Centre Fédéré en Vérification

Technical Report number 2007.86

A Generalization of Cobham's Theorem to Automata over Real Numbers

Bernard Boigelot, Julien Brusten

This work was partially supported by a FRFC grant: 2.4530.02

http://www.ulb.ac.be/di/ssd/cfv

A Generalization of Cobham's Theorem to Automata over Real Numbers

Bernard Boigelot and Julien Brusten*

Institut Montefiore, B28
Université de Liège
B-4000 Liège, Belgium
{boigelot,brusten}@montefiore.ulg.ac.be

Abstract. This paper studies the expressive power of finite-state automata recognizing sets of real numbers encoded positionally. It is known that the sets that are definable in the first-order additive theory of real and integer variables $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ can all be recognized by weak deterministic Büchi automata, regardless of the encoding base $r > 1$. In this paper, we prove the reciprocal property, i.e., that a subset of $\mathbb{R}$ that is recognizable by weak deterministic automata in every base $r > 1$ is necessarily definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This result generalizes to real numbers the well-known Cobham's theorem on the finite-state recognizability of sets of integers. Our proof gives interesting insight into the internal structure of automata recognizing sets of real numbers, which may lead to efficient data structures for handling these sets.

1 Introduction

The verification of infinite-state systems, in particular the reachability analysis of systems modeled as finite-state machines extended with unbounded variables, has prompted the development of symbolic data structures for representing the sets of values that have to be handled during state-space exploration [Boi98].

A simple representation strategy consists in using finite-state automata: The values in the considered domain are encoded as words over a given finite alphabet; a set of values is thus encoded as a language. If this language is regular, then a finite-state automaton that accepts it forms a representation of the set [WB98].

This approach has many advantages: Regular languages are closed under all usual set-theory operators (intersection, union, complement, Cartesian product, projection, ...), and automata are easy to manipulate algorithmically. Deterministic automata can also be reduced to a canonical form, which simplifies comparison operations between sets.

The expressive power of automata is also well suited for verification applications. In the case of programs manipulating unbounded integer variables, it is known for a long time that the sets of integers that can be recognized by

* Research fellow ("aspirant") of the Belgian Fund for Scientific Research (F.R.S.-FNRS).

a finite-state automaton using the positional encoding of numbers in a base $r > 1$ correspond to those definable in an extension of Presburger arithmetic, i.e., the first-order additive theory of the integers $\langle \mathbb{Z}, +, < \rangle$ [Büc62]. Furthermore, the well known Cobham's theorem characterizes the sets that are representable by automata in all bases $r > 1$ as being exactly those that are Presburger-definable [Cob69,BHMV94].

In order to analyze systems relying on integer and real variables, such as timed or hybrid automata, automata-based representations of numbers can be generalized to real values [BBR97]. From a theoretical point of view, this amounts to moving from finite-word to infinite-word automata, which is not problematic. It has been shown that the sets of reals that can be recognized by infinite-word automata in a given encoding base are those definable in an extension of the first-order additive theory of real and integers variables $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ [BRW98].

In practice though, handling infinite-word automata can be difficult, especially if set complementation needs to be performed. It is however known that, for representing the sets definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, the full expressive power of Büchi automata is not required, and that the much simpler subclass of *weak deterministic* automata is sufficient [BJW05]. The advantage is that, from an algorithmic perspective, handling weak automata is similar to manipulating finite-word automata.

A natural question is then to characterize precisely the expressive power of weak deterministic automata representing sets of real numbers. For a given encoding base $r > 1$, it is known that the representable sets form a base-dependent extension of $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This covers, in particular, all the sets definable in $\langle \mathbb{R}, \mathbb{Z}, +, <, P_r \rangle$, where P_r is a predicate that checks whether its argument is a power of r [Bru06].

This paper is aimed at characterizing the subsets of $\mathbb{R}$ that can be represented as weak deterministic automata in multiple bases. Our central result is to show that, for two relatively prime bases r_1 and r_2, the sets that are simultaneously recognizable in bases r_1 and r_2 can be defined in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. As a corollary, such sets are then representable in any base $r > 1$.

The intuition behind our proof is the following. First, we reduce the problem to characterizing the representable subsets of $[0, 1]$. We then introduce the notion of interval boundary points, as points with special topological properties, and establish that a set representable in multiple bases can only contain finitely many such points. Finally, we show that this property implies that S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. The argument used for this last step provides a description of the internal structure of automata representing sets definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This result may help to develop efficient data structures for handling such sets.

2 Representing Sets of Numbers with Automata

In this section, we briefly present the automata-based representations of sets of integer and real values.

