Centre Fédéré en Vérification

Technical Report number 2008.105

On the Sets of Real Numbers Recognized by Finite Automata in Multiple Bases

Bernard Boigelot, Julien Brusten, Véronique Bruyère

This work was partially supported by a FRFC grant: 2.4530.02
and by the MoVES project. MoVES (P6/39) is part of the IAP-Phase VI Interuniversity Attraction Poles Programme funded by the Belgian State, Belgian Science Policy

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

On the Sets of Real Numbers Recognized by Finite Automata in Multiple Bases*

Bernard Boigelot[1], Julien Brusten[1]**, and Véronique Bruyère[2]

[1] Institut Montefiore, B28
Université de Liège
B-4000 Liège, Belgium
{boigelot,brusten}@montefiore.ulg.ac.be
[2] Université de Mons-Hainaut
Avenue du Champ de Mars, 6
B-7000 Mons, Belgium
veronique.bruyere@umh.ac.be

Abstract. This paper studies the expressive power of finite automata recognizing sets of real numbers encoded in positional notation. We consider Muller automata as well as the restricted class of *weak deterministic automata*, used as symbolic set representations in actual applications. In previous work, it has been established that the sets of numbers that are recognizable by weak deterministic automata in two bases that do not share the same set of prime factors are exactly those that are definable in the first order additive theory of real and integer numbers $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This result extends *Cobham's theorem*, which characterizes the sets of integer numbers that are recognizable by finite automata in multiple bases.

In this paper, we first generalize this result to *multiplicatively independent* bases, which brings it closer to the original statement of Cobham's theorem. Then, we study the sets of reals recognizable by Muller automata in two bases. We show with a counterexample that, in this setting, Cobham's theorem does not generalize to multiplicatively independent bases. Finally, we prove that the sets of reals that are recognizable by Muller automata in two bases that do not share the same set of prime factors are exactly those definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. These sets are thus also recognizable by weak deterministic automata. This result leads to a precise characterization of the sets of real numbers that are recognizable in multiple bases, and provides a theoretical justification to the use of weak automata as symbolic representations of sets.

* This work is supported by the *Interuniversity Attraction Poles* program *MoVES* of the Belgian Federal Science Policy Office, and by the grant 2.4530.02 of the Belgian Fund for Scientific Research (F.R.S.-FNRS).

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

1 Introduction

By using the positional notation, real numbers can be encoded as infinite words over an alphabet composed of a fixed number of digits, with an additional symbol for separating their integer and fractional parts. This encoding scheme maps sets of numbers onto languages that describe precisely those sets.

This paper studies the sets of real numbers whose encodings can be accepted by finite automata. The motivation is twofold. First, since regular languages enjoy good closure properties under a large range of operators, automata provide powerful theoretical tools for establishing the decidability of arithmetic theories. In particular, it is known that the sets of numbers that are definable in the first-order additive theory of integers $\langle \mathbb{Z}, +, < \rangle$, also called *Presburger arithmetic*, are encoded by regular finite-word languages [Büc62,BHMV94]. This result translates into a simple procedure for deciding the satisfiability of Presburger formulas. Moving to infinite-word encodings and ω-regular languages, it can be extended to sets of real numbers definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, i.e., the first-order additive theory of real and integer variables. [BBR97,BJW05].

The second motivation is practical. Since finite automata are objects that are easily manipulated algorithmically, they can be used as actual data structures for representing symbolically sets of values. This idea has successfully been exploited in the context of computer-aided verification, leading to representations suited for the sets of real and integer vectors handled during symbolic state-space exploration [WB95,BJW05,EK06]. A practical limitation of this approach is the high computational cost of some operations involving infinite-word automata, in particular language complementation [Saf88,Var07]. However, it has been shown that a restricted form of automata, *weak deterministic* ones, actually suffices for handling the sets definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ [BJW05]. Weak automata can be manipulated with essentially the same cost as finite-word ones [Wil93], which alleviates the problem and leads to an effective representation system.

Whether a set of numbers can be recognized by an automaton generally depends on the chosen encoding base. For integer numbers, it is known that a set $S \subseteq \mathbb{Z}$ is recognizable in a base $r > 1$ iff it is definable in the theory $\langle \mathbb{Z}, +, <, V_r \rangle$, where V_r is a base-dependent function [BHMV94]. Furthermore, the well-known *Cobham's theorem* states that if a set $S \subseteq \mathbb{Z}$ is simultaneously recognizable in two bases $r > 1$ and $s > 1$ that are *multiplicatively independent*, i.e., such that $r^p \neq s^q$ for all $p, q \in \mathbb{N}_{>0}$, then S is *ultimately periodic*, i.e., it differs from a periodic subset of $\mathbb{Z}$ only by a finite set [Cob69]. Equivalently, such a set is definable in $\langle \mathbb{Z}, +, < \rangle$ [BHMV94]. It follows that such a set S is recognizable in every base. Our aim is to generalize as completely as possible Cobham's theorem to automata recognizing real numbers, by precisely characterizing the sets that are recognizable in multiple bases. We first consider the case, relevant for practical applications, of weak deterministic automata. In previous work, it has been established that a set of real numbers is simultaneously recognizable by weak deterministic automata in two bases that do not share the same set of prime factors iff this set is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ [BB07]. As a first contribution, we extend this result to pairs of multiplicatively independent bases. Since recogniz-

