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

Bernard Boigelot

Université de Liège

Joint work with: Pierre Wolper, Louis Bronne, Stéphane Rassart, Sébastien Jodogne, Julien Brusten, Véronique Bruyère, Jérôme Leroux & Jean-François Degbomont.

The starting point of this research

Framework: Verification of infinite-state systems by symbolic state-space exploration.

Main issue: Developing data structures for representing finite and infinite sets of variable values.

Requirements:

- Applicable to the data domains $\mathbb{Z}^n$ and $\mathbb{R}^n$.
- Broad expressive power: linear constraints, discrete periodicities, ...
- Large set of computable operations: $\cap$, $\cup$, $\times$, $\stackrel{?}{=}$, $\stackrel{?}{\subseteq}$, ...
- Efficiency and conciseness.

Automata recognizing numbers

Automata recognizing integer numbers are a powerful tool for

- reasoning about arithmetic theories (e.g., Presburger arithmetic), and
- representing symbolically sets of values.

Advantages:

- Large range of simple, efficient and well-studied manipulation algorithms.
- Expressive power precisely characterized, and suited for many applications.

Question: Can they be generalized into automata recognizing real numbers?

Encoding real numbers

- Real numbers are encoded as infinite words over r digits $0, 1, \ldots, r-1$, and a separator $\star$.

  Example: $Enc_2(3.5) = 0^{+}11 \star 1(0)^{\omega} \cup 0^{+}11 \star 0(1)^{\omega}$.

- Negative numbers are encoded by their r's-complement:
  - Encodings of an integer part $x_I \in \mathbb{Z}_{<0}$ correspond to the last p digits of the encodings of $r^p + x_I$, with $p > 0$.
  - The integer-part length p must be large enough to satisfy

$$-r^{p-1} \leq x_I < r^{p-1}.$$

  - The first digit is thus equal to 0 for positive (or zero) integer parts, and to $r-1$ for negative ones (*sign digit*).

  Example: $Enc_2(-3.5) = 1^{+}00 \star 1(0)^{\omega} \cup 1^{+}00 \star 0(1)^{\omega}$.

Automata recognizing real vectors

- Vectors from $\mathbb{R}^n$ are encoded as infinite words over either
  - $\{0, 1, \ldots, r-1\}^n \cup \{\star\}$ (*synchronous encoding*).

    Example: $Enc_2\left(\left[\begin{array}{r}\frac{1}{3}\\ -\frac{1}{3}\end{array}\right]\right) = \begin{pmatrix}0\\1\end{pmatrix}^+ \star \left[\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\right]^\omega,$

  - or $\{0, 1, \ldots, r-1, \star\}$ (*serial encoding*).

    Example: $Enc_2\left(\left[\begin{array}{r}\frac{1}{3}\\ -\frac{1}{3}\end{array}\right]\right) = (01)^+ \star (0110)^\omega.$

- A Real Vector Automaton (RVA) representing a set $S \subseteq \mathbb{R}^n$ is a Büchi (or deterministic Muller) automaton accepting all the base-r encodings of all the vectors in S.

  A set $S \subseteq \mathbb{R}^n$ is r-recognizable if it can be recognized by some RVA in base r.

The expressive power of number automata

Theorem [Büchi, 60; Bruyère, 85]: The r-recognizable subsets of $\mathbb{Z}^n$ are those that are definable in the first-order theory $\langle \mathbb{Z}, +, \leq, V_r \rangle$, where V_r is a base-dependent function (returning the highest power of r that divides its argument).

Theorem [B., Rassart, Wolper, 98]: The r-recognizable subsets of $\mathbb{R}^n$ are those that are definable in the first-order theory $\langle \mathbb{R}, \mathbb{Z}, +, \leq, X_r \rangle$, where X_r is a base-dependent predicate.

Proof ideas:

- *Definable sets are recognizable:*
  - Atomic formulas can directly be translated into automata.
  - Operators can be applied algorithmically in order to build automata from more complex formulas.

- *Recognizable sets are definable:*
  - The states of a given automaton can be encoded as vectors of base-r digits.
  - This provides a way of encoding sequences of states as real vectors.
  - Thanks to the base-dependent predicate X_r, one can build a formula that checks whether a run accepts a given vector.
  - By quantifying over the run representations, one gets a formula that defines the set of vectors recognized by the automaton.

Notes:

- The theory $\langle \mathbb{R}, \mathbb{Z}, +, \leq, X_r \rangle$ includes $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$, which extends Presburger arithmetic to mixed integer and real variables.
- RVA provide a decision procedure for this arithmetic.

Recognizing numbers in multiple bases

Theorem [Cobham, 69]: The subsets of $\mathbb{Z}$ that are simultaneously recognizable in two multiplicatively independent bases are exactly those that are definable in the first-order theory $\langle \mathbb{Z}, +, \leq \rangle$ (i.e., Presburger arithmetic).