2.1 Number Decision Diagrams

Let $r > 1$ be an integer *base*. A natural number $x \in \mathbb{N}$ can be *encoded* positionally in base r by finite words $b_{p-1}b_{p-2}\ldots b_1b_0$ over the alphabet $\Sigma_r = \{0, 1, \ldots, r-1\}$, such that $x = \sum_{i=0}^{p-1} b_i r^i$. Negative values are encoded by their r's-complement, i.e., the encodings of $x \in \mathbb{Z}$ with $x < 0$ are formed by the last p digits of the encodings of $r^p + x$. The length p of the encodings of a number $x \in \mathbb{Z}$ is not fixed, but must be non-zero and large enough for $-r^{p-1} \leq x < r^{p-1}$ to hold. As a consequence, the most significant digit of encodings, called the *sign digit*, is equal to $r - 1$ for strictly negative numbers, and to 0 for positive numbers.

This encoding scheme maps a subset S of $\mathbb{Z}$ onto a language over Σ_r. If the language containing all the encodings of the elements of S is regular, then a finite-state automaton that accepts it is called a *Number Decision Diagram (NDD)*, and is said to represent, or recognize, the set S. NDDs can be generalized to representing subsets of $\mathbb{Z}^n$, i.e., sets of vectors, for any $n > 0$ [Büc62,WB95,Boi98].

It has been shown [Büc62,Vil92,BHMV94] that the subsets of $\mathbb{Z}$ recognizable by NDDs in a base $r > 1$ are exactly those that can be defined in the first-order theory $\langle \mathbb{Z}, +, <, V_r \rangle$ where $V_r(x)$ is the function mapping an integer $x > 0$ to the greatest power of r dividing it. Moreover, the sets that are recognizable by NDDs in every base $r > 1$ have been characterized by Cobham [Cob69] as being exactly those that are definable in $\langle \mathbb{Z}, +, < \rangle$, i.e., Presburger arithmetic. This result has been extended to subsets of $\mathbb{Z}^n$ by Semenov [Sem77].

Computing the intersection, union, complementation, difference and Cartesian product of sets represented by NDDs reduces to performing the corresponding operations on the languages accepted by the automata. Projection is more tricky, as the resulting automaton has to be completed in order to accept all the encodings of the vectors it recognizes. Finally, since NDDs are finite-word automata, they can be determinized, as well as minimized into a canonical form.

2.2 Real Number Automata

Real numbers can also be encoded positionally. Let $r > 1$ be a base. An encoding w of a number $x \in \mathbb{R}$ is an infinite word $w_I \cdot \star \cdot w_F$ over $\Sigma_r \cup \{\star\}$, where $w_I \in \Sigma_r^*$ encodes the integer part $x_I \in \mathbb{Z}$ of x, and $w_F \in \Sigma_r^\omega$ its fractional part $x_F \in [0, 1]$, i.e., we have $w_F = b_1b_2b_3\ldots$ with $x_F = \Sigma_{i>0} b_i r^{-i}$. Note that some numbers have two distinct encodings with the same integer-part length. For example, in base 10, the number 11/2 has the encodings $0^+ \cdot 5 \cdot \star \cdot 5 \cdot 0^\omega$ and $0^+ \cdot 5 \cdot \star \cdot 4 \cdot 9^\omega$. Such encodings are said to be *dual*. We denote by Λ_r the set of valid prefixes of base-r encodings that include a separator, i.e., $\Lambda_r = \{0, r-1\} \cdot \Sigma_r^* \cdot \star \cdot \Sigma_r^*$.

Similarly to the case of integers, the base-r encoding scheme transforms a set $S \subseteq \mathbb{R}$ into a language $L(S) \subseteq \Lambda_r \cdot \Sigma_r^\omega$. A *Real Number Automaton (RNA)* is defined as a Büchi automaton that accepts the language containing all the base-r encodings of the elements of S. This representation can be generalized into *Real Vector Automata (RVA)*, suited for subsets of $\mathbb{R}^n$ $(n > 0)$ [BBR97].

The expressiveness of RVA (and RNA) has been studied [BRW98]: The subsets of $\mathbb{R}^n$ that are representable in a base $r > 1$ are exactly those that are definable in the first-order theory $\langle \mathbb{R}, \mathbb{Z}, +, <, X_r \rangle$, where $X_r(x, u, k)$ is a base-dependent predicate that is true iff u is an integer power of r, and there exists an encoding of x in which the digit at the position specified by u is equal to k. The predicate X_r can alternatively be replaced by a function V_r analogous to the one defined in the integer case [Bru06]: We say that $x \in \mathbb{R}$ *divides* $y \in \mathbb{R}$ iff there exists an integer k such that $kx = y$. The function V_r is then defined such that $V_r(x)$ returns the greatest power of r dividing x, if it exists, and 1 otherwise.

2.3 Weak Deterministic RNA

As in the case of integers, applying most set-theory operators to RNA (or RVA) reduces to carrying out the same operations on their accepted language. This is somehow problematic, since operations like set complementation are typically costly and tricky to implement on infinite-word automata [KV05].

