-rational. In addition to that the condition (iii) is crucial. It guarantees, as we shall soon see, that an ω-word function is always defined. It also allows to define the *degree* of a state of a WFA as follows: The states with loops of weight one are of degree 0, and a state q is of degree i if there exists a transition from that state to a state of degree $i - 1$ with a nonzero weight, and moreover all transitions from it with nonzero weights go to states of degree at most $i - 1$. The degree of the level automaton is defined to be the maximum of the degrees of its states.

Part of the motivation of introducing the above restricted class of WFA is demonstrated in the following lemma and its consequences. A further motivation comes from the fact, as we shall see, that already this class allows a number of nontrivial and illustrative results and examples.

Let $\mathcal{A} = (Q, A, W, I, T)$ be a level automaton. For any $q \in Q$, let T_q denote the final distribution where state q is assigned value 1 and all other states value 0. We define an ω-word function $f_q : A^\omega \to \mathbb{R}_+$ by

$$f_q(w) = \lim_{n\to\infty} I W_{\mathrm{pref}_n(w)} T_q$$

and prove

Lemma 2.1. *The functions f_q are always defined. Moreover, if q is of positive degree then f_q is the zero function.*

Proof. We use essentially the conditions of a level automaton. They allow to define a majorizing level automaton $\mathcal{A}_{\mathrm{maj}} = (Q, A, W_{\mathrm{maj}}, I, T_{\mathrm{maj}})$ as follows. The weight matrices are obtained from $\mathcal{A}$ by

- keeping the weights of the loops in states of degree 0 as they are in $\mathcal{A}$, that is equal to one;

- replacing all other weights of the loops by
$$\alpha = \max\{W(p,a,p)\} < 1\,,$$
where $p \in Q$ has a positive degree and $a \in A$;
- all other weights by
$$\beta = \max\{W(p,a,q)\}\,,$$
where $p, q \in Q$ and $a \in A$.

The final distribution T_{maj} assigns each state value one. It follows that for each finite $w \in A^*$ and state $q \in Q$

$$F_q(w) = IW_wT_q \leqslant F_{\mathcal{A}_{\text{maj}}}(w)\,. \tag{2.1}$$

Also, by descending induction on the degree i of state q, we conclude that for all $w \in A^\omega$ we have

$$f_{\mathcal{A}_{\text{maj}}}(w) = 0, \quad \text{for } i = n, \dots, 1, \text{and} \tag{2.2}$$
$$f_{\mathcal{A}_{\text{maj}}}(w) \text{ is defined for } i = 0\,. \tag{2.3}$$

In particular, the second sentence of Lemma 2.1 follows from (2.1) and (2.2). Note that to conclude (2.3) we need the well known fact of monotonous sequences of numbers: each such sequence has a limit if it is bounded. To complete the proof we apply the above fact together with conditions (2.1) and (2.3). □

As an immediate consequence of Lemma 2.1 we formulate

Corollary 2.2. *The ω-word function of a level automaton is always defined.*

Actually, we can state even a stronger statement. We recall that the set of all infinite words A^ω becomes a metric space when we define distance function $d : A^\omega \to \mathbb{R}_+$ as follows:
$$d(u,v) = 2^{-|u \wedge v|}\,,$$
where the function $\wedge$ defines the maximal common prefix of infinite words. By convention $d(u,u) = 0$. Now we can formulate

Theorem 2.3. *For each level automaton $\mathcal{A}$ the ω-word function $f_{\mathcal{A}}$ is continuous, in fact, even uniformly continuous.*

Proof. Follows from considerations of Lemma 2.1 and the triangular inequality:
$$\big|f_{\mathcal{A}}(u) - f_{\mathcal{A}}(v)\big| \leqslant \big|f_{\mathcal{A}}(u) - F_{\mathcal{A}}(u \wedge v)\big| + \big|f_{\mathcal{A}}(v) - F_{\mathcal{A}}(u \wedge v)\big|\,. \qquad \square$$

The above shows that WFA, or at least their restriction to level automata, behave quite smoothly as generators of ω-word functions. The picture changes drastically when they are viewed as tools to define images, or to compute real functions. This is an essential point of this chapter.

As shown earlier we can transform a four letter alphabet into coordinates of a plane, or accordingly a binary alphabet into digits of numbers. We are mainly dealing with the binary case and hence analyzing how automata can be used to compute real functions.

Let $A = \{0, 1\}$ be a binary alphabet and $\mathcal{A}$ a WFA with A as the alphabet. An infinite word $\omega = a_1a_2\dots$ over A is viewed as a binary representation

$$0.a_1a_2\dots$$

of a real number. This representation is not unique since for finite word $u \in A^*$ the words

$$u10^\omega \text{ and } u01^\omega$$

represent the same number. The uniqueness is achieved by restricting to infinite words in the set

$$X = A^\omega \setminus A^*1^\omega = \{w \in A^\omega \mid w \text{ contains infinitely many 0's}\}\,.$$

Now, every real $x \in [0, 1)$ possesses the unique binary representation $\mathrm{bin}(x) \in X$. In other words $\mathrm{bin}(x) = a_1a_2\dots$ is the unique word in X satisfying

$$x = \sum_{i=1}^{\infty} a_i 2^{-i}\,.$$

For a binary infinite word ω in X we denote by $\hat{\omega}$ the unique binary number it represents, that is $\omega = \mathrm{bin}(\hat{\omega})$. A class of real numbers, namely those which possess a finite binary representation, are of special interest in our considerations. They are all rational numbers referred to as *dyadic rationals*. Clearly, each dyadic rational $r \in [0, 1)$ is of the form $r = \frac{m}{2^\ell}$ for some nonnegative integers m and ℓ with $m < 2^\ell$.

The real function $\hat{f}_\mathcal{A} : [0, 1) \to \mathbb{R}$ computed by WFA $\mathcal{A}$ is defined by

$$\hat{f}_\mathcal{A}(x) = f_\mathcal{A}\big(\mathrm{bin}(x)\big)\,.$$

Of course, the above allows $\hat{f}_\mathcal{A}$ to be only partial, although for level automaton it is always total by Corollary 2.2. Central questions of our considerations will be those where $\hat{f}_\mathcal{A}$ is either *continuous* or even *smooth*. We recall that a real function is smooth if and only if it has all derivatives on the considered interval.

Theorem 2.3 has the following interesting corollary stating that functions $\hat{f}_\mathcal{A}$ are continuous everywhere except possibly in the denumerable set of dyadic points.

Corollary 2.4. *For any level automaton $\mathcal{A}$, the real function $\hat{f}_\mathcal{A}$ is continuous at any non-dyadic $x \in [0, 1)$, and right continuous at every dyadic point $x \in [0, 1)$.*

Proof. Let $w = \mathrm{bin}(x)$. If x is non-dyadic, w does not end in 0^ω or 1^ω. Let d be the metric of A^ω defined earlier, and d_E the normal Euclidean metric in the interval $[0, 1)$. It is easy to verify (relying on the fact that w does not end in 0^ω or 1^ω) that for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d_E(x, y) < \delta \Longrightarrow d\big(\mathrm{bin}(x), \mathrm{bin}(y)\big) < \varepsilon\,. \tag{2.4}$$

The result then immediately follows from the continuity of $f_\mathcal{A}$ (Theorem 2.3). The proof of the right continuity at dyadic rationals $x \in [0, 1)$ is similar, except that the implication (2.4) is only guaranteed to hold for $y \geqslant x$. □

We call the subinterval

$$\left[\frac{m}{2^\ell}, \frac{m+1}{2^\ell}\right) \subseteq [0, 1)$$

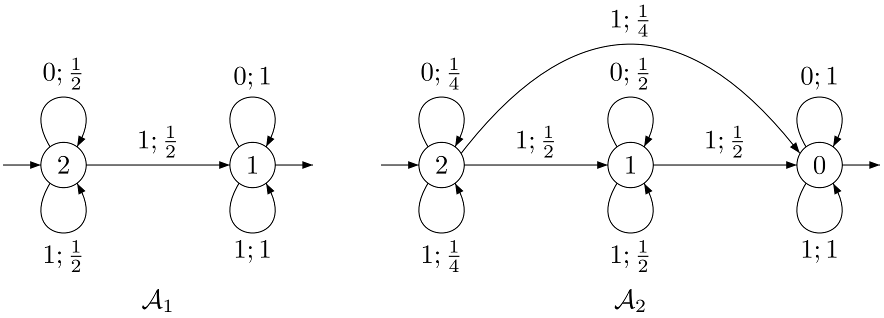


Figure 7. Two level automata

a *dyadic interval*, and consistently to our earlier explanations define its *address* to be the prefix of $\mathrm{bin}\left(\frac{m}{2^\ell}\right)$ of length ℓ. In other words the address of the above interval is the length ℓ binary expansion of the integer $m < 2^\ell$.

Example 2.1. Consider the WFA depicted in Figure 7. In these illustrations we use incoming arrows to represent initial distributions, in this example $(1, 0)$ and $(1, 0, 0)$, respectively, and outgoing arrows to represent final distributions, here $(0, 1)$ and $(0, 0, 1)$, respectively.

Let us analyze the automaton $\mathcal{A}_1$. Clearly, the weight of $w = 0110^\omega$ is

$$\frac{1}{4} + \frac{1}{8} = \frac{3}{8}.$$

In other words $\hat{f}_{\mathcal{A}_1}\left(\frac{3}{8}\right) = \frac{3}{8}$. Interestingly, we also have

$$f_{\mathcal{A}_1}(0101^\omega) = \frac{1}{4} + \frac{1}{16}\sum_{i=0}^{\infty}\frac{1}{2^i} = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}.$$

Actually, simple general computations show that $\hat{f}_{\mathcal{A}_1}(x) = x$ for all $x \in [0, 1)$. Similarly, we can conclude that $\hat{f}_{\mathcal{A}_2}(x) = x^2$ for all $x \in [0, 1)$. These verifications are left to the reader.

In the above examples the functions computed by WFA were not only continuous but also smooth. As we shall see, and will illustrate in the next example, this happens very rarely.

As we hinted in the above example, a necessary condition for the continuity of a function $\hat{f}_{\mathcal{A}}$ at point $x = \frac{1}{2}$ is the equality

$$f_{\mathcal{A}}(10^\omega) = f_{\mathcal{A}}(01^\omega).$$

The next theorem formulates the continuity of $\hat{f}_{\mathcal{A}}$ at dyadic rationals in terms of simple conditions on the ω-word function $f_{\mathcal{A}}$. Later we see that these conditions can be effectively checked.

Theorem 2.5. *For any level automaton $\mathcal{A}$ the following conditions are equivalent:*

(i) *$\hat{f}_{\mathcal{A}}$ is continuous on the interval* $[0,1)$;
(ii) $f_{\mathcal{A}}(u01^{\omega}) = f_{\mathcal{A}}(u10^{\omega})$ *for all words* $u \in A^*$.