Theorem [Semenov, 77]: The subsets of $\mathbb{Z}^n$ that are simultaneously recognizable in two multiplicatively independent bases are exactly those that are Presburger-definable.

Questions:

- Can those theorems be generalized to the real domain?
- Are RVA really usable in practice?

Weak RVA

Definitions:

- A Büchi automaton is weak if each strongly connected component of its transition graph contains only either accepting or non-accepting states.
- A set that can be recognized by a weak deterministic RVA in a base $r > 1$ is said to be weakly r-recognizable.

Advantages of weak deterministic RVA:

- Simple and efficient manipulation algorithms.
- Existence of an easily computable canonical form.

Theorem [B., Jodogne, Wolper, 01]: Every subset of $\mathbb{R}^n$ that is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ is weakly recognizable in every base $r > 1$.

Proof sketch:

- Define two metric topologies over resp. vectors and words:
  - $dist(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$.
  - $dist(w, w') = \dfrac{1}{|\mathit{commonprefix}(w, w')| + 1}$.
- In each topology, consider the class $F_\sigma \cap G_\delta$ of sets that can be expressed both as countable unions of closed sets and countable intersections of open sets.
- The subsets of $\mathbb{R}^n$ definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ belong to $F_\sigma \cap G_\delta$ (in the vector topology):
  - Such sets can be expressed as countable unions of sets defined in $\langle \mathbb{R}, +, \leq, 1 \rangle$.
  - This theory admits quantifier elimination, and thus defines sets that are finite Boolean combinations of open and closed sets.

- The sets definable in $\langle \mathbb{R}, +, \leq, 1 \rangle$ thus belong to F_σ, hence their countable unions as well.
- Since $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ is closed under complementation, the sets definable in that theory also belong to G_δ.

- The encodings of sets that belong to $F_\sigma \cap G_\delta$ (in the vector topology) form languages that belong to $F_\sigma \cap G_\delta$ (in the word topology):
  - The open and closed subsets of $\mathbb{R}^n$ are both encoded by languages that are finite Boolean combinations of open and closed sets, hence belong to $F_\sigma \cap G_\delta$.
  - Countable unions and intersections of such sets therefore belong to $F_\sigma \cap G_\delta$ as well.

- The ω-regular languages that belong to $F_\sigma \cap G_\delta$ correspond to those that can be accepted by weak deterministic automata.

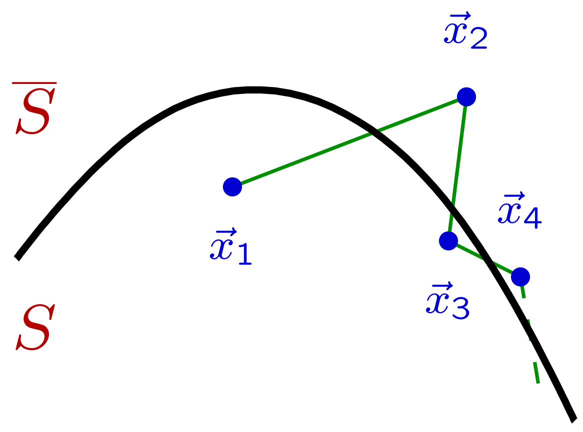
Beyond weak recognizability

Definition: A set $S \subseteq \mathbb{R}^n$ satisfies the dense oscillating property if

$$\exists \vec{x}_1 : \forall \varepsilon_1 : \exists \vec{x}_2 : \forall \varepsilon_2 : \exists \vec{x}_3 : \forall \varepsilon_3 : \cdots$$

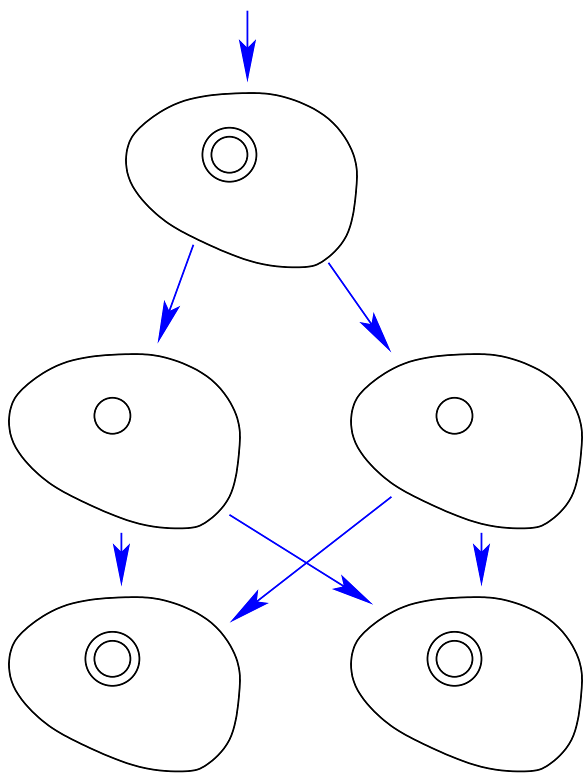
such that

- $dist(\vec{x}_i, \vec{x}_{i+1}) \leq \varepsilon_i$ for all i,
- $\vec{x}_i \in S$ for all odd i, and
- $\vec{x}_i \notin S$ for all even i.