In order to alleviate this problem, it has been shown that the full expressive power of Büchi automata is not needed for representing the subsets of $\mathbb{R}^n$ ($n \geq 0$), that are definable in the first-order additive theory $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ of mixed integer and real variables [BJW05]. Such sets can indeed be represented by *weak deterministic* RVA, i.e., deterministic RVA such that their set of states can be partitioned into disjoint subsets $Q_1, \ldots, Q_m$, where each Q_i contains only either accepting or non-accepting states, and there exists a partial order $\leq$ on the sets $Q_1, \ldots, Q_m$ such that for every transition (q, a, q') of the automaton, with $q \in Q_i$ and $q' \in Q_j$, we have $Q_j \leq Q_i$.

As remarked in [Wil93], weak deterministic automata are infinite-word automata that can be manipulated essentially in the same way as finite-word ones. There exist efficient algorithms for applying to weak deterministic RVA all classical set-theory operators (intersection, union, complement, Cartesian product, projection, ...) [BJW05]. Furthermore, such RVA can be minimized into a canonical form.

It is worth mentioning that expressiveness of weak deterministic RVA is clearly not limited to the sets that are definable in the first-order additive theory of the integers and reals. For instance, the set of (negative and positive) integer powers of the representation base is clearly recognizable. Let $r > 1$ be a base, and $P_r(x)$ be a predicate that holds iff x is an integer power of r. It has been shown, using a quantifier elimination result for $\langle \mathbb{R}, 1, +, \leq, P_r \rangle$ [vdD85,AY07], that all the sets definable in $\langle \mathbb{R}, \mathbb{Z}, +, <, P_r \rangle$ can also be represented by weak deterministic RVA in base r [Bru06].

3 Problem Reduction

Let $S \subseteq \mathbb{R}$ be a set recognizable by a weak deterministic RNA $\mathcal{A}_1$, assumed to be in canonical form, in a base $r > 1$. Each accepting path of $\mathcal{A}$ contains exactly one occurrence of the separator symbol $\star$. Each transition labeled by $\star$ thus links two distinct strongly connected components of $\mathcal{A}$. Since there are only finitely many such transitions, the language L accepted by $\mathcal{A}$ is of the form $\bigcup_i L_i^I \cdot \star \cdot L_i^F$, where the union is finite, and for all i, $L_i^I \subseteq \Sigma_r^*$ encodes the integer part, and $L_i^F \subseteq \Sigma_r^\omega$ the fractional part, of the encodings of numbers $x \in S$. More precisely, for every i, let $S_i^I \subseteq \mathbb{Z}$ denote the set encoded by L_i^I and let $S_i^F \subseteq [0,1]$ denote the set encoded by $0^+ \cdot \star \cdot L_i^F$. We have $S = \bigcup_i (S_i^I + S_i^F)$. Note that each L_i^I is recognizable by a NDD in base r and that, similarly, each language of the form $0^+ \cdot \star \cdot L_i^F$ is recognizable by a RNA (except for the dual encodings of 0 and 1, which can be explicitly added to the language if needed).

The decomposition of S into sets S_i^I and S_i^F of integer and fractional parts does not depend on the representation base. Therefore, if S is recognizable in two relatively prime bases r_1 and r_2, then so are S_i^I and S_i^F for every i. From Cobham's theorem, each S_i^I must then be definable in $\langle \mathbb{Z}, +, < \rangle$. In order to show that S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, it is hence sufficient to prove that each S_i^F is definable in that theory. We have thus reduced the problem of characterizing the subsets of $\mathbb{R}$ that are simultaneously recognizable in two relatively prime bases to the same problem over the subsets of $[0,1]$.

4 Interval Boundary Points

We now consider a set $S \subseteq [0,1]$ represented by a weak deterministic RNA $\mathcal{A}$. We define the *interval boundary points* of S as points with specific topological properties, and establish a relation between the existence of such points and some structures in the transition graph of $\mathcal{A}$.

4.1 Definitions

A *neighborhood* $N_\varepsilon(x)$ of a point $x \in \mathbb{R}$, with $\varepsilon > 0$, is the set $N_\varepsilon(x) = \{y \mid |x - y| < \varepsilon\}$. A point $x \in \mathbb{R}$ is a *boundary point* of S iff all its neighborhoods contain points from S as well as from its complement $\overline{S}$, i.e., $\forall \varepsilon > 0 : N_\varepsilon(x) \cap S \neq \emptyset \wedge N_\varepsilon(x) \cap \overline{S} \neq \emptyset$.