Proof. If $\hat{f}_{\mathcal{A}}$ is continuous then for every $u \in A^*$

$$\begin{aligned} f_{\mathcal{A}}(u10^{\omega}) &= \hat{f}_{\mathcal{A}}(\widehat{u10^{\omega}}) \\ &= \lim_{n\to\infty} \hat{f}_{\mathcal{A}}(\widehat{u01^n0^{\omega}}) \\ &= \lim_{n\to\infty} f_{\mathcal{A}}(u01^n0^{\omega}) \\ &= f_{\mathcal{A}}(u01^{\omega}) \,. \end{aligned}$$

Conversely, assume (ii). Continuity of $\hat{f}_{\mathcal{A}}$ at non-dyadic rationals and 0 was established in Corollary 2.4. Consider then an arbitrary positive dyadic rational $x = \widehat{u10^{\omega}}$. As in the proof of Corollary 2.4, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d_E(x,y) < \delta \Longrightarrow d(\text{bin}(y), w) < \varepsilon \text{ for } w = u01^{\omega} \text{ or } w = u10^{\omega} \,.$$

The result now follows from the continuity of $f_{\mathcal{A}}$ and the assumption that $f_{\mathcal{A}}(u01^{\omega}) = f_{\mathcal{A}}(u10^{\omega})$. □

We call a WFA *strongly* continuous, if it computes a continuous function for all initial distributions. For a WFA $\mathcal{A}$, we denote by $\mathcal{A}_I$ the automaton obtained by changing the initial distribution into I. The following corollary is obtained from Theorem 2.5, by considering all initial distributions.

Corollary 2.6. *For any level automaton $\mathcal{A}$ the following conditions are equivalent:*

(i) *$\mathcal{A}$ is strongly continuous on the interval* $[0,1)$;
(ii) $f_{\mathcal{A}_I}(01^{\omega}) = f_{\mathcal{A}_I}(10^{\omega})$ *for all initial distributions I.*
(iii) $f_{\mathcal{A}_I}(01^{\omega}) = f_{\mathcal{A}_I}(10^{\omega})$ *for all unit coordinate vectors* $I = (\ldots 0,1,0,\ldots)$.

Proof. Implication (i) $\Longrightarrow$ (ii) follows from Theorem 2.5, and (ii) $\Longrightarrow$ (iii) is trivial. Let us prove that (iii) implies (i). Let I be an arbitrary initial distribution. By Theorem 2.5, to prove (i) it is enough to show that $f_{\mathcal{A}_I}(u01^{\omega}) = f_{\mathcal{A}_I}(u10^{\omega})$ for all $u \in A^*$. Denote $IW_u = (a_1, a_2, \ldots a_n)$, and let I_i be the i-th unit coordinate vector. Then, by (iii),

$$f_{\mathcal{A}_I}(u01^{\omega}) = \sum_{i=1}^{n} a_i f_{\mathcal{A}_{I_i}}(01^{\omega}) = \sum_{i=1}^{n} a_i f_{\mathcal{A}_{I_i}}(10^{\omega}) = f_{\mathcal{A}_I}(u10^{\omega}) \,. \qquad \square$$

After the above results, it is illustrative to consider the following general example.

Example 2.2. Consider the level automaton shown in Figure 8. Here, we have assumed, consistently to our definition of level automata, that $\alpha, \beta \in [0,1)$, $\gamma, \delta \in \mathbb{R}_+$, the weights of the loops on state q_1 are equal to 1, and that q_0 and q_1 correspond to initial and final states, respectively.

In this special case the continuity and the strong continuity clearly coinside. By (iii) of Corollary 2.6 we see that $\hat{f}_{\mathcal{A}}$ is continuous if and only if

$$f_{\mathcal{A}}(01^{\omega}) = f_{\mathcal{A}}(10^{\omega}) \,,$$

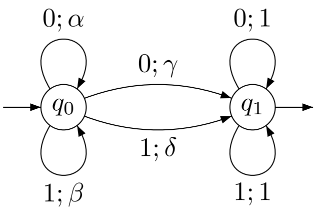


Figure 8. A simple two state level automaton.

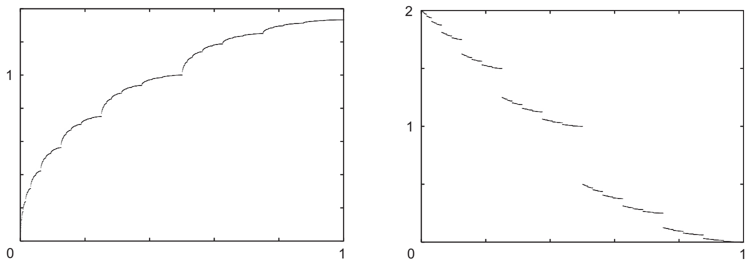


Figure 9. Two examples of functions computed by the automaton of Figure 8.

which can be rewritten as

$$\gamma + \alpha\delta \sum_{i=0}^{\infty} \beta^i = \delta + \beta\gamma \sum_{i=0}^{\infty} \alpha^i \,,$$

or equivalently, as

$$(\alpha + \beta - 1)\big(\delta(1-\alpha) - \gamma(1-\beta)\big) = 0\,.$$

As a conclusion we have shown that the automaton $\mathcal{A}$ of Figure 8 computes a continuous function if and only if one of the following conditions is satisfied:

(i) $\alpha + \beta = 1$;

or

(ii) $\delta(1-\alpha) = \gamma(1-\beta)$.

At this point a few comments are in order. First, for a random choice of parameters α, β, γ and δ the probability that $\mathcal{A}$ computes a continuous function is zero! This justifies our view that WFA compute very rarely nicely behaving functions of analysis. Secondly, one can continue the above analysis, for example, by requiring that $\hat{f}_{\mathcal{A}}$ would have a derivative at point $x = \frac{1}{2}$. A necessary condition would be $\alpha = \beta = \frac{1}{2}$, see [8]. An example of such an automaton is that shown in Figure 7, where $\alpha = \beta = \frac{1}{2}$, $\gamma = 0$ and $\delta = \frac{1}{2}$. That computed the identity function. In Figure 9 two other functions computed by automaton $\mathcal{A}$ of Figure 8 are shown. The first one, with $\alpha = \frac{3}{4}$, $\beta = \frac{1}{4}$, $\gamma = 0$ and $\delta = 1$ computes a continuous function while the second one, with $\alpha = \frac{1}{2}$, $\beta = \frac{1}{4}$, $\gamma = 1$ and $\delta = 0$, a noncontinuous one. Further, as analyzed in [8], the latter function has no derivative in nondenumerably many points.

3 Normal forms, minimality and decidability

In this section we continue our search for basic properties of WFA. We concentrate on level automata, although some results are formulated in a more general setting. In consistence to our approach the automata are over a binary alphabet, that is $A = \{0, 1\}$, however, this is not crucial.

Our first results state important normal forms. For clarity we assume here that the final distribution is $T = (1, \ldots, 1)$. By Lemma 2.1, this is no restriction when level automata are used to compute real functions.

Let

$$\mathcal{A} = (Q, A, W, I, T)$$

with $A = \{0, 1\}$ and $T = (1, \ldots, 1)$, be a WFA. We say that $\mathcal{A}$ is *average preserving*, resp. *0-faithful*, if

$$\sum_{q \in Q, a \in A} W(p, a, q) = 2 \quad \text{for all } p \in Q\,, \tag{3.1}$$

resp.

$$\sum_{q} W(p, 0, q) = 1 \quad \text{for all } p \in Q\,.$$

The above local properties of $\mathcal{A}$ can be defined equivalently as global properties of functions computed by $\mathcal{A}$ as follows:

Lemma 3.1. *If WFA is average preserving, resp. 0-faithful, then*

$$F_{\mathcal{A}}(w) = \frac{1}{2} \sum_{a \in A} F_{\mathcal{A}}(wa) \quad \textit{for all } w \in A^*\,,$$

resp.

$$F_{\mathcal{A}}(w) = f_{\mathcal{A}}(w0^{\omega}) \quad \textit{for all } w \in A^*\,.$$

Proof. Property (3.1) means that 2 is an eigenvalue of the matrix $W_0 + W_1$, and that T is an associated eigenvector. Hence we can compute, for all $w \in A^*$:

$$\begin{aligned}
\frac{1}{2}\big(F(w0) + F(w1)\big) &= \frac{1}{2}(IW_w W_0 T + IW_w W_1 T) \\
&= \frac{1}{2} IW_w (W_0 + W_1) T \\
&= \frac{1}{2} IW_w 2T = IW_w T = F_{\mathcal{A}}(w)\,.
\end{aligned}$$

Therefore, the first claim follows. The second claim is proved in a similar way. □

With suitable conditions of the function W also the reverse of Lemma 3.1 holds. For us, however, the important properties are those formulated in the lemma. The interpretations of these properties are as follows. Being average preserving means that the automaton preserves the average of the function it computes, that is the average value of $F_{\mathcal{A}}$ over words of a fixed length is constant, as is that of the functions $f_{\mathcal{A}}$ and $\hat{f}_{\mathcal{A}}$. This was already explained in Introduction, and it "holds" also for the automaton computing Cantor's Dust. The 0-faithfulness, in turn, means that the values of the function $\hat{f}_{\mathcal{A}}$ on binary dyadic points x are already obtained as the value $F_{\mathcal{A}}(w)$ where $w0^{\omega} = \text{bin}(x)$, and hence they can be effectively computed.

Our next goal is to show that the above special classes of automata are actually normal forms for level WFA. Apart from effectiveness, an analogous result about average preserving WFA holds also for the non-level case, see [19].

Theorem 3.2. *For each level automaton $\mathcal{A}$, one can effectively find an average preserving level automaton $\mathcal{A}_{ap}$ and a 0-faithful level automaton $\mathcal{A}_0$ such that $f_{\mathcal{A}} = f_{\mathcal{A}_{ap}} = f_{\mathcal{A}_0}$.*

Proof. Let $\mathcal{A} = (Q, \{0,1\}, W, I, T)$ be a level automaton. The idea is to transform $\mathcal{A}$ into $\mathcal{A}_{\text{ap}}$ state by state, from low degree states to high degree states. All states of degree zero satisfy (3.1), so they do not need to be transformed.

Let $p \in Q$ have degree at least one, and assume that all states of lower degree already satisfy (3.1). We set

$$k = \frac{2 - W(p,0,p) - W(p,1,p)}{\sum_{q \neq p} \big(W(p,0,q) + W(p,1,q)\big)}$$

and modify $\mathcal{A}$ to $\mathcal{A}' = (Q, \{0,1\}, W', I, T)$ as follows:

(i) the weight of each nonloop transition starting from p is multiplied by k;
(ii) the weight of each nonloop transition entering in p is divided by k;
(iii) the initial distribution value of state p is divided by k.

Now, clearly k is positive and its choice makes the state p average preserving, that is (3.1) is valid for p. Moreover, since transitions from lower degree states were not modified, (3.1) still holds for all lower degree states.