ability in two multiplicatively dependent bases is equivalent to recognizability in only one of them [BRW98], this result provides a complete characterization of the sets that are recognizable in multiple bases by weak deterministic automata.

Then, we move to sets recognized by Muller automata. We prove that there exists a set of real numbers recognizable in two multiplicatively independent bases that share the same set of prime factors, but that is not definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This shows that Cobham's theorem does not directly generalize to Muller automata recognizing sets of real numbers. Finally, we establish that a set $S \subseteq \mathbb{R}$ is simultaneously recognizable in two bases that do not share the same set of prime factors iff S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. As a corollary, such a set must then be recognizable by a weak deterministic automaton. Our result thus provides a theoretical justification to the use of weak automata, by showing that their expressive power corresponds precisely to the sets of reals recognizable by infinite-word automata in every encoding base.

2 Representing sets of real numbers with finite automata

Let $r > 1$ be an integer numeration *base* and let $\Sigma_r = \{0, \ldots, r-1\}$ be the corresponding set of *digits*. We encode a real number x in base r, most significant digit first, by words of the form $w_I \star w_F$, where $w_I \in \{0, r-1\}\Sigma_r^*$ encodes the integer part x_I of x and $w_F \in \Sigma_r^\omega$ encodes its fractional part x_F. Negative numbers are represented by their r's-complement. The length p of w_I is not fixed but has to be large enough for $-r^{p-1} \leq x_I < r^{p-1}$ to hold; thus, the most significant digit of an encoding of a real number is equal to 0 for positive numbers and to $r-1$ for negative ones [BBR97]. Some numbers have two distinct encodings with the same integer-part length, e.g., in base 10, the number $11/2$ admits the encodings $0^+5\star50^\omega$ and $0^+5\star49^\omega$. For a word $w = b^I_{p-1}b^I_{p-2}\ldots b^I_1 b^I_0 \star b^F_1 b^F_2 b^F_3 \ldots \in \{0, r-1\}\Sigma_r^* \star \Sigma_r^\omega$, we denote by $[w]_r$ the real number encoded by w in base r, i.e.,

$$[w]_r = \sum_{i=0}^{p-2} b^I_i r^i + \sum_{i>0} b^F_i r^{-i} + \begin{cases} 0 & \text{if } b^I_{p-1} = 0, \\ -r^{p-1} & \text{if } b^I_{p-1} = r-1. \end{cases}$$

For finite words $w \in \Sigma_r^*$, we denote by $[w]_r$ the natural number encoded by w, i.e., $[w]_r = [0w \star 0^\omega]_r$.

If the language formed by all the base-r encodings of the elements of a set $S \subseteq \mathbb{R}$ is ω-regular, then it can be accepted by a (non-unique) infinite-word automaton, called a *Real Number Automaton (RNA)* recognizing S. Such a set S is then said to be *r-recognizable*. RNA can be generalized into *Real Vector Automata (RVA)*, suited for subsets of $\mathbb{R}^n$, with $n > 0$ [BBR97].

RNA and RVA have originally been defined as Büchi automata [BBR97]. In this paper, we will instead consider them to be *deterministic Muller automata*. This adaptation can be made without loss of generality, since both classes of automata share the same expressive power [McN66,PP04]. The fact that RNA have a deterministic transition relation will simplify technical developments.

The subsets of $\mathbb{R}$ that are r-recognizable 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 holds whenever 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 [BRW98].

It is known that the full expressive power of infinite-word automata is not needed for representing the subsets of $\mathbb{R}$ that are definable in the first-order theory $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ [BJW05]. Indeed, such sets can be recognized by *weak deterministic automata*, i.e., deterministic Büchi automata 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 $Q_j \leq Q_i$.

A set recognized by a weak deterministic automaton in base r is said to be *weakly r-recognizable* and such an automaton is called a *weak RNA*.

It has been established [BJW05] that the r-recognizable sets $S \subseteq \mathbb{R}$ that are not weakly r-recognizable are exactly those that satisfy the *dense oscillating property*: One has $\exists x_1 \forall \varepsilon_1 \exists x_2 \forall \varepsilon_2 \exists x_3 \forall \varepsilon_3 \cdots$ such that $|x_{i+1} - x_i| < \varepsilon_i$ for all $i \geq 1$, $x_i \in S$ for all odd i, and $x_i \notin S$ for all even i.