A *left neighborhood* $N_\varepsilon^<(x)$ of a point $x \in \mathbb{R}$, with $\varepsilon > 0$, is the set $N_\varepsilon^<(x) = \{y \mid x - \varepsilon < y < x\}$. Similarly, a *right neighborhood* $N_\varepsilon^>(x)$ of x is defined as $N_\varepsilon^>(x) = \{y \mid x < y < x + \varepsilon\}$. A boundary point x of S is a *left interval boundary point* of S iff it admits a left neighborhood $N_\varepsilon^<(x)$ that is entirely contained in either S or $\overline{S}$, i.e., $\exists \varepsilon > 0 : N_\varepsilon^<(x) \subseteq S \vee N_\varepsilon^<(x) \subseteq \overline{S}$. *Right interval boundary points* are defined in the same way. A point $x \in S$ is an *interval boundary point* of S iff it is a left or a right interval boundary point of S.

Each interval boundary point x of S is thus characterized by its direction (left or right), its polarity w.r.t. S (i.e., whether $x \in S$ or $x \notin S$), and the polarity of its left or right neighborhoods of sufficiently small size (i.e., whether they are subsets of S or of $\overline{S}$). The possible combinations define eight *types* of interval boundary points, that are illustrated in Figure 1.

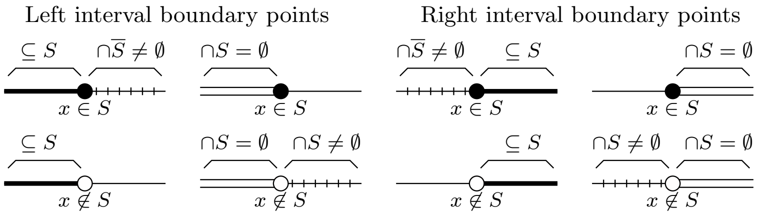

Fig. 1. Types of interval boundary points.

4.2 Recognizing Interval Boundary Points

Recall that $\mathcal{A}$ is a weak deterministic RNA recognizing a set $S \subseteq [0, 1]$. We assume w.l.o.g. that $\mathcal{A}$ is canonical and complete, in the sense that from each state q and alphabet symbol a, there exists an outgoing transition from q labeled by a. Consider a path π of $\mathcal{A}$ that reads an encoding w of a left interval boundary point x of S. Since $\mathcal{A}$ is weak, π eventually reaches a strongly connected component C that it does not leave. The accepting status of C corresponds to the polarity of x w.r.t. S.

Since x is a left interval boundary point, all its sufficiently small left neighborhoods are either subsets of S or subsets of $\overline{S}$, depending on the type of x. Hence, from each state s of C visited infinitely many times by π, its outgoing transitions labeled by smaller digits than the one read in π must necessarily lead to either the universal or the empty strongly connected component of $\mathcal{A}$. It follows that, after having reached some state s in C, the path π follows the transitions within C that are labeled by the smallest possible digits, hence it eventually cycles through a loop. A similar result holds for right interval boundary points, which are read by paths that eventually follow the largest possible digits in their terminal strongly connected component.

As a consequence, every base-r encoding w of an interval boundary point of S is necessarily *ultimately periodic*, i.e., such that $w = u \cdot v^{\omega}$, with $u \in \Lambda_r$ and $v \in \Sigma_r^+$. Besides, each ultimate period v of such encodings can be uniquely determined from a suitable state of $\mathcal{A}$ associated with a direction (left or right). We therefore have the following results.

Theorem 1. *Each interval boundary point of a subset of* $[0, 1]$ *that is recognizable by a weak deterministic RNA is a rational number.*

Theorem 2. *Let* $S \subseteq [0, 1]$ *be a set recognizable by a weak deterministic RNA in a base* $r > 1$*. The set of ultimate periods of the base-*r *encodings of the interval boundary points of* S *is finite.*

4.3 Recognizing Interval Boundary Points in Multiple Bases

Consider now a set $S \subseteq [0, 1]$ that is simultaneously recognizable by weak deterministic RNA in two relatively prime bases $r_1 > 1$ and $r_2 > 1$. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ denote, respectively, such RNA.

Suppose that S has infinitely many interval boundary points. From Theorem 2, there must exist some ultimate period $v \in \Sigma_{r_1}^{+}$ such that infinitely many interval boundary points of S have base-r_1 encodings of the form $u_i \cdot v^{\omega}$, with $\forall i : u_i \in \Lambda_{r_1}$. Moreover, the language L of the words u_i for which $u_i \cdot v^{\omega}$ encodes an interval boundary point of S, and such that u_i and v do not end with the same digit, is infinite and regular. (The restriction on the last digit of u_i and v expresses that u_i is the smallest aperiodic prefix of $u_i \cdot v^{\omega}$.) Indeed, each $u_i \in L$ can be recognized by a path from the initial state of $\mathcal{A}$ to a state from which v can be read as the ultimate period of an encoding of an interval boundary point.