We claim that

$$f_{\mathcal{A}} = f_{\mathcal{A}'}\,.$$

To prove this, for every $q \in Q$, let T_q be the final distribution vector assigining value 1 to state q and value 0 to all other states. If we denote by $\mathcal{A}_q = (Q, \{0,1\}, W, I, T_q)$ and $\mathcal{A}'_q = (Q, \{0,1\}, W', I, T_q)$ the level WFA with final distribution T_q and the original and modified weight functions, respectively, then for every infinite word w,

$$\begin{aligned} f_{\mathcal{A}_q}(w) &= f_{\mathcal{A}'_q}(w) \quad \text{if } q \neq p\,,\\ f_{\mathcal{A}_q}(w) &= \frac{1}{k} f_{\mathcal{A}'_q}(w) \quad \text{if } q = p\,. \end{aligned}$$

It follows from Lemma 2.1 that for $q = p$ we have $f_{\mathcal{A}_q}(w) = f_{\mathcal{A}'_q}(w) = 0$, so we have the equality for all q. Since $\mathcal{A}$ and $\mathcal{A}'$ have the same final distribution $T = (1, 1, \ldots, 1)$, the claim now follows from

$$f_{\mathcal{A}}(w) = \sum_{q \in Q} f_{\mathcal{A}_q}(w) = \sum_{q \in Q} f_{\mathcal{A}'_q}(w) = f_{\mathcal{A}'}(w)\,.$$

Now the final conclusion is easy. We transform the automaton as described above and at each step make one more state to satisfy (3.1).

The proof for 0-faithfulness is analogous. Only the coefficient

$$k = \frac{1 - W(p,0,p)}{\sum_{q \neq p} W(p,0,q)}$$

is used when transforming state p. □

Now we see that values of $\hat{f}_{\mathcal{A}}$ can be effectively computed at dyadic points.

Corollary 3.3. *There is an algorithm to compute $f_{\mathcal{A}}(u0^{\omega})$ for a given level WFA $\mathcal{A}$ and word $u \in A^*$.*

Proof. First convert $\mathcal{A}$ into an equivalent 0-faithful automaton $\mathcal{A}_0$ and compute $f_{\mathcal{A}}(u0^{\omega}) = f_{\mathcal{A}_0}(u0^{\omega}) = F_{\mathcal{A}_0}(u)$. □

Next we associate to any WFA $\mathcal{A} = (Q, A, W, I, T)$ the linear subspaces

$$\mathcal{L} = \langle IW_u \mid u \in A^* \rangle, \text{ and}$$
$$\mathcal{R} = \langle W_u T \mid u \in A^* \rangle$$

of $\mathbb{R}^{1\times n}$ and $\mathbb{R}^{n\times 1}$, respectively, called the *left* and the *right* generated subspace. One can effectively find words $u_1, \ldots, u_m$ such that vectors IW_{u_i} generate $\mathcal{L}$ simply by testing longer and longer words u, until a number ℓ is found such that the vectors IW_u, for all u of length ℓ, are linear combinations of vectors corresponding to shorter words. See [14] for more details. Analogously, generators of $\mathcal{R}$ can be effectively found. By Theorem 2.5, to test continuity of $\hat{f}_{\mathcal{A}}$, it is enough to verify $f_{\mathcal{A}}(u01^{\omega}) = f_{\mathcal{A}}(u10^{\omega})$ for all words $u \in A^*$. But this is equivalent to testing $f_{\mathcal{A}}(u_i01^{\omega}) = f_{\mathcal{A}}(u_i10^{\omega})$ for words $u_1, \ldots, u_m$ such that IW_{u_i} generate the subspace $\mathcal{L}$. Such words u_i can be effectively constructed, as discussed above, and values $f_{\mathcal{A}}(u_i01^{\omega})$ and $f_{\mathcal{A}}(u_i10^{\omega})$ can be effectively computed by Corollary 3.3, so we have justified the first claim of the following Theorem.

Theorem 3.4. *There are algorithms to test whether a given level WFA $\mathcal{A}$ is continuous or strongly continuous.*

Proof. The effectiveness of testing continuity was shown above. By Corollary 2.6, to test strong continuity it is enough to test whether $f_{\mathcal{A}_I}(01^{\omega}) = f_{\mathcal{A}_I}(10^{\omega})$ for all unit coordinate vectors $I = (\ldots 0, 1, 0, \ldots)$. By Corollary 3.3 values $f_{\mathcal{A}_I}(10^{\omega})$ and $f_{\mathcal{A}_I}(01^{\omega})$ can be effectively computed. For the second value, we note that it is the value given to a dyadic point by the WFA that is obtained from $\mathcal{A}$ by swapping matrices W_0 and W_1. □

The following important application of the generated subspaces $\mathcal{L}$ and $\mathcal{R}$ allows us to minimize a given WFA. Here the minimality is with respect to the number of states, as usual. We call a WFA $\mathcal{A}$ *minimal* if there exist no *equivalent* WFA with a fewer number of states. The equivalence is interpreted here as the equivalence of the word function $F_{\mathcal{A}}$ computed by the automata. The WFA considered are not necessarily level automata.

Theorem 3.5. *WFA $\mathcal{A}$ is minimal if and only if the dimensions of $\mathcal{L}$ and $\mathcal{R}$ are equal to n, the number of states. One can effectively test minimality and find an equivalent minimal WFA for any given WFA.*

Proof. Let n be the number of states in WFA $\mathcal{A} = (Q, A, W, I, T)$. Suppose first that $\dim(\mathcal{L}) = \dim(\mathcal{R}) = n$. Then there are words $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ such that IW_{u_i}, $i = 1, \ldots n$, form a basis of $\mathbb{R}^{1\times n}$, and $W_{v_j}T$, $j = 1, \ldots n$, form a basis of $\mathbb{R}^{n\times 1}$.

Let L be the $n \times n$ matrix whose rows are the vectors IW_{u_i} and let R be the $n \times n$ matrix whose columns are the vectors $W_{v_j}T$. Both L and R have full rank n, so their product LR has rank n as well. Element (i,j) of the product LR is $IW_{u_i}W_{v_j}T = F_{\mathcal{A}}(u_iv_j)$.

Let $\mathcal{B} = (Q', A, W', I', T')$ be an m-state WFA such that $F_{\mathcal{B}} = F_{\mathcal{A}}$. Let L' and R' be the matrices whose rows and columns are vectors $I'W'_{u_i}$ and $W'_{v_j}T'$, respectively. These are matrices of sizes $n \times m$ and $m \times n$, respectively, so their ranks are at most m. Their product $L'R'$ is identical to LR because the element (i,j) is $I'W'_{u_i}W'_{v_j}T' = F_{\mathcal{B}}(u_iv_j)$. The rank of product $L'R'$ is at most m, so we must have $m \geqslant n$. Hence we have shown that $\mathcal{A}$ is a minimum state WFA.

Conversely, suppose that $\dim(\mathcal{L}) = m < n$. The case when $\dim(\mathcal{R}) < n$ is analogous. Construct an $m \times n$ matrix B whose rows constitute an orthonormal basis of $\mathcal{L}$. Then

$$IW_uB^TB = IW_u$$

for every $u \in A^*$. Consider now a new WFA with m states defined as:

- initial distribution vector $I' = IB^T$,
- transition matrices $W'_a = BW_aB^T$ for all $a \in A$, and
- final distribution vector $T' = BT$.

Then it easily follows that

$$I'W'_u = IW_uB^T$$

for every $u \in A^*$. Consequently,

$$I'W'_uT' = IW_uT\,,$$

which means that the new WFA with m states computes the same word function as $\mathcal{A}$. Hence $\mathcal{A}$ is not a minimal WFA.

Because orthogonal basis can be effectively constructed for $\mathcal{L}$ and $\mathcal{R}$, one can effectively test $\mathcal{A}$ for minimality and, in the case $\mathcal{A}$ is not minimal, one can effectively find a smaller, equivalent WFA. Hence the minimization is effective. □

Note, however, that minimizing a level automaton may create a WFA that is not a level automaton. But if we have a minimal level WFA then continuity and the strong continuity coincide:

Corollary 3.6. *A minimal level WFA $\mathcal{A}$ is strongly continuous if and only if $\hat{f}_{\mathcal{A}}$ is continuous.*

Proof. By Theorem 3.5, every initial distribution I is a linear combination of vectors IW_u for some $u \in A^*$. If $\hat{f}_{\mathcal{A}}$ is continuous then by Theorem 2.5 we have $f_{\mathcal{A}}(u01^\omega) = f_{\mathcal{A}}(u10^\omega)$ for all words $u \in A^*$. Then, for any vector I, we have $f_{\mathcal{A}_I}(01^\omega) = f_{\mathcal{A}_I}(10^\omega)$, so by Corollary 2.6 the WFA is strongly continuous. □

Assume that a real function $f : [0,1) \to \mathbb{R}$ is computed by a strongly continuous level automaton $\mathcal{A}$, that is $\hat{f}_{\mathcal{A}} = f$. Then a "naturally defined" subautomaton defines also continuous functions, where "naturally defined" refers to the process where some set of states and all high degree states leading to those are deleted. It is not difficult to see that the assumptions of strong continuity or minimality are needed here.

In the next result we analyze the reverse situation, namely how to extend a WFA computing a continuous function to a larger, and more complex, WFA which still would compute a continuous function. In this spirit we introduce in the next theorem a *continuity construction*. Let $\mathcal{A} = (Q, \{0,1\}, W, I, T)$, with $T = (1, 1, \dots, 1)$ and card $Q = n$, be a strongly continuous level automaton, and denote by $\mathcal{A}_q$ the automaton $\mathcal{A}$ with q as the initial state. Consequently, $\mathcal{A}_q$ computes a continuous function for each q in Q. For two numbers $\alpha, \beta \in [0, 1)$ we define an extension $\mathcal{A}(\alpha, \beta)$ of $\mathcal{A}$ as follows. Each of those is reduced, contains one new state r, which is the only initial one, and the following transitions

$$r \xrightarrow{0;\alpha} r, \quad r \xrightarrow{1;\beta} r$$

and

$$r \xrightarrow{i;W(i,q)} q, \quad \text{for } i \in \{0,1\} \text{ and } q \in Q\,.$$

Here $W(i, q)$'s are allowed to be arbitrary *real* numbers. It follows that automata in $\mathcal{A}(\alpha, \beta)$ are well defined level automata (up to the point that nonlooping weights might be negative), and moreover completely specified by $2n$-dimensional vector

$$\big(W(0,q), W(1,q)\big) \in \mathbb{R}^{2n}\,. \tag{3.2}$$

The next result characterizes when $\mathcal{A}_{\text{ext}}$ in $\mathcal{A}(\alpha, \beta)$ defines a continuous function, or in fact is strongly continuous.