In the technical sections of this paper, we will need to apply transformations to sets represented by RNA (or weak RNA), or to the chosen encoding base. The following results are immediate corollaries of [BRW98] and [BJW05].

Theorem 1. *Let $S \subseteq \mathbb{R}$, $r \in \mathbb{N}_{>1}$ and $a, b \in \mathbb{Q}$. If S is (resp. weakly) r-recognizable then the sets $aS + b$ and $S \cap [a, b]$ are (resp. weakly) r-recognizable as well.*

Theorem 2. *Let $S \subseteq \mathbb{R}$, $r \in \mathbb{N}_{>1}$, and $k \in \mathbb{N}_{>0}$. The set S is (resp. weakly) r-recognizable iff it is (resp. weakly) r^k-recognizable.*

3 Problem reductions

In the next sections, we will consider sets $S \subseteq \mathbb{R}$ that are simultaneously recognizable, either by RNA or by weak RNA, in two bases r and s that satisfy some conditions. We will then tackle the problem of proving that such sets are definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. In this section, we reduce this problem, by restricting the domain to the interval $[0, 1]$, and introducing the notion of boundary point.

3.1 Reduction to $[0, 1]$

Let $S \subseteq \mathbb{R}$ be a set recognized by a (resp. weak) RNA $\mathcal{A}$. Each accepting path of $\mathcal{A}$ reads exactly one occurrence of the symbol $\star$. Since $\mathcal{A}$ is finite-state, its accepted language $L(\mathcal{A})$ has the form $\bigcup_i L_i^I \star L_i^F$, where the union is finite, and the languages L_i^I and L_i^F contain, respectively, integer and fractional parts of the encodings of the elements of S. This induces a decomposition of S into a finite union $\bigcup_i (S_i^I + S_i^F)$, where for each i, we have $S_i^I \subseteq \mathbb{Z}$ and $S_i^F \subseteq [0, 1]$. It has been shown [BB07] that this decomposition is independent from the encoding base.

Besides, every set S_i^I and S_i^F is recognizable by the same type of automaton as $\mathcal{A}$.

Assume now that $S \subseteq \mathbb{R}$ is simultaneously (resp. weakly) r- and s-recognizable, with respect to bases r and s that are multiplicatively independent. By Cobham's theorem [Cob69], each set S_i^I is thus definable in $\langle \mathbb{Z}, +, < \rangle$. This reduces the problem of establishing that S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ to the same problem for each set S_i^F. Since we have $S_i^F \subseteq [0, 1]$ for all i, the problem has thus been reduced from the domain $\mathbb{R}$ to the interval $[0, 1]$.

3.2 Boundary points

The following notions are adapted from [BB07]. Given a point $x \in \mathbb{R}$ and a value $\varepsilon > 0$, a *neighborhood* of x is the set $N_\varepsilon(x) = \{y \in \mathbb{R} \mid |x - y| < \varepsilon\}$. A point $x \in \mathbb{R}$ is a *boundary point* of a set $S \subseteq \mathbb{R}$ iff all its neighborhoods contain at least one point from S as well as from its complement $\overline{S} = \mathbb{R} \setminus S$.

Lemma 1. *If a (resp. weakly) r-recognizable set $S \subseteq \mathbb{R}$ has only finitely many boundary points, then it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

Proof sketch. If $S \subseteq \mathbb{R}$ has only finitely many boundary points, then it can be decomposed into a finite union of intervals. In order to prove that S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, it is sufficient to show that the extremities of these intervals are rational numbers. Since S is (resp. weakly) r-recognizable, it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, <, X_r \rangle$, and so is the set S' containing only those interval extremities. The set S' is thus finite and r-recognizable, and its elements are encoded by words sharing a finite number of fractional parts. These are necessarily ultimately periodic, from which the elements of S' are rational. □

4 Multiplicatively independent bases

Let $r, s \in \mathbb{N}_{>1}$ be two *multiplicatively independent* bases, i.e., such that $r^p \neq s^q$ for all $p, q \in \mathbb{N}_{>0}$. We consider a set $S \subseteq [0, 1]$ that is both (resp. weakly) r- and s-recognizable. In the next section, we derive some properties under the assumption that S has infinitely many boundary points. We will see that this assumption leads to a contradiction in the case of weak RNA, showing that S is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ by Lemma 1. This will be no longer true for RNA.

4.1 Product stability

Let $\mathcal{A}_r$ be a (resp. weak) RNA recognizing S in base r. We assume w.l.o.g. that the transition relation of $\mathcal{A}_r$ is complete.