Hence, there exist $w_1 \in \Lambda_{r_1}$ and $w_2, w_3 \in \Sigma_{r_1}^{*}$, with $|w_2| > 0$, such that $\forall k : w_1 \cdot (w_2)^k \cdot w_3 \in L$. Furthermore, we have that $w_2 \cdot w_3$ and v do not end with the same digit.

Thus, for each $k \geq 0$, there exists an interval boundary point of S with a base-r_1 encoding of the form $w_1 \cdot (w_2)^k \cdot w_3 \cdot v^{\omega}$. Each word in this language is ultimately periodic, thus it encodes in base r_1 a rational number that can also be encoded by an ultimately periodic word in base r_2. We use the following lemma.

Lemma 1. *Let $r_1 > 1$ and $r_2 > 1$ be relatively prime bases, and let $w_1 \in \Lambda_{r_1}, w_2, w_3, w_4 \in \Sigma_{r_1}^{*}$, with $|w_2| > 0$, $|w_4| > 0$, such that the words $w_2 \cdot w_3$ and w_4 do not end with the same digit. The subset of $\mathbb{Q}$ encoded in base r_1 by the language $w_1 \cdot (w_2)^{*} \cdot w_3 \cdot (w_4)^{\omega}$ cannot be encoded in base r_2 with only a finite number of ultimate periods.*

Proof. The proof is given in Appendix A. □

Together with Theorem 2, this lemma contradicts our assumption that S has infinitely many interval boundary points. We thus have the following theorem.

Theorem 3. *If a set $S \subseteq [0, 1]$ is simultaneously recognizable by weak deterministic RNA in two relatively prime bases, then it has finitely many interval boundary points.*

We therefore call a set that satisfies the hypotheses of Theorem 3 a *finite-boundary set*.

5 Finite-Boundary Sets

Our goal is now to characterize the structure of the transition graph of RNA that recognize finite-boundary sets. We start by establishing some properties that hold for all weak deterministic RNA, and then focus on the specific case of finite-boundary sets.

5.1 Properties of Weak Deterministic RNA

Let $\mathcal{A}$ be a weak deterministic RNA, which we assume to be complete and canonical, recognizing a subset of $\mathbb{R}$ in a base $r > 1$. Consider a strongly connected component C of $\mathcal{A}$ such that each of its outgoing transitions leads to either the universal or the empty strongly connected component, i.e., those accepting respectively the languages Σ_r^{ω} and $\emptyset$.

Lemma 2. *Let π be a minimal (resp. maximal) infinite path within C, i.e., a path that follows from each visited state the transition of C labeled by the smallest (resp. largest) possible digit. The destination of all outgoing transitions from states visited by π, and that are labeled by a smaller (resp. larger) digit than the one read in π, is identical.*

Proof. We first study the case of two transitions t_1 and t_2 originating from the same state s visited by π, that are respectively labeled by digits d_1, d_2 smaller that the digit d read from s in π. Among the digits that satisfy this condition, one can always find consecutive values, hence it is sufficient to consider the case where $d_2 = d_1 + 1$.

Let σ be a finite path that reaches s from the initial state of $\mathcal{A}$. By appending to σ suffixes that read $d_1 \cdot (r-1)^\omega$ and $d_2 \cdot 0^\omega$, one obtains paths that recognize dual encodings of the same number, hence these paths must be either both accepting or both non-accepting. Therefore, t_1 and t_2 share the same destination.

Consider now transitions t_1 and t_2 from distinct states s_1 and s_2 visited by π, labeled by smaller digits than those – respectively denoted d_1 and d_2 – read in π. We can assume w.l.o.g. that s_1 and s_2 are consecutive among the states visited by π that have such outgoing transitions. In other words, the subpath of π that links s_1 to s_2 is labeled by a word of the form $d_1 \cdot 0^k$, with $d_1 > 0$ and $k \geq 0$.

Let σ' be a finite path that reaches s_1 from the initial state of $\mathcal{A}$. Appending to σ' suffixes that read $(d_1 - 1) \cdot (r-1)^\omega$ and $d_1 \cdot 0^\omega$ yields paths that read dual encodings of the same number, hence these paths must be either both accepting or both non-accepting. The destinations of the transitions that leave C from s_1 and s_2 must thus be identical.

The case of maximal paths is handled in the same way. □

The following result now expresses a constraint on the trivial (acyclic) strongly connected components of the fractional part of $\mathcal{A}$ (i.e., the part of $\mathcal{A}$ reached after reading an occurrence of the symbol $\star$).