Theorem 3.7. *Let $\mathcal{A}$ be a strongly continuous level automaton, and let $\alpha, \beta \in [0, 1)$. Let $\mathcal{A}_{ext}$ be in the above family $\mathcal{A}(\alpha, \beta)$, defined by choosing vector (3.2). Then $\mathcal{A}_{ext}$ computes a continuous real function if and only if it is strongly continuous. This happens if and only if one of the following conditions is satisfied*

(i) $f_{\mathcal{A}_q}(0^\omega) = f_{\mathcal{A}_q}(1^\omega) = 0$ *for all* $q \in Q$,

(ii) $\alpha + \beta = 1$ *and* $f_{\mathcal{A}_q}(0^\omega) = f_{\mathcal{A}_q}(1^\omega)$ *for all* $q \in Q$,

(iii)
$$\sum_{q\in Q} \lambda_q W(0,q) + \sum_{q \in Q} \mu_q W(1,q) = 0 \tag{3.3}$$

for some fixed numbers λ_q and μ_q such that at least one of those is not zero.

Proof. To begin with we note that the strong continuity and the continuity of $\mathcal{A}_{\text{ext}}$ on the interval $[0, 1)$ are both equivalent to

$$f_{\mathcal{A}_{\text{ext}}}(01^\omega) = f_{\mathcal{A}_{\text{ext}}}(10^\omega)\,. \tag{3.4}$$

Indeed, if $\mathcal{A}_{\text{ext}}$ is continuous then by Theorem 2.5, condition (3.4) must hold. On the other hand, by Corollary 2.6, condition (3.4) together with the strong continuity of $\mathcal{A}$ are enough to conclude that $\mathcal{A}_{\text{ext}}$ is strongly continuous. Potentially negative weights on connecting transitions is not an additional problem.

To make use of the above we compute

$$f_{\mathcal{A}_{\text{ext}}}(01^\omega) = \alpha\left(\frac{1}{1-\beta}\sum_{q\in Q} W(1,q) f_{\mathcal{A}_q}(1^\omega)\right) + \sum_{q\in Q} W(0,q) f_{\mathcal{A}_q}(1^\omega)\,.$$

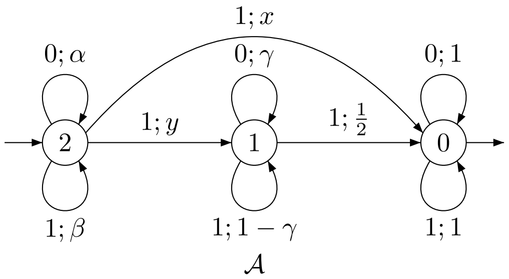


Figure 10. A level automaton of degree two.

Computing the similar formula for $f_{\mathcal{A}_{\text{ext}}}(10^\omega)$, and using (3.4) we derive (3.3) with

$$\lambda_q = f_{\mathcal{A}_q}(1^\omega) - \frac{\beta}{1-\alpha} f_{\mathcal{A}_q}(0^\omega)\,, \text{ and}$$
$$\mu_q = \frac{\alpha}{1-\beta} f_{\mathcal{A}_q}(1^\omega) - f_{\mathcal{A}_q}(0^\omega)\,.$$

If here both $\lambda_q = 0$ and $\mu_q = 0$, for all q, then necessarily $f_{\mathcal{A}_q}(0^\omega) = f_{\mathcal{A}_q}(1^\omega)$. If, moreover, $f_{\mathcal{A}_q}(0^\omega) \neq 0$ for some state q then $\alpha + \beta = 1$. So we are in the case (i) or (ii), where all other values of $W(i,q)$ can be arbitrary. Otherwise we are in the case (iii), where the vector (3.2) belongs to a fixed $2n-1$ dimensional hyperplane of $\mathbb{R}^{2n}$. □

Theorem 3.7 further confirms that WFA compute very rarely continuous functions. Even a more peculiar feature of such automata is analyzed in the next example.

Example 3.1. Consider the three state automaton $\mathcal{A}$ illustrated in Figure 10. Now, since $\gamma + (1-\gamma) = 1$, the automaton $\mathcal{A}_1$, that is $\mathcal{A}$ with only state 1 as initial one, computes a continuous function. Further the inequality $0 = f_{\mathcal{A}_1}(0^\omega) \neq f_{\mathcal{A}_1}(1^\omega)$ guarantees that there exists the unique hyperplane in $\mathbb{R}^2$, that is a line going through the origin, such that $\mathcal{A}$ defines a continuous function if and only if the pair (x, y) is on this line. In other words, $f_{\mathcal{A}}$ is continuous if and only if $\frac{x}{y}$ is fixed.

If we choose $\alpha = \beta = \frac{1}{4}$ and $\gamma = \frac{1}{2}$, the above ratio yields the value $\frac{1}{2}$. If we further fix $y = \frac{1}{2}$, then necessarily $x = \frac{1}{4}$ and we are in the automaton $\mathcal{A}_2$ of Figure 7. As we saw, this computes the parabola $f(x) = x^2$. Now, after fixing α, β and γ as above, the requirement that $\mathcal{A}$ computes a continuous function leads to the unique, up to the ratio of x and y, automaton $\mathcal{A}_2$. This indeed is the unique minimal level automaton for the function $f(x) = x^2$. This is justified as follows. Any such automaton has three states, as a consequence of Corollary 3.10 below. Moreover, by considerations in Section 4, necessarily $\alpha = \beta = \frac{1}{4}$ and $W(1,0) = W(0,0) = \frac{1}{2}$.

A striking phenomenon occurs when we try to decompose the above (essentially) unique level automaton computing the parabola $f(x) = x^2$. The only, at least automata theoretic, way of doing this is to compute $\hat{f}_{\mathcal{A}_2}$ as the sum of two functions

$$\hat{f}_{\mathcal{A}_2} = \hat{f}_{\mathcal{A}'_2} + \hat{f}_{\mathcal{A}'_1}\,,$$

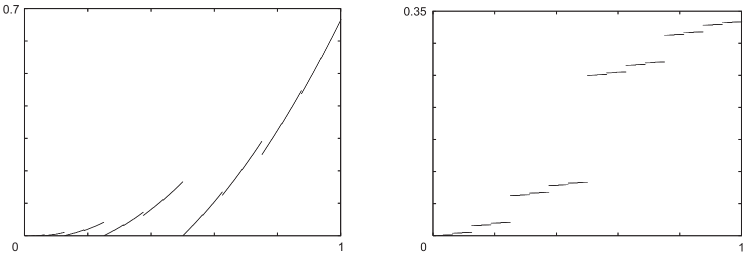


Figure 11. Functions computed by WFA's $\mathcal{A}_2'$ and $\mathcal{A}_1'$.

where $\mathcal{A}_2'$ is obtained from $\mathcal{A}_2$ by deleting the edge $2 \xrightarrow{1;\frac{1}{4}} 0$ and $\mathcal{A}_1'$ is obtained from $\mathcal{A}_2$ by deleting the state 1. It follows that while $\hat{f}_{\mathcal{A}_2}$ is very nicely behaving, both $f_{\mathcal{A}_2'}$ and $f_{\mathcal{A}_1}$ are complicated noncontinuous fractal type functions, as illustrated in Figure 11.

We have analyzed above when WFA, and in particular the level variants, define continuous functions. In the next results we require not only the continuity, but even more regularity. Namely, we characterize when a WFA defines a *smooth* function. We start with the following result from [14].

Theorem 3.8. *Let $\mathcal{A} = (Q, A, W, I, T)$ be a minimal (not necessarily level) WFA whose ω-word function $f_{\mathcal{A}}$ is everywhere defined and possesses continuous k first derivatives. If the k'th derivative is not zero, then W_0 has $k+1$ distinct eigenvalues $\frac{1}{2^i}$ for $i = 0, 1, \dots k$.*

Proof. Let $Q = \{q_1, q_2, \dots, q_n\}$. For any $q_i \in Q$ we denote by $\hat{f}_i$ the real function computed by the WFA obtained from $\mathcal{A}$ by making q_i the only initial state. The weight matrices W_0 and W_1 state a mutually recursive linear decomposition of functions $\hat{f}_i$:

$$\begin{pmatrix} \hat{f}_1(x) \\ \vdots \\ \hat{f}_n(x) \end{pmatrix} = \begin{cases} W_0 \begin{pmatrix} \hat{f}_1(2x) \\ \vdots \\ \hat{f}_n(2x) \end{pmatrix} & \text{for } x \in [0, \frac{1}{2}) \ , \\ W_1 \begin{pmatrix} \hat{f}_1(2x-1) \\ \vdots \\ \hat{f}_n(2x-1) \end{pmatrix} & \text{for } x \in [\frac{1}{2}, 1) \ . \end{cases} \tag{3.5}$$

It follows from the characterization of minimal WFA in Theorem 3.5 that each $\hat{f}_i$ is a linear composition of translated and dilated versions of restrictions of $\hat{f}_{\mathcal{A}}$ on some dyadic subsegments. Hence each $\hat{f}_i$ is everywhere defined and has k continuous derivatives. By differentiating k times the first equality in (3.5) we obtain

$$\begin{pmatrix} \hat{f}_1^{(k)}(x) \\ \vdots \\ \hat{f}_n^{(k)}(x) \end{pmatrix} = 2^k W_0 \begin{pmatrix} \hat{f}_1^{(k)}(2x) \\ \vdots \\ \hat{f}_n^{(k)}(2x) \end{pmatrix} \qquad \text{for } x \in [0, \tfrac{1}{2}) \, .$$

Substituting $x = 0$ gives

$$Z = 2^k W_0 Z, \tag{3.6}$$

where

$$Z = \begin{pmatrix} \hat{f}_1^{(k)}(0) \\ \vdots \\ \hat{f}_n^{(k)}(0) \end{pmatrix}.$$

If Z is the zero vector then $\hat{f}_i^{(k)}(0) = 0$ for $i = 1, 2, \ldots, n$. This implies that $\hat{f}_{\mathcal{A}}^{(k)}(x) = 0$ for every dyadic rational x. Dyadic rationals are dense in $[0, 1)$, so the k'th derivative of $\hat{f}_{\mathcal{A}}$ is zero, contradicting the assumption in the theorem statement.

We conclude that Z is an eigenvector of W_0 corresponding to eigenvalue $\frac{1}{2^k}$. This reasoning can be applied with any $i \leqslant k$ in place of k. □

Since an $n \times n$ matrix can have at most n eigenvalues we have

Corollary 3.9. *Each WFA computing a real function with non-vanishing k'th derivative contains at least $k + 1$ states.*

If an n-state WFA computes a smooth function then, by the corollary, the n'th derivative of the function must be zero, so the function must be a polynomial of degree at most $n - 1$.

Corollary 3.10. *Any smooth function computed by a WFA is a polynomial. A WFA that computes a polynomial of degree k must have at least $k + 1$ states.*

We conclude this section with decidability questions. An important basic result is the decidability of the equivalence.

Theorem 3.11. *For two given level automata $\mathcal{A}$ and $\mathcal{A}'$ it is decidable whether or not*

$$f_{\mathcal{A}} \overset{?}{=} f_{\mathcal{A}'}\,, \tag{3.7}$$

and, respectively

$$\hat{f}_{\mathcal{A}} \overset{?}{=} \hat{f}_{\mathcal{A}'}\,.$$

Proof. By definitions of the considered functions the questions are equivalent from the decidability point of view.