Since S is (resp. weakly) r-recognizable, it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, <, X_r \rangle$, and so is the set B_S of all boundary points of S, which is thus r-recognizable. Let $\mathcal{A}_r^B$ be a RNA recognizing B_S.

By hypothesis, S has infinitely many boundary points, hence there exist infinitely many distinct paths of $\mathcal{A}_r^B$ that end up cycling in the same set of

accepting states. One can thus extract from $\mathcal{A}_r^B$ an infinite language $L = 0 \star uv^*tw^\omega$, where $t, u, v, w \in \Sigma_r^*$, $|v| > 0$, $|w| > 0$, and L encodes an infinite subset of the boundary points of S. We then define $y = [0 \star uv^\omega]_r$ and, for each $k \in \mathbb{N}_{>0}$, $y_k = [0 \star uv^k tw^\omega]_r$. The sequence $y_1, y_2, y_3, \ldots \in \mathbb{Q}^\omega$ forms an infinite sequence of distinct boundary points of S, converging towards $y \in \mathbb{Q}$. If we have $y_k > y$ for infinitely many k, then we define $S^1 = (S - y) \cap [0, 1]$. Otherwise, we define $S^1 = (-S + y) \cap [0, 1]$. From Theorem 1, the set S^1 is both (resp. weakly) r- and s-recognizable. Moreover, this set admits an infinite sequence of distinct boundary points that converges to 0.

Let $\mathcal{A}_r^1$ and $\mathcal{A}_s^1$ be (resp. weak) RNA recognizing S^1 in the respective bases r and s. The path π_0 of $\mathcal{A}_r^1$ that reads $0 \star 0^\omega$ is composed of a prefix labeled by $0\star$, followed by an acyclic path of length $p \geq 0$, and finally by a cycle of length $q > 0$. It follows that a word of the form $0 \star 0^p t$, with $t \in \Sigma_r^\omega$, is accepted by $\mathcal{A}_r^1$ iff the word $0 \star 0^{p+q} t$ is accepted as well. Remark that the set S^1 admits infinitely many boundary points with a base-r encoding beginning with $0 \star 0^p$. Similar properties hold for $\mathcal{A}_s^1$. In this automaton, the path π_0' recognizing $0 \star 0^\omega$ reads the symbols 0 and $\star$, and then follows an acyclic sequence of length p' before reaching a cycle of length q'.

We now define $S^2 = r^p S^1 \cap [0, 1]$. Like S^1, the set S^2 admits an infinite sequence of boundary points that converges to 0. Moreover, by Theorem 1, S^2 is both (resp. weakly) r- and s-recognizable. Let $\mathcal{A}_r^2$ be a (resp. weak) RNA recognizing S^2 in base r. For every $t \in \Sigma_r^\omega$, the word $0 \star t$ is accepted by $\mathcal{A}_r^2$ iff the word $0 \star 0^q t$ is accepted as well. In other words, the fact that a number $x \in [0, 1]$ belongs or not to S^2 is not influenced by the insertion of q zero digits in its encodings, immediately after the symbol $\star$. This amounts to dividing the value of x by r^q, which leads to the following definition.

Definition 1. *Let $D \subseteq \mathbb{R}$ be a domain, and let $f \in \mathbb{R}_{>0}$. A set $S \subseteq D$ is* f-product-stable *in the domain D iff for all $x \in D$ such that $fx \in D$, we have $x \in S \Leftrightarrow fx \in S$.*

From the previous discussion, we have that S^2 is r^q-product-stable in $[0, 1]$. We then define $S^3 = s^{p'} S^2 \cap [0, 1]$. The set S^3 is r^q-product-stable in $[0, 1]$ as well. By Theorem 1, S^3 is also both (resp. weakly) r- and s-recognizable. Besides, since $S^3 = r^p s^{p'} S^1 \cap [0, 1]$, the set S^3 can alternatively be obtained by first defining $S^4 = s^{p'} S^1 \cap [0, 1]$, which is both (resp. weakly) r- and s-recognizable by Theorem 1. Then, one has $S^3 = r^p S^4 \cap [0, 1]$. By a similar reasoning in base s, we get that S^3 is $s^{q'}$-product-stable in $[0, 1]$. Like S^2, the set S^3 admits an infinite sequence of distinct boundary points that converges to 0.

Finally, we replace the bases r and s by $r' = r^q$ and $s' = s^{q'}$, thanks to Theorem 2. The results of this section are then summarized by the following lemma.

Lemma 2. *Let r and s be two multiplicatively independent bases, and let $S \subseteq [0, 1]$ be a set that is both (resp. weakly) r- and s-recognizable, and that admits infinitely many boundary points. There exist powers $r' = r^i$ and $s' = s^j$ of r and s, with $i, j \in \mathbb{N}_{>0}$, and a set $S' \subseteq [0, 1]$ that is both (resp. weakly) r'- and*

s'-recognizable, both r'- and s'-product-stable in $[0,1]$, and that admits infinitely many boundary points.