Lemma 3. *From any trivial strongly connected component of the fractional part of $\mathcal{A}$, there must exist a reachable strongly connected component that is neither empty, trivial, nor universal.*

Proof. The proof is by contradiction. Let $\{s\}$ be a trivial strongly connected component of the fractional part of $\mathcal{A}$. Assume that all paths from s eventually reach the universal or the empty strongly connected component, after passing only through trivial components. As a consequence, the language accepted from s is of the form $L \cdot \Sigma_r^\omega$, where $L \subset \Sigma_r^*$ is finite. We can require w.l.o.g. that all words in L share the same length l. Note that L cannot be empty or equal to Σ_r^l, since s does not belong to the empty or universal components.

Each word in Σ_r^l can be seen as the base-r encoding of an integer in the interval $[0, r^l - 1]$. Since L is neither empty nor universal, there exist two words $w_1, w_2 \in \Sigma_r^l$ that do not both belong to L or to $\Sigma_r^l \setminus L$, and that encode two consecutive integers n and $n+1$. Then, $u \cdot w_2 \cdot 0^\omega$ and $u \cdot w_1 \cdot (r-1)^\omega$ encode the same number in base r, where u is the label of an arbitrary path from the initial state of $\mathcal{A}$ to s. This contradicts the fact that $\mathcal{A}$ accepts all the encodings of the numbers it recognizes. □

5.2 Properties of RNA Recognizing Finite-Boundary Sets

Theorem 4. *Let $\mathcal{A}$ be a weak deterministic RNA, supposed to be in complete and canonical form, recognizing a finite-boundary set $S \subseteq [0,1]$. Each non-trivial, non-empty and non-universal strongly connected component of the fractional part of $\mathcal{A}$ takes the form of a single cycle. Moreover, from each such component, the only reachable strongly connected components besides itself are the empty or the universal ones.*

Proof. Let C be a non-trivial, non-empty and non-universal strongly connected component of the fractional part of $\mathcal{A}$, and let s be an arbitrary state of C. The path π from s that stays within C and follows the transitions with the smallest possible digits is cyclic, and determines the ultimate period of encodings of some interval boundary points of S. If C contains other cycles, or if C is reachable from other non-trivial strongly connected components in the fractional part, then π can be prefixed by infinitely many reachable paths from an entry state of the fractional part of $\mathcal{A}$ to s. This contradicts the fact that S has only finitely many interval boundary points. That no trivial strongly connected component can be reachable from C then follows from Lemma 3. □

This result characterizes quite precisely the shape of the fractional part of a weak deterministic RNA recognizing a finite-boundary set: Its transition graph is first composed of a bottom layer of strongly connected components containing only the universal and the empty one, and then a (possibly empty) layer of single-cycle components leading to the bottom layer. Thanks to Lemma 2, the transitions that leave a single-cycle component with a smaller (or larger) digit all lead to the same empty or universal component (which may differ for the smaller and larger cases). Thus, each single-cycle component can simply be characterized by its label and the polarity of its smaller and greater alternatives. Finally, the two layers of non-trivial strongly connected components can be reached through an acyclic structure of trivial components, such that from each of them, there is at least one outgoing path leading to a single-cycle component.

As a consequence, we are now able to describe the language accepted by such a RNA.

Theorem 5. *Let $\mathcal{A}$ be a weak deterministic RNA recognizing a finite-boundary set $S \subseteq [0,1]$ encoded in a base $r > 1$. The language L accepted by $\mathcal{A}$ can be expressed as*

$$L = \bigcup_i L' \cdot w_i \cdot \Sigma_r^\omega \,\cup\, \bigcup_i L' \cdot w'_i \cdot (v_i)^\omega \,\cup\, \bigcup_i L' \cdot w''_i \cdot (\Sigma_r^\omega \setminus (v'_i)^\omega) \cup L_0 \cup L_1,$$

where each union is finite, $\forall i : w_i, w'_i, w''_i, v_i, v'_i \in \Sigma_r^*$ with $|v_i| > 0$, $|v'_i| > 0$, $L' = 0^+ \cdot \star$, L_0 is either empty or equal to $(r-1)^+ \cdot \star \cdot (r-1)^\omega$, and L_1 is either empty or equal to $0^+ \cdot 1 \cdot \star \cdot 0^\omega$.

(The terms L_0 and L_1 are introduced in order to deal with the dual encodings of 0 and 1.)