The decidability of the first question follows from our normal forms and the corresponding result of rational series. Indeed, we first transform $\mathcal{A}$ and $\mathcal{A}'$ into equivalent 0-faithful automata, $\mathcal{A}_0$ and $\mathcal{A}_0'$, respectively. By Theorem 3.2 this can be effectively done. Then the question (3.7) becomes the question

$$F_{\mathcal{A}} \overset{?}{=} F_{\mathcal{A}'}\,,$$

which is just a special case of the equivalence problem for finite automata with multiplicities, cf. e.g., [15]. □

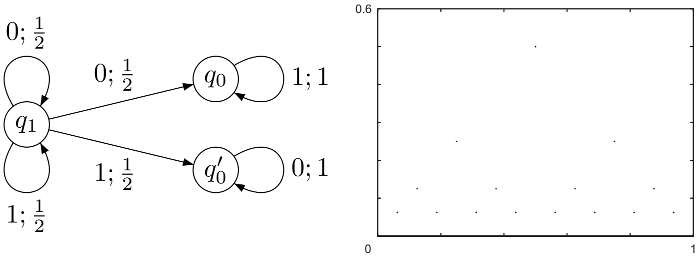


Figure 12. A 3-state automaton and the function it computes.

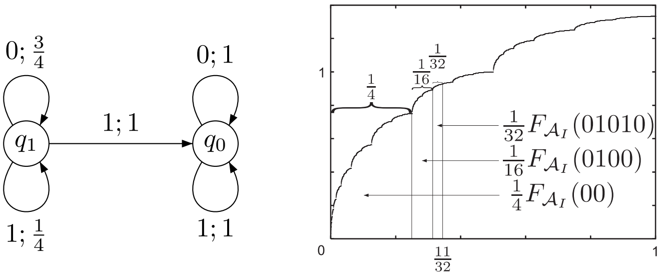


Figure 13. A 2-state automaton and the continuous fractal type function it computes. The additional information presented here will be explained in Section 5.

4 Examples

In this section we give several examples showing what can, and also what cannot be done, by our model of WFA. Several such examples are already shown, and a complicated monster type of automaton is given in Section 6. In Section 5 theoretical foundations of these examples, and more generally, image and function manipulation is created.

By this point it should be clear that WFA are powerful tools to compute efficiently many fractal type of functions. We will have two more examples, one computing a noncontinuous and the other continuous such function.

Example 4.1. The 3-state automaton shown in Figure 12 computes the fractal type function which is noncontinuous at any dyadic rational point.

Example 4.2. The 2-state automaton shown in Figure 13 computes, by Example 2.2, a continuous fractal type of function which does not have the right derivative at a countably infinite set of points, including the point 0.

In the next example we combine several automata to create a particular shadowing effect to a picture.

Example 4.3. Let us start from the fourth automaton $\mathcal{A} = (Q, A, W, I, T)$ of Figure 4 in Section 2 computing Sierpinski's triangle. Note that the automaton is not strictly speaking a level automaton, since there are loops in states of positive degree with weights equal to

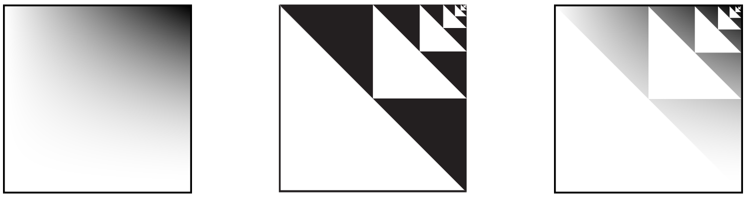


Figure 14. A shadowed square, Sierpinski's triangle and as their combination a shadowed variant of Sierpinski's triangle.

1. However, the convergence at any point is guaranteed due to the determinism of the underlying automaton.

Our goal is to create from this a picture which is horizontally linearly darkening and vertically quadratically darkening, that is a picture shown in Figure 14 on the right. We make use of automata $\mathcal{A}_1 = (Q_1, A, W_1, I_1, T_1)$ and $\mathcal{A}_2 = (Q_2, A, W_2, I_2, T_2)$ computing the linear function $f(x) = x$ and the parabola $f(y) = y^2$, respectively. By the construction analyzed in Section 5, we find an automaton

$$\mathcal{A}_1 \times_c \mathcal{A}_2 = (Q_1 \times Q_2, A \times A, W_1 \times W_2, I_1 \times I_2, T_1 \times T_2)\,,$$

where

$$W_1 \times W_2 \;:\; (p_1, p_2) \xrightarrow{(2j+i);W_1(p_1,i,q_1)W_2(p_2,j,q_2)} (q_1, q_2)\,,$$

which computes the shadowed square shown in Figure 14. In the square the picture turns linearly (resp. quadratically) darker, when moving horizontally (resp. vertically). In other words, the automaton computes the two unknown function

$$f(x, y) = xy^2 \quad \text{for } x, y \in [0, 1]\,.$$

Now, finally the right side picture of Figure 14 is obtained, again according to constructions of the next section, by computing the direct product of automata $\mathcal{A}$ and $\mathcal{A}_1 \times_c \mathcal{A}_2$, that is by the automaton

$$\mathcal{B} = \mathcal{A} \times (\mathcal{A}_1 \times_c \mathcal{A}_2)\,.$$

Note that the automaton $\mathcal{A}_1 \times_c \mathcal{A}_2$ has 6 states, and the automaton $\mathcal{B}$ has 24 states.

In what follows we consider how polynomials can be computed by level automata. We already saw in the previous section that any automaton computing a polynomial of degree n must have at at least $n+1$ states. The following example shows that $n+1$ states are enough, even among level automata. The level automaton constructed has one state of each degree, and the loops on states of degree j are of the form

$$q_j \xrightarrow{i,\frac{1}{2^j}} q_j\,.$$

Example 4.4. Consider the automaton $\mathcal{A}_n$ shown in Figure 15. Clearly, for $n = 1$ and $n = 2$, we have the automata $\mathcal{A}_1$ and $\mathcal{A}_2$ of Figure 7. One can straightforwardly show, e.g., by induction on n, that the automaton $\mathcal{A}_n$ with $I = (1, 0, \ldots, 0)$ and $T = (0, \ldots, 0, 1)$ computes the nth power: $f(x) = x^n$. Hence, any polynomial, even when

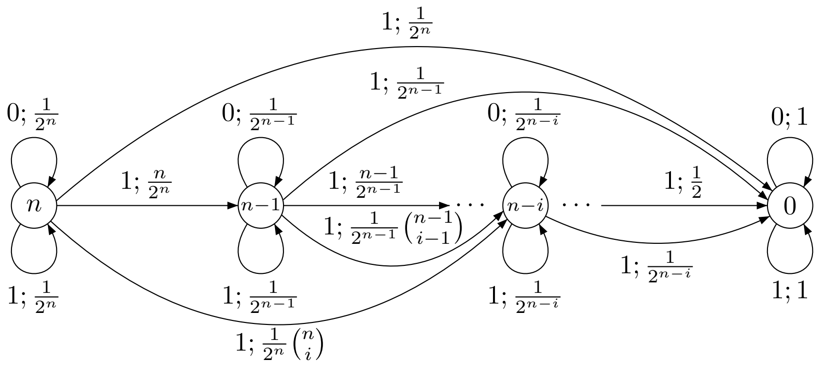


Figure 15. The automaton $\mathcal{A}_n$ computing the function $f(x) = x^n$, or, in fact, any polynomial of degree n with a suitable initial distribution.

negative coefficients are allowed, can be computed by the above fixed automaton having the vector of coefficients of the polynomial as initial distribution.

By the argument we used in Example 4.3 the above can be extended to polynomials with several variables.

Example 4.5. Each polynomial $P(x_1, \dots, x_n)$ with real coefficients can be computed by a fixed level automaton with a suitable initial distribution. The number of states of such an automaton is

$$\prod_{i=1}^{n} (d_i + 1)\,,$$

where d_i is the maximal power of x_i occurring in P.

We conclude this section with an observation that automatic real functions are not closed under the composition of functions – another property deviating our theory from those in classical analysis.

Example 4.6. Consider the functions $f_1, f_2 : [0, 1) \to \mathbb{R}$ defined by

$$f_1(x) = \begin{cases} 0 & \text{if } 0 \leqslant x < \frac{1}{2}, \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

and

$$f_2(x) = \frac{3}{4}x \quad \text{for } 0 \leqslant x < 1\,.$$

Clearly, both of these are real automatic functions, and $f_1 \circ f_2 : [0, 1) \to \mathbb{R}$ is well defined:

$$f_1 \circ f_2(x) = \begin{cases} 0 & \text{if } 0 \leqslant x < \frac{2}{3}, \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

Hence, $f_1 \circ f_2$ is not continuous at point $\frac{2}{3} = 0,1010\dots$ But $\frac{2}{3}$ is not a dyadic rational, so that, by Corollary 2.4, the composition $f_1 \circ f_2$ is not real automatic.

The above considerations rely on the fact that we use binary representation for numbers. Indeed, in ternary notations $\frac{2}{3}$ is a dyadic rational and the above conclusion would not hold. However, if we take $f_2'(x) = x^2$, we could obtain the composition

$$f_1 \circ f_2'(x) = \begin{cases} 0 & \text{if } 0 \leqslant x < 2\sqrt{2}, \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

for which our above arguments would work independently of the arity of the number representation.

It is not difficult to extend the above nonclosure property for continuous real automatic functions, see [17].

5 Image manipulation

This section intends to further explain what is achievable in our theory of WFA, and even more why these are doable. In other words, we explain what some basic properties of automata theory, in particular closure properties, tell if translated into images represented by weighted finite automata. A major goal is to show what kind of operations on images can be done directly on the level of their compressed automata representation. We believe that this indeed is the main strong point of the theory.

We continue with a few simple observations. Here it is more natural to think pictures as real two dimensional pictures rather than functions.

Example 5.1 (Zooming). A straightforward operation which is easy to implement is to zoom a given part of the picture. Indeed, what is needed here is to compute the distribution IW_u of the address u of the subpicture and set this to be the initial distribution of the automaton.

Example 5.2 (Complementation). To make a negative of a picture is even simpler: change the initial distribution to its negative. Similarly, a paling or darkening of a picture can be achieved by multiplying the initial distribution by a constant.

Example 5.3 (Combining pictures). This again is a trivial task since our automata are allowed to be nondeterministic: a natural way of defining a union of WFA can be used. Note also that this combining of pictures can be done without overlaps or by allowing overlaps. A special case is the one where two pictures are put on top of each other. This can be used also as an operation of creating a background to a given picture.

After the preliminary remarks we focus on different types of products of automata and their relations to operations on pictures. At least the following types of products of automata are natural: Hadamard product (or direct product), complete direct product and Cauchy product. We deal each of those separately. For the clarity of the presentation we do not pay attention to convergency considerations – for our models of level automata this is indeed no problem.