4.2 Recognizability by weak RNA

We are now ready to prove that the sets $S \subseteq [0,1]$ that are recognizable by weak RNA in two multiplicatively independent bases r and s can only have finitely many boundary points.

By contradiction, suppose that such a set S has infinitely many boundary points. By Lemma 2, we can assume w.l.o.g. that S is r- and s-product-stable in $[0,1]$.

Hence, there exist $\alpha, \beta \in (0,1]$ such that $\alpha \in S$ and $\beta \notin S$. For every $i, j \in \mathbb{Z}$ such that $r^i s^j \alpha \in (0,1]$, we thus have $r^i s^j \alpha \in S$. Similarly, for every $i, j \in \mathbb{Z}$ such that $r^i s^j \beta \in (0,1]$, we have $r^i s^j \beta \notin S$.

Let γ be an arbitrary point in the open interval $(0,1)$. Since r and l are multiplicatively independent, it follows from Kronecker's approximation theorem [HW85] that any open interval of $\mathbb{R}_{>0}$ contains some number of the form r^i/s^j with $i, j \in \mathbb{N}_{>0}$ [Per90]. Hence, for every sufficiently small $\varepsilon > 0$ and $\delta \in \{\alpha, \beta\}$, there exist $i, j \in \mathbb{N}_{>0}$ such that

$$0 < \gamma - \varepsilon < (r^i/s^j)\delta < \gamma + \varepsilon < 1$$

showing that every sufficiently small neighborhood $N_\varepsilon(\gamma)$ of γ contains one point from S as well as from $\overline{S}$. The latter property leads to a contradiction, since it implies that S satisfies the dense oscillating property, and therefore cannot be recognized by a weak RNA.

Taking into account the problem reductions introduced in Sections 3.1 and 3.2, we thus have established the following result, that fully generalizes Cobham's theorem to weak RNA.

Theorem 3. *Let r and s be two multiplicatively independent bases. If a set $S \subseteq \mathbb{R}$ is weakly r- and s-recognizable, then it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

Thanks to the above mentioned reductions, we can rephrase this theorem as follows. If a set $S \subseteq \mathbb{R}$ is weakly r- and s-recognizable in two multiplicatively independent bases, then it is a finite union $\bigcup_i (S_i^I + S_i^F)$, where each $S_i^I \subseteq \mathbb{Z}$ is ultimately periodic and each $S_i^F \subseteq [0,1]$ is a finite union of intervals with rational extremities. It is worth mentioning that, as observed in [Wei99], such a structural description of subsets S of $\mathbb{R}$ is equivalent to the definability of S in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.

4.3 Recognizability by RNA

Theorem 3 cannot be directly generalized to automata that are not restricted to be weak and deterministic. Indeed, with RNA, a set can be recognizable in two multiplicatively independent bases without being definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. This property is established by the following theorem.

Theorem 4. *For every pair of bases r and s that share the same set of prime factors, there exists a set S that is both r- and s-recognizable, and that is not definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

Proof sketch. A counterexample is provided by the set $S = \{n/(f_1^{i_1} f_2^{i_2} \cdots f_k^{i_k}) \mid n \in \mathbb{Z}, i_1, i_2, \ldots, i_k \in \mathbb{N}\}$, where $f_1, f_2, \ldots f_k$ are the prime factors of r and s. Indeed, in either base $t \in \{r, s\}$, this set is encoded by the language $L_t = \{0, t-1\}\Sigma_t^* \star \Sigma_t^*(0^\omega \cup (t-1)^\omega)$. This language is clearly ω-regular, hence S is both r- and s-recognizable. However, S satisfies the dense oscillating property, which prevents it from being recognized by a weak RNA. It follows that S is not definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. □

The case of bases that do not share the same set of prime factors is investigated in the next section.

5 Bases with different sets of prime factors

We now consider a subset of $[0, 1]$ that is recognizable by RNA in two bases that have different sets of prime factors. Recall that according to Lemma 1, in order to prove that the set is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$, it is sufficient to show that this set has only finitely many boundary points. Like in Section 4, we proceed by contradiction, and assume that the set has infinitely many boundary points. By Lemma 2, there exist bases r and s with different sets of prime factors, and a set $S \subseteq [0, 1]$ that is both r- and s-recognizable, both r- and s-product-stable in $[0, 1]$, and that has infinitely many boundary points. Without loss of generality, we assume that there is a prime factor of s that does not divide r.