In the expression given by Theorem 5, each term of the union encodes a subset of $[0, 1]$ that is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$: $L' \cdot w_i \cdot \Sigma_r^\omega$ defines an interval $[a, b]$, with $a, b \in \mathbb{Q}$, the terms $L' \cdot w'_i \cdot (v_i)^\omega$, L_0 and L_1 correspond to single rational numbers $c \in \mathbb{Q}$, and $L' \cdot w''_i \cdot (\Sigma_r^\omega \setminus (v'_i)^\omega)$ recognizes a set $[a, b] \setminus \{c\}$ with $a, b, c \in \mathbb{Q}$. This shows that the set $S \subseteq [0, 1]$ recognized by $\mathcal{A}$ is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. Combining this result with Theorem 3, as well as the reduction discussed in Section 3, we get our main result:

Theorem 6. *If a set $S \subseteq \mathbb{R}$ is simultaneously recognizable by weak deterministic RNA in two relatively prime bases, then it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

Corollary 1. *A set $S \subseteq \mathbb{R}$ is recognizable by weak deterministic RNA in every base $r > 1$ iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

6 Conclusions and Future Work

The main contribution of this work is to show that the subsets of $\mathbb{R}$ that can be recognized by weak deterministic RNA in all integer bases $r > 1$ are exactly those that are definable in the first-order additive theory of the real and integer numbers $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. Our central result is actually stronger, stating that recognizability in two relatively prime bases r_1 and r_2 is sufficient for forcing definability in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. Using the same argument as in the proof of Lemma 1, this result can directly be extended to bases r_1 and r_2 that do not share the same set of prime factors. This differs slightly from the statement of Cobham's original theorem, which considers instead bases that are multiplicatively independent, i.e., that cannot be expressed as integer powers of the same integer [Cob69,BHMV94]. Unfortunately, our approach does not easily generalize to multiplicatively independent bases, since Theorem 3 then becomes invalid. Addressing this issue is an interesting open problem.

Another contribution is a detailed characterization of the transition graph of weak deterministic RNA that represent subsets of $\mathbb{R}$ defined in first-order additive arithmetic. This characterization could be turned into efficient data structures for handling such RNA. In particular, since their fractional parts recognize a finite union of interval and individual rational values, an efficient representation might be based on symbolic data structures such as BDDs for handling large but finite enumerations. Another possible application is the extraction of formulas from automata-based representations of sets [Lat05,Ler05].

Finally, another goal will be to extend our results to sets in higher dimensions, i.e., to generalize Semenov's theorem [Sem77] to automata over real vectors.

References

[AY07] J. Avigad and Y. Yin. Quantifier elimination for the reals with a predicate for the powers of two. *Theoretical Computer Science*, 370:48–59, 2007.

[BBR97] 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, Haifa, June 1997. Springer.

[BHMV94] V. Bruyère, G. Hansel, C. Michaux, and R. Villemaire. Logic and p-recognizable sets of integers. *Bulletin of the Belgian Mathematical Society*, 1(2):191–238, March 1994.

[BJW05] B. Boigelot, S. Jodogne, and P. Wolper. An effective decision procedure for linear arithmetic over the integers and reals. *ACM Transactions on Computational Logic*, 6(3):614–633, 2005.

[Boi98] B. Boigelot. *Symbolic methods for exploring infinite state spaces.* PhD thesis, Université de Liège, 1998.

[Bru06] J. Brusten. *Etude des propriétés des RVA.* Graduate thesis, Université de Liège, May 2006.

[BRW98] B. Boigelot, S. Rassart, and P. Wolper. On the expressiveness of real and integer arithmetic automata. In *Proc. 25th ICALP*, volume 1443 of *Lecture Notes in Computer Science*, pages 152–163, Aalborg, July 1998. Springer.

[Büc62] J. R. Büchi. On a decision method in restricted second order arithmetic. In *Proc. International Congress on Logic, Methodoloy and Philosophy of Science*, pages 1–12, Stanford, 1962. Stanford University Press.

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

[KV05] O. Kupferman and M.Y. Vardi. Complementation constructions for nondeterministic automata on infinite words. In *Proc. 11th TACAS*, volume 3440 of *Lecture Notes in Computer Science*, pages 206–221, Edinburgh, April 2005. Springer.

[Lat05] L. Latour. *Presburger arithmetic: from automata to formulas.* PhD thesis, Université de Liège, 2005.

[Ler05] J. Leroux. A polynomial time Presburger criterion and synthesis for number decision diagrams. In *Proc. 20th LICS*, pages 147–156, Chicago, June 2005. IEEE Computer Society.

[Sem77] A.L. Semenov. Presburgerness of predicates regular in two number systems. *Siberian Mathematical Journal*, 18:289–299, 1977.

[vdD85] L. van den Dries. The field of reals with a predicate for the powers of two. *Manuscripta Mathematica*, 54:187–195, 1985.

[Vil92] R. Villemaire. The theory of $\langle \mathbb{N}, +, V_k, V_l \rangle$ is undecidable. *Theoretical Computer Science*, 106(2):337–349, 1992.

[WB95] P. Wolper and B. Boigelot. An automata-theoretic approach to Presburger arithmetic constraints. In *Proc. 2nd SAS*, volume 983 of *Lecture Notes in Computer Science*, Glasgow, September 1995. Springer.