Let $\mathcal{A}_i = (Q_i, A, W_i, I_i, T_i)$, for $i = 1, 2$, be two weighted finite automata over the same alphabet A. Their *cross product* or *Hadamard product*, $\mathcal{A}_1 \times \mathcal{A}_2 = (Q, A, W, I, T)$ is defined by conditions

$$\begin{array}{rcl} Q & = & Q_1 \times Q_2\,, \\ I(p_1, p_2) & = & I_1(p_1) I_2(p_2)\,, \\ T(p_1, p_2) & = & T_1(p_1) T_2(p_2) \end{array}$$

and

$$W\big((p_1, p_2), a, (q_1, q_2)\big) = W_1(p_1, a, q_1) \cdot W_2(p_2, a, q_2)\,,$$

for $p_i, q_i \in Q_i$ and $a \in A$. Note that the cross product of average preserving WFA is not necessarily average preserving. In the case of finite automata this construction recognizes the intersection of the component languages, in the connection of rational formal power series their componentwise (or Hadamard) product. Consequently, for WFA and images we have

Theorem 5.1. *For weighted finite automata $\mathcal{A}_1$ and $\mathcal{A}_2$ we have*

$$f_{\mathcal{A}_1 \times \mathcal{A}_2}(w) = f_{\mathcal{A}_1}(w) \cdot f_{\mathcal{A}_2}(w) \quad \textit{for each } w \in A^\omega\,.$$

That is the cross product of two WFA define the pointwise product of the original images.

In above $f_{\mathcal{A}_1}$, $f_{\mathcal{A}_2}$ and $f_{\mathcal{A}_1 \times \mathcal{A}_2}$ were all defined over the same alphabet. We can, however, define also a more general cross product of finite automata where also the input symbols are paired. The *complete cross product* of automata $\mathcal{A}_i = (Q_i, A_i, W_i, I_i, T_i)$, for $i = 1, 2$, is the automaton

$$\mathcal{A}_1 \times_c \mathcal{A}_2 = (Q_1 \times Q_2, A_1 \times A_2, W, I, T)\,,$$

where I and T are as in the cross product above, and

$$W\big((p_1, p_2), (a_1, a_2), (q_1, q_2)\big) = W_1(p_1, a_1, q_1) \cdot W_2(p_2, a_2, q_2)\,,$$

for $p_i, q_i \in Q_i$ and $a_i \in A_i$.

Interestingly, this seldomly used product of automata is very useful in image manipulation, as observed in [8]. It follows straightforwardly:

Theorem 5.2. *For weighted finite automata $\mathcal{A}_1$ and $\mathcal{A}_2$ we have*

$$f_{\mathcal{A}_1 \times_c \mathcal{A}_2}(w) = f_{\mathcal{A}_1}(w_1) \cdot f_{\mathcal{A}_2}(w_2)\,,$$

where w_i, for $i = 1, 2$, is the projection of w into A_i^ω.

In the case of images, the above becomes more intuitive as follows. Assume that $A_1 = A_2 = \{0, 1\}$, that is both $\hat{f}_{\mathcal{A}_1}$ and $\hat{f}_{\mathcal{A}}$ are one unknown real functions. Now, as earlier, we rename $A = A_1 \times A_2 = \{(0,0), (0,1), (1,0), (1,1)\}$ as $(0,0) \leftrightarrow 0$, $(0,1) \leftrightarrow 1$, $(1,0) \leftrightarrow 2$, $(1,1) \leftrightarrow 3$. Further let $\pi_i : A^* \to A_i^*$ be the natural projection. Then each $w \in A^n$, via the pair $\big(\pi_1(w), \pi_2(w)\big)$ determines uniquely a pixel in $2^n \times 2^n$ -resolution of the unit interval. And as earlier when $n \to \infty$ the correspondence between $w \in A^\omega$ and $\big(\widehat{\pi_1(w)}, \widehat{\pi_2(w)}\big)$ becomes one-to-one if we exclude those w's for which $\pi_1(w) \in A^*1^\omega$ or $\pi_2(w) \in A^*1^\omega$. Consequently, $\mathcal{A}_1 \times_c \mathcal{A}_2$ can be viewed as an automaton computing a real function : $[0,1)^2 \to \mathbb{R}$, and Theorem 5.2 with the above notation can be rephrased as:

Corollary 5.3. *We have*

$$\hat{f}_{\mathcal{A}_1 \times_c \mathcal{A}_2}\big(\widehat{\pi_1(w)}, \widehat{\pi_2(w)}\big) = \hat{f}_{\mathcal{A}_1}\big(\widehat{\pi_1(w)}\big)\hat{f}_{\mathcal{A}_2}\big(\widehat{\pi_2(w)}\big)\,.$$

Our previous results justify the correctness of Examples 4.3 and 4.5 in Section 4.

Next we turn to so-called Cauhcy product which is defined for formal power series $s : A^* \to \mathbb{R}$. For two such series s and s' their Cauchy product $s \cdot_C s'$ is defined by

$$s \cdot_C s'(w) = \sum_{w=uv} s(u) \cdot s'(v)\,.$$

Since in A^* each word has only finitely many different factorizations the above is well defined. It is well known that if the above series are rational, that is defined by finite automata with multiplicities, so is their Cauchy product, see e.g., [22].

In our presentation the functions $F_{\mathcal{A}} : A^* \to \mathbb{R}$ are rational formal power series, so that for two such functions, say $F_{\mathcal{A}}$ and $F_{\mathcal{B}}$, we can find a weighted finite automaton $\mathcal{C}$ such that

$$F_{\mathcal{C}}(w) = F_{\mathcal{A}}(w) \cdot_C F_{\mathcal{B}}(w)\,.$$

However, our main interest is in computations on infinite words, that is in the cases when the above sum is not finite. What can be done in this setting is analyzed in [17]. What we intend to do here is to look at a special case which is very much in the spirit of computing the above product, namely how to integrate a given real automatic function.

Theorem 5.4. *For each level automaton $\mathcal{A}$ of degree n there exists a level automaton $\mathcal{A}_I$ of degree $n+1$ such that*

$$\hat{f}_{\mathcal{A}_I}(x) = \int_0^x \hat{f}_{\mathcal{A}}(t)\,dt\,.$$

Proof. (Outline) We give here the main construction, and refer to [8] for details. We start from a level automaton of degree n computing $\hat{f}_{\mathcal{A}} : [0,1) \to \mathbb{R}$. First, by Theorem 3.2 we search for an equivalent average preserving level automaton $\mathcal{A}_{\text{ap}}$ of the same degree as $\mathcal{A}$. Next we add to that one more state, say t, and modify the transitions as follows:

(i) t contains loops of weight 1;

(ii) weights of old transitions of $\mathcal{A}_{\text{ap}}$ are divided by 2;

(iii) from each old state q of $\mathcal{A}_{\text{ap}}$ there is a transition to t labelled with 1 and with the weight $\frac{1}{2}\sum_{p\in Q} W(p,0,q)$.

Let $\mathcal{A}_I$ be the level automaton of degree $n+1$ thus defined. Now, the proof of Theorem 5.4 is a consequence of the following two claims.

Claim I. For each $w = a_1a_2\ldots$ with $a_i \in \{0,1\}$ and $w \notin A^*1^\omega$ we have

$$f_{\mathcal{A}_I}(w) = \sum_{i=1}^{\infty} a_i 2^{-i} F_{\text{ap}}(\text{pref}_{i-1}(w) \cdot 0)\,.$$

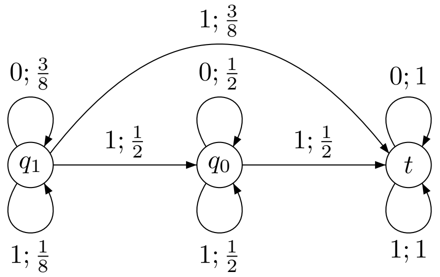


Figure 16. An integrated automaton $\mathcal{A}_I$.

Claim II. For each $w = a_1a_2\dots$ with $a_i \in \{0,1\}$ and $w \notin A^*1^\omega$ we have

$$\int_0^x \hat{f}_{\mathcal{A}_{\text{ap}}}(t)\,dt = \sum_{i=1}^{\infty} a_i 2^{-i} F_{\text{ap}}(\text{pref}_{i-1}(w)\cdot 0)\,,$$

where $x = \text{bin}(w)$.

Claim I follows straightforwardly, while the Claim II is strongly based on the average preserving property of $\mathcal{A}_{\text{ap}}$. □

Few remarks of the above theorem are in order. First it, together with Example 5.1, proves that the automaton of Figure 15 computes the function $f(x) = x^n$. Second, the next example explains more of the construction.

Example 5.4. Consider the automaton $\mathcal{A}$ of Figure 13. As we noted (without a proof) it is not at all smooth. On the other hand it is average preserving so that we can directly apply the construction of the proof of the previous theorem. We obtain the automaton $\mathcal{A}_I$ in Figure 16. We claim that

$$\hat{f}_{\mathcal{A}_I}\left(\frac{11}{32}\right) = \frac{1}{4}F_{\mathcal{A}}(00) + \frac{1}{16}F(0100) + \frac{1}{32}F_{\mathcal{B}}(01010)\,,$$

which shows that $\int_0^{\frac{11}{32}} \hat{f}_{\mathcal{A}}(t)\,dt = \hat{f}_{\mathcal{A}_I}\left(\frac{11}{32}\right)$. We have the verification of this to the reader, cf. the illustration of Figure 13.

A feature of WFA representations of functions is that one can perform many useful operations using finite state transducers [10]. The transformations can be done directly on the WFA representation. A *weighted finite transducer* (WFT) from alphabet A to alphabet B is a 6-tuple $\mathcal{T} = (Q, A, B, W, I, T)$ where the state set Q, the initial distribution I and the final distribution T are as in WFA. The weight function W is a mapping

$$W : Q \times (A \cup \{\varepsilon\}) \times (B \cup \{\varepsilon\}) \times Q \longrightarrow \mathbb{R}\,.$$

The weight function W is conveniently viewed as a collection of real matrices $W_{a,b}$, indexed by $a \in A\cup\{\varepsilon\}$ and $b \in B\cup\{\varepsilon\}$. The element (q,p) of matrix $W_{a,b}$ is $W(q,a,b,p)$, and we call it the weight of the transition from q to p with input a and output b. The transducer is called ε*-free* if the weight matrices $W_{a,\varepsilon}$, $W_{\varepsilon,b}$ and $W_{\varepsilon,\varepsilon}$ are zero matrices for all $a \in A, b \in B$.

WFT $\mathcal{T}$ defines a (partial) *weighted relation*

$$\rho : A^* \times B^* \longrightarrow \mathbb{R}$$

as follows: For every $u \in A^*$, $v \in B^*$ we have

$$\rho(u, v) = IW_{u,v}T\,,$$

where

$$W_{u,v} = \sum_{\substack{a_1 \ldots a_m = u \\ b_1 \ldots b_m = v}} W_{a_1,b_1} \ldots W_{a_m,b_m}\,,$$

if the sum converges. The sum is over all decompositions of u and v into symbols $a_i \in A \cup \{\varepsilon\}$ and $b_i \in B \cup \{\varepsilon\}$. The sum is finite (and hence converges) if the WFT does not contain any cycles whose input and output words are both empty. If the WFT is ε-free then clearly

$$\rho(a_1 \ldots a_k, b_1 \ldots b_k) = IW_{a_1,b_1} \ldots W_{a_k,b_k}T\,,$$

where all $a_i \in A$ and $b_i \in B$, and $\rho(u, v) = 0$ when $|u| \neq |v|$.