5.1 Sum stability

Our strategy consists in exploiting Cobham's theorem so as to derive additional properties of S. The first step is to build from S a set $S' \subseteq \mathbb{R}_{\geq 0}$ that coincides with S over $[0, 1]$, shares the same recognizability and product-stability properties, and contains numbers with non-trivial integer parts.

Lemma 3. *Let $r, s \in \mathbb{N}_{>1}$ be two bases with different sets of prime factors, and let $S \subseteq [0, 1]$ be a set that is r- and s-recognizable, r- and s-product-stable in $[0, 1]$, and that has infinitely many boundary points. There exists a set $S' \subseteq \mathbb{R}_{\geq 0}$ that is r- and s-recognizable, r- and s-product-stable in $\mathbb{R}_{\geq 0}$, and that has infinitely many boundary points.*

Proof sketch. Let $S' = \{r^k x \mid x \in S \wedge k \in \mathbb{N}\}$. This set is clearly r-product-stable in $\mathbb{R}_{\geq 0}$. Since S is r-product-stable in $[0, 1]$, we have $S' \cap [0, 1] = S$ showing that S' has infinitely many boundary points. We build a RNA $\mathcal{A}'_r$ recognizing S' in base r from a RNA $\mathcal{A}_r$ recognizing S as follows. The automaton $\mathcal{A}'_r$ is similar to $\mathcal{A}_r$, except that it delays arbitrarily the reading of the symbol $\star$.

In order to prove that S' is s-recognizable, notice that, since S is both r- and s-product-stable in $[0, 1]$, we have $S' = \{r^i s^j x \mid x \in S \wedge i, j \in \mathbb{Z}\}$. The set S' can

therefore be expressed as $S' = \{s^k x \mid x \in S \wedge k \in \mathbb{N}\}$. By the same reasoning as in base r, this set is s-recognizable, and s-product-stable in $\mathbb{R}_{\geq 0}$. □

Consider now a set S' obtained from S by Lemma 3. As discussed in Section 3.1, this set can be expressed as a finite union $S' = \bigcup_i (S_i^I + S_i^F)$, where for each i, we have $S_i^I \subseteq \mathbb{N}$ and $S_i^F \subseteq [0,1]$. Moreover, for each i, the set S_i^I is both r- and s-recognizable, and it follows from Cobham's theorem that this set is definable in $\langle \mathbb{N}, +, < \rangle$. Since such a set is ultimately periodic, there exists $n_i \in \mathbb{N}_{>0}$ for which $\forall x \in \mathbb{N}, x \geq n_i : x \in S_i^I \Leftrightarrow x + n_i \in S_i^I$. By defining $n = \operatorname{lcm}_i(n_i)$, we have $\forall x \in \mathbb{R}_{\geq 0}, x \geq n : x \in S' \Leftrightarrow x + n \in S'$. This prompts the following definition.

Definition 2. *Let $D \subseteq \mathbb{R}$ be a domain, and let $t \in \mathbb{R}$. A set $S \subseteq D$ is* t-sum-stable *in D iff for all $x \in D$ such that $x + t \in D$, we have $x \in S \Leftrightarrow x + t \in S$.*

Let us show that the set $S'' = (1/n)S'$ is 1-sum-stable in $\mathbb{R}_{>0}$. For every $x \geq 1$, we have $x \in S'' \Leftrightarrow x + 1 \in S''$. For $x < 1$, we choose $k \in \mathbb{N}$ such that $r^k x \geq 1$. Exploiting the properties of S' (transposed to S''), we get $x \in S'' \Leftrightarrow r^k x \in S'' \Leftrightarrow r^k x + r^k \in S'' \Leftrightarrow x + 1 \in S''$. Lemma 3 can thus be refined as follows.

Lemma 4. *Let $r, s \in \mathbb{N}_{>1}$ be two bases with different sets of prime factors, and let $S \subseteq [0,1]$ be a set that is r- and s-recognizable, r- and s-product-stable in $[0,1]$, and that has infinitely many boundary points. There exists a set $S' \subseteq \mathbb{R}_{>0}$ that is r- and s-recognizable, has infinitely many boundary points, and is r-product-, s-product- and 1-sum-stable in $\mathbb{R}_{>0}$.*

Note that Lemmas 3 and 4 still hold if the bases r and s are multiplicatively independent.

5.2 Exploiting sum-stability properties

Consider a set $S' \subseteq \mathbb{R}_{>0}$ that satisfies the properties expressed by Lemma 4. It remains to show that these properties lead to a contradiction. The hypothesis on the prime factors of r and s is explicitly used in this section.