[WB98] P. Wolper and B. Boigelot. Verifying systems with infinite but regular state spaces. In *Proc. 10th CAV*, volume 1427 of *Lecture Notes in Computer Science*, pages 88–97, Vancouver, June 1998. Springer.

[Wil93] T. Wilke. Locally threshold testable languages of infinite words. In *Proc. 10th STACS*, volume 665 of *Lecture Notes in Computer Science*, pages 607–616, Würzburg, 1993. Springer.

A Proof of Lemma 1

For a base $r > 1$ and a word $w \in \Lambda_r \cdot \Sigma_r^\omega$, let $[w]_r$ denote the real number encoded by w in that base. Similarly, for $w \in \{0, r-1\} \cdot \Sigma_r^*$, let $[w]_r$ denote the integer number encoded by w, i.e., $[w]_r = [w \cdot \star \cdot 0^\omega]_r$. For every $k \geq 0$, we define $x_k = [w_1 \cdot (w_2)^k \cdot w_3 \cdot (w_4)^\omega]_{r_1}$.

The prefix w_1 can be decomposed into $w_1 = w_1' \cdot \star \cdot w_1''$, with $w_1' \in \{0, r_1 - 1\} \cdot \Sigma_{r_1}^*$ and $w_1'' \in \Sigma_{r_1}^*$. We then have for every $k > 0$,

$$x_k = \frac{y_k}{r_1^{|w_1''|+k|w_2|+|w_3|}(r_1^{|w_4|} - 1)}, \tag{1}$$

with $y_k = (r_1^{|w_4|} - 1)[w_1' \cdot w_1'' \cdot w_2^k \cdot w_3]_{r_1} + [0 \cdot w_4]_{r_1}$. Remark that y_k is an integer, but cannot be a multiple of r_1. Indeed, we have $y_k \bmod r_1 = ([0 \cdot w_4]_{r_1} - [w_1' \cdot w_1'' \cdot w_2^k \cdot w_3]_{r_1}) \bmod r_1$, which is non-zero thanks to the hypothesis on the last digits of $w_2 \cdot w_3$ and w_4. For every $k > 0$, we have

$$y_k = \frac{z_k}{r_1^{|w_2|} - 1},$$

with $z_k = a r_1^{k|w_2|} + b$, $a = r_1^{|w_3|}(r_1^{|w_4|} - 1)((r_1^{|w_2|} - 1)[w_1' \cdot w_1'']_{r_1} + [0 \cdot w_2]_{r_1})$, and $b = -r_1^{|w_3|}(r_1^{|w_4|} - 1)[0 \cdot w_2]_{r_1} + (r_1^{|w_2|} - 1)(r_1^{|w_4|} - 1)[0 \cdot w_3]_{r_1} + (r_1^{|w_2|} - 1)[0 \cdot w_4]_{r_1}$.

Substituting in (1), we get

$$x_k = \frac{z_k}{r_1^{|w_1''|+k|w_2|+|w_3|}(r_1^{|w_2|} - 1)(r_1^{|w_4|} - 1)}. \tag{2}$$

Since $z_k = (r_1^{|w_2|} - 1)y_k$ and $y_k \bmod r_1 \neq 0$, we have $z_k \bmod r_1 \neq 0$, hence $b \neq 0$. Consider a prime factor f of r_1, and define l as the greatest integer such that f^l divides b. For every $k > l$, we have $z_k \bmod f^l = 0$ and $z_k \bmod f^{l+1} = b \bmod f^{l+1} \neq 0$. It follows that the reduced rational expression of x_k, i.e., $x_k = n_k/d_k$ with $n_k, d_k \in \mathbb{Z}$, $d_k > 0$ and $\gcd(n_k, d_k) = 1$, is such that f^{k-l} divides d_k for every $k > l$. Indeed, the numerator of (2) is not divisible by f^{l+1} whereas its denominator is divisible by f^{k+1}.

Assume now, by contradiction, that the set $\{x_k \mid k \geq 0\}$ can be represented in base r_2 using only a finite number of ultimate periods. Then, there exists an ultimate period $v \in \Sigma_{r_2}^+$ such that for infinitely many values of k, we have $x_k = [u_k' \cdot \star \cdot u_k'' \cdot v^\omega]_{r_2}$, with $u_k' \in \{0, r_2 - 1\} \cdot \Sigma_{r_2}^*$ and $u_k'' \in \Sigma_{r_2}^*$. We then have, for these values of k,

$$x_k = \frac{[u_k' \cdot u_k'' \cdot v]_{r_2} - [u_k' \cdot u_k'']_{r_2}}{r_2^{|u_k''|}(r_2^{|v|} - 1)}.$$

Since $(r_2^{|v|} - 1)$ is bounded, and r_2 is relatively prime with r_1 by hypothesis, the denominator of this expression can only be divisible by a bounded number of powers of f, which contradicts our previous result. □