Next we define the action of a weighted relation ρ on word functions $F : A^* \longrightarrow \mathbb{R}$. The result is a function $G : B^* \longrightarrow \mathbb{R}$, defined by

$$G(w) = \sum_{u \in A^*} f(u)\rho(u, w), \text{ for all } w \in B^*,$$

if the sum converges. We denote $G = \rho(F)$. The sum is finite if the WFT is ε-free or, more generally, if the weight matrices $W_{a,\varepsilon}$ are zero for all $a \in A \cup \{\varepsilon\}$. In this case the sum is over all words u whose length is not greater than the length of w.

It is easy to see that

$$\rho(r_1F_1 + r_2F_2) = r_1\rho(F_1) + r_2\rho(F_2)$$

for all word functions F_1 and F_1 and $r_1, r_2 \in \mathbb{R}$, so the operation ρ is linear. The following two examples show some natural linear image transformations that can be implemented as a WFT over the alphabets $A = B = \{0, 1, 2, 3\}$. See [10] for more examples.

Example 5.5. Let $w = a_1a_2 \ldots a_n$ be a fixed word of length n. Consider the two WFT's in Figure 17. The first transducer computes the following weighted relation: $\rho(u, wu) = 1$ for every $u \in A^*$, and $\rho(u, v) = 0$ if $v \neq wu$. The effect is to insert the input image, appropriately scaled, into the sub-square addressed by w.

By swapping the input and output symbols we obtain the second WFT in Figure 17. It assigns $\rho(wu, u) = 1$ for every $u \in A^*$, and $\rho(v, u) = 0$ if $v \neq wu$. The corresponding action on images is to crop and zoom the sub-square addressed by w from the input image.

Example 5.6. The WFT in Figure 18 rotates the image 90° clockwise. This is established by a simple permutation $0 \mapsto 2 \mapsto 3 \mapsto 1 \mapsto 0$ of the alphabet.

An application of an ε-free WFT $\mathcal{T} = (Q, A, B, W, I, T)$ on a WFA $\mathcal{A} = (P, A, W', I', T')$ is defined to be the WFA $\mathcal{T}(\mathcal{A}) = (P \times Q, B, W'', I'', T'')$ where $I''(p, q) =$

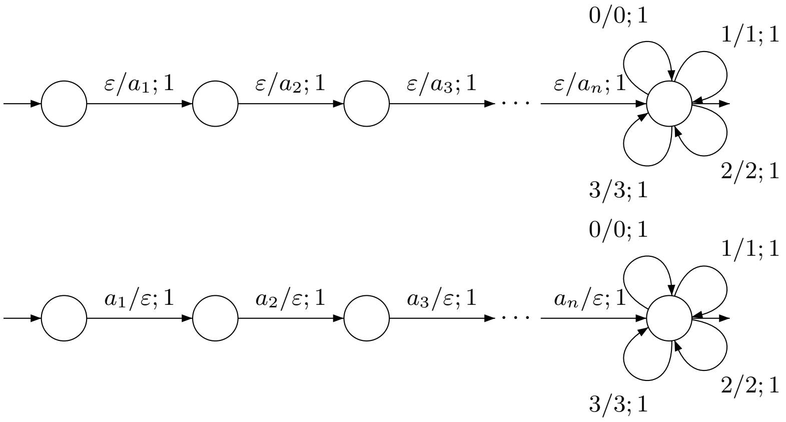


Figure 17. The transducers in Example 5.5.

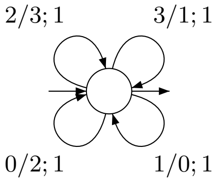


Figure 18. A transducer that defines a 90° rotation clockwise.

$I'(p)I(q)$ and $T''(p,q) = T'(p)T(q)$ for all $p \in P$ and $q \in Q$, and

$$W''((p,q),b,(s,t)) = \sum_{a\in A} W'(p,a,s)W(q,a,b,t)$$

for all $p, s \in P$, $q, t \in Q$ and $b \in B$. This definition is a straightforward generalization of the usual application of a non-weighted transducer on a finite automaton. It is easy to see that

$$F_{\mathcal{T}(\mathcal{A})} = \rho(F_{\mathcal{A}}),$$

where ρ is the weighted relation determined by $\mathcal{T}$. In other words, WFT transformations of WFA word functions can be conveniently done directly in the automaton, the resulting word function is again defined by a WFA.

We finish the section with an example showing how the integral function, discussed in Theorem 5.4, can be defined using a WFT.

Example 5.7. Figure 19 shows a WFT whose input and output alphabet is $A = \{0,1\}$. For any $u = a_1a_2\dots a_n$, where each $a_i \in A$, let us denote by $\hat{u}$ the number whose binary expansion is $0.a_1a_2\dots a_n$, that is,

$$\hat{u} = \sum_{i=1}^{n} a_i 2^{-i}.$$

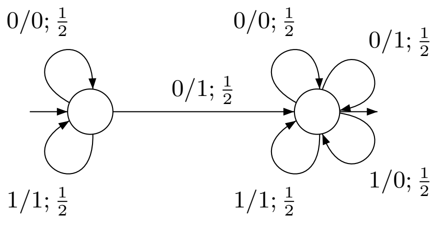


Figure 19. A transducer that computes the integral.

The transducer defines the following relation:

$$\rho(u,v) = \frac{1}{2^n}, \text{ if } u,v \in A^n \text{ and } \hat{u} < \hat{v}\,,$$

and $\rho(u,v) = 0$ otherwise. For a function $F : A^* \longrightarrow \mathbb{R}$ we have $\rho(F) = G$, where for each $v \in A^n$

$$G(v) = \frac{1}{2^n} \sum_{u \in A^n,\ \hat{u} < \hat{v}} F(u)\,.$$

Let $\mathcal{A}$ be an average preserving level automaton that computes the word function $F = F_{\mathcal{A}}$, and let $G = \rho(F) = F_{\rho(\mathcal{A})}$. It follows easily from the average preserving property of $\mathcal{A}$ that for any $w = a_1 a_2 \ldots a_n$, with $a_i \in A$, we have

$$G(a_1 a_2 \ldots a_n) = \sum_{i=1}^{n} \frac{a_i}{2^i} F(a_1 \ldots a_{i-1} 0)\,.$$

Analogously to the Claim II in the proof of Theorem 5.4 we then have that

$$\hat{f}_{\rho(\mathcal{A})}(x) = \int_0^x \hat{f}(t)\, dt\,,$$

for all $x \in [0,1)$.

6 A monster function

The goal of this section is to further emphasize how simple WFA can define complicated functions – at least in the spirit of classical analysis. More precisely we introduce a four state level automaton which defines a continuous but nowhere derivable real function. The material of this section comes from [13] and master's thesis of T. Sallinen.

We consider the automaton $\mathcal{A}(t)$ of Figure 20. We first note that $\mathcal{A}(t)$ is strongly continuous. Indeed subautomaton constituting of states 0 and 1 (or 0 and 1′) defines a continuous function by Example 2.2. And the whole automaton does the same by symmetry. Our second observation is that the subautomaton, say $\mathcal{A}_1$, constituting of states 0, 1 and 2 is strongly continuous, and hence by Corollary 2.6, the continuity of $\hat{f}_{\mathcal{A}_1}$ is

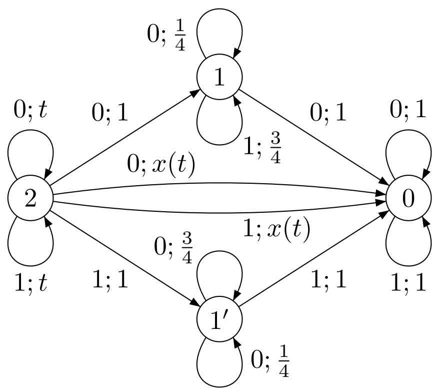


Figure 20. A four state level automaton $\mathcal{A}(t)$.

equivalent to the condition

$$f_{\mathcal{A}_1}(10^\omega) = f_{\mathcal{A}_1}(01^\omega)\,.$$

Simple calculations show that this is equivalent to

$$t\left(\frac{1}{1-t}x(t) + \frac{1}{1-t}\cdot 1\cdot \frac{4}{3}\right) = x(t)\,,$$

and further to

$$x(t) = \frac{4t}{2t-1}, \quad \text{provided } t \neq \frac{1}{2}\,. \tag{6.1}$$

As a conclusion, for $t = \frac{1}{2}$, $\mathcal{A}_1$ never defines a continuous function but for other values of $t \in (0,1)$, there exists the unique value $x(t)$ which makes $\hat{f}_{\mathcal{A}_1}$ continuous. Of course, this value can be also negative, so that the automaton is not strictly within the class of our level automata, but since $x(t)$ is a weight of a noncycle this is just a technicality. By symmetry, the subautomaton constituting of states 0, $1'$ and 2 leads to the same condition (6.1).

From now on we fix $\mathcal{A}(t)$, for $t \in (0,1)$, $t \neq \frac{1}{2}$, to mean the automaton of Figure 20, where $x(t)$ is fixed by (6.1). The functions $\hat{f}_{\mathcal{A}(t)}(x)$ for $t = \frac{1}{4}$, $\frac{2}{3}$ and $\frac{3}{4}$ are illustrated in Figure 21. While functions $\hat{f}_{\mathcal{A}(1/4)}(x)$ and $\hat{f}_{\mathcal{A}(3/4)}(x)$ are already quite complicated looking fractaltype functions, $\hat{f}_{\mathcal{A}(2/3)}(x)$ seems to be even more chaotic, likewise all functions $\hat{f}_{\mathcal{A}(t)}$ with $t \neq \frac{1}{4}, \frac{3}{4}$. Formally, this is shown in the following results.

Theorem 6.1. *Functions $\hat{f}_{\mathcal{A}(1/4)}$ and $\hat{f}_{\mathcal{A}(3/4)}$ are continuous, but they possess a derivative equal to zero at dyadic rational points.*

Theorem 6.2. *The function $\hat{f}_{\mathcal{A}(t)}$, for $t \in (0,1)$, $t \neq \frac{1}{4}, \frac{3}{4}$, is continuous, but does not have a derivative at any point.*

Proof. We outline the proof of Theorem 6.2 – that of Theorem 6.1 is implicitly in these considerations.