We proceed by characterizing the numbers $t \in \mathbb{R}$ for which S' is t-sum-stable in $\mathbb{R}_{>0}$. These form the set $T_{S'} = \{t \in \mathbb{R} \mid \forall x \in \mathbb{R}_{>0} : x + t \in \mathbb{R}_{>0} \Rightarrow (x \in S' \Leftrightarrow x + t \in S')\}$. Since S' is r-recognizable, it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, <, X_r \rangle$, and so is $T_{S'}$, that is therefore r-recognizable as well.

The set $T_{S'}$ enjoys interesting closure properties:

Property 1. *For every $t, u \in T_{S'}$ and $a, b \in \mathbb{Z}$, we have $at + bu \in T_{S'}$.*

The set $T_{S'}$ is also r- and s-product stable in $\mathbb{R}$. Since $1 \in T_{S'}$, this yields the following property.

Property 2. *For every $k \in \mathbb{Z}$, we have $r^k \in T_{S'}$ and $s^k \in T_{S'}$.*

Intuitively, being able to add or subtract r^k from a number, for any k, makes it possible to change in an arbitrary way finitely many digits in its base-r encodings, without influencing the fact that this number belongs or not to S'. Our next step will be to show that this property can be extended to all digits of base-r encodings, implying either $S' = \emptyset$ or $S' = \mathbb{R}_{>0}$. This would then contradict our assumption that S' has infinitely many boundary points.

Property 3. *There exist $l, m \in \mathbb{N}_{>0}$ such that, for every $k \in \mathbb{N}_{>0}$, we have*

$$m/(r^{lk} - 1) \in T_{S'}.$$

Proof. By Property 2, we have $1/s^k \in T_{S'}$ for all $k \in \mathbb{N}$. The base-r encodings of $1/s^k$ are of the form $0^+ \star v_k u_k^\omega$, where u_k is their *period*. Hence, $1/s^k = a_k/(r^{|v_k|}(r^{|u_k|} - 1))$, with $a_k \in \mathbb{N}_{>0}$. Recall that, by hypothesis, there exists a prime factor f of s that does not divide r. Thus f^k must divide $r^{|u_k|} - 1$. It follows that the length of the periods u_k must be unbounded w.r.t. k.

Consider a RNA $\mathcal{A}_r^T$ recognizing $T_{S'}$ in base r. We study the rational numbers accepted by $\mathcal{A}_r^T$, which have base-r encodings of the form $v \star wu^\omega$. We assume w.l.o.g. that the considered periods u are the shortest possible ones. It follows from the unboundedness of u_k that $T_{S'}$ contains rational numbers with infinitely many distinct periods. As a consequence, there exist u, u', v, v', w, w' such that u^ω is not a suffix of $(u')^\omega$, the words $v \star wu^\omega$ and $v' \star w'(u')^\omega$ are both accepted by $\mathcal{A}_r^T$, and the paths π and π' of $\mathcal{A}_r^T$ reading them end up cycling in exactly the same subset of accepting states. (Recall that RNA are deterministic Muller automata.)

Let q be one of these states, and $u_1, u_2 \in \Sigma_r^+$ be periods of the (respective) words read by π and π' after reaching q in their final cycle. These periods can be repeated arbitrarily, hence we can assume w.l.o.g. that $|u_1| = |u_2|$. Moreover we can assume w.l.o.g. that $[u_2]_r > [u_1]_r$, otherwise u^ω would be a suffix of $(u')^\omega$. Besides, there exist $v, w \in \Sigma_r^*$ such that $v \star w$ reaches q. From the structure of $\mathcal{A}_r^T$, it follows that for every $k \geq 0$, the word $v \star w(u_1^k u_2)^\omega$ is accepted by $\mathcal{A}_r^T$.

For each $k \geq 0$, we thus have $[v \star w(u_1^k u_2)^\omega]_r \in T_{S'}$. Developing, we get $d_k/r^{|w|} + [vw \star 0^\omega]_r/r^{|w|} \in T_{S'}$, with $d_k = [\star(u_1^k u_2)^\omega]_r$. Thanks to Properties 1 and 2, and the r-product-stability property of $T_{S'}$, this implies $d_k \in T_{S'}$. We now express d_k in terms of $[u_1]_r$, $[u_2]_r$, and k:

$$d_k = \frac{[u_1^k u_2]_r}{r^{l(k+1)} - 1} = \frac{[u_2]_r - [u_1]_r}{r^{l(k+1)} - 1} + \frac{[u_1]_r}{r^l - 1}, \text{ where } l = |u_1| = |u_2|.$$

The next step will consist in getting rid of the second term of this expression. By Properties 1 and 2, we have for all $k \in \mathbb{N}$,

$$(r^l - 1)d_k - [u_1]_r = \frac{m}{r^{l(k+1)} - 1} \in T_{S'},$$

where $m = (r^l - 1)([u_2]_r - [u_1]_r)$ is such that $m \in \mathbb{N}_{>0}$. For all $k > 0$, we thus have $m/(r^{lk} - 1) \in T_{S'}$. □

We are now ready to conclude. Given l and m by Property 3, we define $S'' = (1/m)S'$. Like S', this set has infinitely many boundary points. The set $T_{S''}$ of the values t for which S'' is t-sum-stable in $\mathbb{R}_{>0}$ is given by $T_{S''} = (1/m)T_{S'}$. This set is thus r-recognizable. From Properties 1 and 2, we have for every $k \in \mathbb{N}$, $1/r^k \in T_{S''}$. Finally, from Property 3, we have for every $k > 0$, $1/(r^{lk}-1) \in T_{S''}$.

Property 4. *The set $T_{S''}$ is equal to $\mathbb{R}$.*

Proof. Since $T_{S''}$ and $\mathbb{R}$ are both r-recognizable, and two ω-regular languages are equal iff they share the same subset of ultimately periodic words [PP04], it is actually sufficient to show $T_{S''} \cap \mathbb{Q} = \mathbb{Q}$. Every rational t admits a base-r encoding of the form $v \star wu^\omega$, where $|u| = lk$ for some $k \in \mathbb{N}_{>0}$. We have

$$t = \frac{[vw \star 0^\omega]_r}{r^{|w|}} + \frac{[u]_r}{r^{|w|}(r^{lk}-1)}.$$

Since $1/r^{|w|} \in T_{S''}$ and $1/(r^{lk}-1) \in T_{S''}$, the closure and product-stability properties of $T_{S''}$ imply $t \in T_{S''}$. □

As a consequence, we either have $S'' = \emptyset$ or $S'' = \mathbb{R}_{>0}$, which contradicts the hypothesis that this set has infinitely many boundary points. We thus finally have the following theorem.

Theorem 5. *Let r and s be two bases that do not share the same set of prime factors. If a set $S \subseteq \mathbb{R}$ is r- and s-recognizable, then it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$.*

6 Conclusions

In this paper, we have established that the sets of real numbers that can be recognized by finite automata in two sufficiently different bases are exactly those that are definable in the first-order additive theory of real and integer variables $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. In the case of weak deterministic automata, used in actual implementations of symbolic representation systems, the condition on the bases turns out to be multiplicative independence. It is worth mentioning that recognizability in multiplicatively dependent bases is equivalent to recognizability in one of them, and that definability in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$ implies recognizability in every base. We have thus obtained a complete characterization of the sets of numbers recognizable in multiple bases, similar to the one known for the integer domain [Cob69].

For Muller automata, we have demonstrated that multiplicative independence of the bases is not a strong enough condition, and that the bases must have different sets of prime factors in order to force definability of the represented sets in $\langle \mathbb{R}, \mathbb{Z}, +, < \rangle$. Recall that the sets definable in that theory can all be recognized by weak deterministic automata. We have thus established that the sets of real numbers that can be recognized by infinite-word automata in all encoding bases are exactly those that are recognizable by weak deterministic automata. This result provides a theoretical justification to the use of weak automata as symbolic data structures for representing sets of real and integer numbers.

References

[BB07] B. Boigelot and J. Brusten. A generalization of Cobham's theorem to automata over real numbers. In *Proc. 34th ICALP*, volume 4596 of *Lecture Notes in Computer Science*, pages 813–824, Wroclaw, July 2007. Springer.

[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.

[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.

[EK06] J. Eisinger and F. Klaedtke. Don't care words with an application to the automata-based approach for real addition. In *Proc. 18th CAV*, volume 4144 of *Lecture Notes in Computer Science*, pages 67–80, Seattle, August 2006. Springer.

[HW85] G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers*. Oxford University Press, 5th edition, 1985.

[McN66] R. McNaughton. Testing and generating infinite sequences by a finite automaton. *Information and Control*, 9(5):521–530, 1966.

[Per90] D. Perrin. Finite automata. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, pages 1–57. Elsevier and MIT Press, 1990.

[PP04] D. Perrin and J.E. Pin. *Infinite words*, volume 141 of *Pure and Applied Mathematics*. Elsevier, 2004.

[Saf88] S. Safra. On the complexity of ω-automata. In *Proc. 29th Symposium on Foundations of Computer Science*, pages 319–327. IEEE Computer Society, October 1988.

[Var07] M. Vardi. The Büchi complementation saga. In *Proc. 24th. STACS*, volume 4393 of *Lecture Notes in Computer Science*, pages 12–22, Aachen, February 2007. Springer.

[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*, pages 21–32, Glasgow, September 1995. Springer.

[Wei99] V. Weispfenning. Mixed real-integer linear quantifier elimination. In *Proc. ACM SIGSAM ISSAC*, pages 129–136, Vancouver, July 1999. ACM Press.

[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.