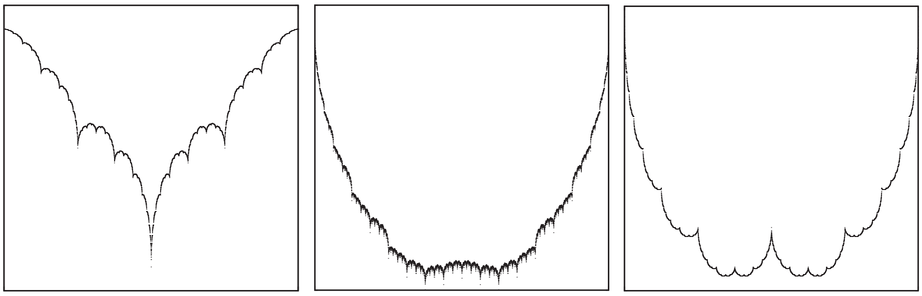


Figure 21. Functions computed by the automaton $\mathcal{A}(t)$ of Figure 20 for values $t=\frac{1}{4}$, $t=\frac{2}{3}$ and $t=\frac{3}{4}$, respectively. The middle one is an example of a continuous nowhere differentiable function.

Let $t \neq \frac{1}{4}, \frac{3}{4}$ and $x(t)$ be fixed. We choose an arbitrary $w \in \{0,1\}^\omega \setminus \{0,1\}^*1^\omega$ and show that, for some constant c, there exist infinitely many words $w_i \in \{0,1\}^\omega \setminus \{0,1\}^*1^\omega$ such that

$$|f_{\mathcal{A}(t)}(w) - f_{\mathcal{A}(t)}(w_i)| \geqslant c|\hat{w} - \hat{w}_i| \,. \tag{6.2}$$

This indeed proves the theorem.

Let $w_n = \mathrm{pref}_n(w)$ and define the following four sequences of words:

$$w_0(n) = w_n 0^\omega \,, \qquad w_1(n) = w_n 10^\omega \,,$$
$$w_2(n) = w_n 110^\omega \,, \qquad w_3(n) = w_n 1^\omega \,.$$

Actually, the word $w_3(n)$ is illegal but we can use it since the function is continuous. It follows that

$$|\hat{w} - \widehat{w_i(n)}| \leqslant \frac{1}{2^n} \quad \text{for } i = 0,1,2,3 \,.$$

The weight matrices associated to $\mathcal{A}(t)$ are

$$M_0 = \begin{pmatrix} t & 1 & 0 & x(t) \\ 0 & \frac{1}{4} & 0 & 1 \\ 0 & 0 & \frac{3}{4} & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } M_1 = \begin{pmatrix} t & 0 & 1 & x(t) \\ 0 & 0 & \frac{1}{4} & 1 \\ 0 & \frac{3}{4} & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and let us denote $(\alpha_n, \beta_n, \gamma_n, \delta_n)$ the distribution given by the prefix w_n, that is

$$(\alpha_n, \beta_n, \gamma_n, \delta_n) = (1,0,0,0)M_{w_n} \,.$$

In particular, $\alpha_n = t^n$ for all n. We need the limits $\lim_{n\to\infty} M_0^n$ and $\lim_{n\to\infty} M_1^n$. Straightforward computations show that these exist, and moreover

$$\lim_{n\to\infty} M_0^n = \begin{pmatrix} 0 & 0 & 0 & r(t) \\ 0 & 0 & 0 & \frac{4}{3} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } \lim_{n\to\infty} M_1^n = \begin{pmatrix} 0 & 0 & 0 & r(t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{4}{3} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where

$$r(t) = \frac{x(t)}{1-t} - \frac{4}{3(t-1)}\,.$$

With the help of above we can compute

$$\begin{aligned}
f_{\mathcal{A}(t)}\big(w_0(n)\big) &= r(t)\cdot\alpha_n + \frac{4}{3}\beta_n + \delta_n\,,\\
f_{\mathcal{A}(t)}\big(w_1(n)\big) &= \big(t\cdot r(t) + x(t)\big)\alpha_n + \beta_n + \gamma_n + \delta_n\,,\\
f_{\mathcal{A}(t)}\big(w_2(n)\big) &= \big(t^2\cdot r(t) + t\cdot x(t) + x(t) + 1\big)\alpha_n + \frac{3}{4}\beta_n + \frac{5}{4}\gamma_n + \delta_n\,,\\
f_{\mathcal{A}(t)}\big(w_3(n)\big) &= r(t)\cdot\alpha_n + \frac{4}{3}\gamma_n + \delta_n\,,
\end{aligned}$$

and further conclude that

$$f_{\mathcal{A}(t)}\big(w_0(n)\big) - f_{\mathcal{A}(t)}\big(w_3(n)\big) = \frac{4}{3}(\beta_n - \gamma_n)\,,$$

and

$$f_{\mathcal{A}(t)}\big(w_1(n)\big) - f_{\mathcal{A}(t)}\big(w_2(n)\big) = \left(\frac{4}{3}t - 1\right)\alpha_n + \frac{1}{4}(\beta_n - \gamma_n)\,.$$

Consequently,

$$\begin{aligned}
&4\Big(f_{\mathcal{A}(t)}\big(w_1(n)\big) - f_{\mathcal{A}(t)}\big(w_2(n)\big)\Big)\\
&\qquad = \frac{3}{4}\Big(f_{\mathcal{A}(t)}\big(w_0(n)\big) - f_{\mathcal{A}(t)}\big(w_3(n)\big)\Big) + 4\left(\frac{4}{3}t - 1\right)t^n,
\end{aligned}$$

implying that

$$\begin{aligned}
&\max\big\{\,\big|f_{\mathcal{A}(t)}\big(w_1(n)\big) - f_{\mathcal{A}(t)}\big(w_2(n)\big)\big|\,,\big|f_{\mathcal{A}(t)}\big(w_0(n)\big) - f_{\mathcal{A}(t)}\big(w_3(n)\big)\big|\,\big\}\\
&\qquad\qquad \geqslant \frac{1}{2}\left|\frac{4}{3}t - 1\right| t^n\,.
\end{aligned}$$

This, in turn, means that there exists an infinite subset $I_0 \subseteq \mathbb{N}$ such that

$$\frac{\big|f_{\mathcal{A}(t)}(w) - f_{\mathcal{A}(t)}\big(w_i(n)\big)\big|}{|\hat{w} - \widehat{w_i(n)}|} \geqslant \left|\frac{1}{3}t - \frac{1}{4}\right|(2t)^n \quad \text{for all } i \in I_0\,.$$

This confirms (6.2) when $t > \frac{1}{2}$ and $t \neq \frac{3}{4}$. The case $t < \frac{1}{2}$, with $t \neq \frac{1}{4}$, is similar, in fact, symmetric. The above reasoning also explains why the theorem does not hold for $t = \frac{3}{4}$, and even with some extra efforts why Theorem 6.1 holds true. □

As the final remark we emphasize the following. The functions $\hat{f}_{\mathcal{A}(t)}$ we defined are certainly quite complicated in terms of classical analysis. However, their automata-theoretic representation is very simple, and also provides a fast method to compute their values at any point. Since $\mathcal{A}(t)$ contains only four states, which is the minimal number of states needed to compute a cubic polynomial, as we have seen, computationally $\hat{f}_{\mathcal{A}(t)}$ is not harder than any cubic polynomial in our formalism.

Acknowledgement. The authors are grateful to Turo Sallinen for helping in finalizing this chapter. In particular, Theorem 6.2 in the presented form comes from his master's thesis [21].

References

[1] J. Albert and J. Kari. Digital image compression. In M. Droste, W. Kuich, and H. Vogler, editors, *Handbook of weighted automata*, pages 453–479. Springer-Verlag, 2009. 991

[2] J.-P. Allouche and J. Shallit. *Automatic sequences*. Cambridge University Press, 2003. 991

[3] M. F. Barnsley. *Fractals everywhere*. Academic Press, 1988. 990, 991

[4] J. Berstel and M. Morcrette. Compact representation of patterns by finite automata. In *Proc. Pixim '89, Paris*, pages 387–402, 1989. 991

[5] J. Berstel and C. Reutenauer. Rational series and their languages. In *EATCS Monographs on theoretical computer science*. Springer-Verlag, 1988. 996

[6] K. Culik, II and S. Dube. Affine automata and related techniques for generation of complex images. In *Proceedings of MFCS 1990*, Lecture notes in computer science 452, pages 224–231. Springer-Verlag, 1990. 991

[7] K. Culik, II and S. Dube. Rational and affine expressions for image processing. *Discrete Appl. Math.*, 41:85–120, 1993. 991

[8] K. Culik, II and J. Karhumäki. Finite automata computing real functions. *SIAM J. Comput.*, 23:789–814, 1994. 1002, 1015, 1016

[9] K. Culik, II and J. Kari. Image compression using weighted finite automata. *Computers & Graphics*, 17(3):305–313, 1993. 991

[10] K. Culik, II and J. Kari. Finite state transformations of images. *Computers & Graphics*, 20:125–135, 1996. 1017, 1018

[11] K. Culik, II and J. Kari. Computational fractal geometry with WFA. *Acta Inform.*, 34(2):151–166, 1997. 991

[12] K. Culik, II and J. Kari. *Handbook of formal languages*, volume 3, chapter Digital images and formal languages. Springer-Verlag, 1997. 991

[13] D. Derencourt, J. Karhumäki, M. Latteux, and A. Terlutte. On continuous functions computed by finite automata. *RAIRO-Theor. Inf. Appl.*, 29:387–403, 1994. 1020

[14] M. Droste, J. Kari, and P. Steinby. Observations on the smoothness properties of real functions computed by weighted finite automata. *Fund. Inform.*, 73(1,2):99–106, 2006. 1005, 1009

[15] S. Eilenberg. *Automata, languages and machines, Vol. A*. Academic Press, 1974. 990, 996, 1010

[16] T. Harju and J. Karhumäki. *Handbook of formal languages*, volume 1, chapter Morphisms. Springer-Verlag, 1997. 991

[17] J. Karhumäki, J. Kari, and J. Kupke. Binary operations on automatic functions. *RAIRO-Theor. Inf. Appl.*, 42(2):217–236, 2008. 1014, 1016

[18] J. Kari. Image processing using finite automata. In Z. Esik, C. Martin-Vide, and V. Mitrana, editors, *Recent advances in formal languages and applications*, Studies in computational intelligence, Vol. 25. Springer-Verlag, 2006. 991

[19] J. Kari, A. Kazda, and P. Steinby. On continuous weighted finite automata. Submitted for publication. 1004

[20] P. Prusinkiewicz and A. Lindenmayer. *The algorithmic beauty of plants*. Springer-Verlag, 1990. 990, 991

[21] T. Sallinen. Reaalifunktioiden laskennasta automaateilla. Master's thesis, University of Turku, 2009. 1024

[22] A. Salomaa and M. Soittola. *Automata-theoretic aspects of formal power series*. Springer-Verlag, 1978. 996, 1016

[23] M. P. Schützenberger. On the definition of a family of automata. *Information and Control*, 4:245–270, 1961. 990

[24] J. von Neumann. *Theory of self-reproducing automata*. University of Illinois Press, 1966. 990

[25] S. Wolfram. *A new kind of science*. Wolfram Media, 2002. 990