Theorem: The r-recognizable subsets of $\mathbb{R}^n$ that satisfy the dense oscillating property are not weakly r-recognizable.

Proof intuition: In order to recognize such a set, a weak automaton would need to switch infinitely many times between its accepting and non-accepting components, which is not possible.



Extending Cobham's theorem to subsets of $\mathbb{R}$

Cobham's theorem: A subset of $\mathbb{Z}$ is simultaneously recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{Z}, +, \leq \rangle$.

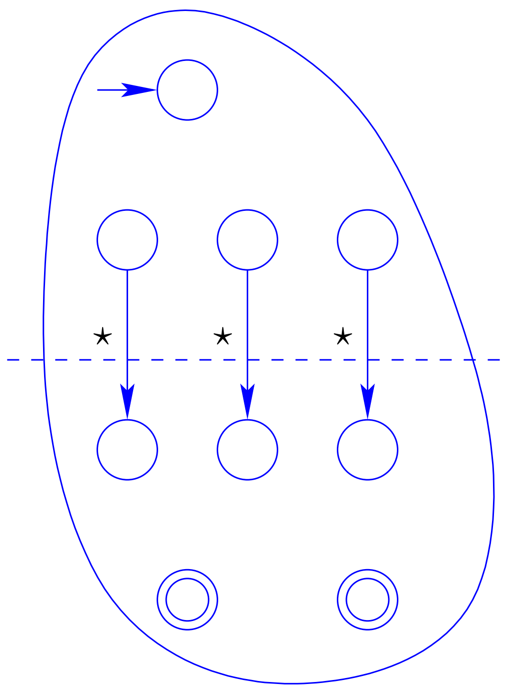
This work:

- A subset of $\mathbb{R}$ is simultaneously weakly recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ [B., Brusten, 07].
- A subset of $\mathbb{R}$ is simultaneously recognizable in two bases with different sets of prime factors iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ [B., Brusten, Bruyère, 08].

Step 1: Reduction to $[0,1]$

Let $S \subseteq \mathbb{R}$ be recognizable (resp. weakly recognizable).

We have $S = \bigcup_{i=1\ldots m} S^i_{\mathit{Int}} + S^i_{\mathit{Fract}}$, with $\forall i : S^i_{\mathit{Int}} \subseteq \mathbb{Z}$, $S^i_{\mathit{Fract}} \subseteq [0,1]$.



Note: The decomposition of S into $S^1_{\mathit{Int}}, S^2_{\mathit{Int}}, \ldots, S^1_{\mathit{Fract}}, S^2_{\mathit{Fract}}, \ldots$ does not depend on the representation base.

Proof:

- Given S and $x_I \in \mathbb{Z}$, we define

$$S_{\mathit{Fract}}(x_I) = \{x_F \in [0,1] \mid x_I + x_F \in S\}.$$

- Since S is recognized by a finite automaton, the set

$$\{S_{\mathit{Fract}}(x_I) \mid x_I \in \mathbb{Z}\} = \{S^1_{\mathit{Fract}}, S^2_{\mathit{Fract}}, \ldots\}$$

  is finite.

- For every element S^k_{Fract} in this set, one has

$$S^k_{\mathit{Int}} = \{x_I \in \mathbb{Z} \mid S_{\mathit{Fract}}(x_I) = S^k_{\mathit{Fract}}\}.$$

- Those definitions are independent from r.

Corollary: The set $S \subseteq \mathbb{R}$ is recognizable (resp. weakly recognizable) in two bases $r > 1$ and $s > 1$ iff for all $i \in \{1, \ldots, m\}$, the sets $S^i_{Int} \subseteq \mathbb{Z}$ and $S^i_{Fract} \subseteq [0, 1]$ are recognizable (resp. weakly recognizable) as well in these two bases r and s.

By Cobham's theorem, if a set $S^i_{Int} \subseteq \mathbb{Z}$ is recognizable in two multiplicatively independent bases, then it is definable in $\langle \mathbb{Z}, +, \leq \rangle$.

Thus, it is sufficient to prove that:

- a subset of $[0, 1]$ is simultaneously weakly recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{R}, +, \leq, 1 \rangle$.

- a subset of $[0, 1]$ is simultaneously recognizable in two bases with different sets of prime factors iff it is definable in $\langle \mathbb{R}, +, \leq, 1 \rangle$.

Step 2: Characterize sets by their boundary points

Definition: 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
- at least one point from $\mathbb{R} \setminus S$.

Lemma: If a r-recognizable set $S \subseteq \mathbb{R}$ has finitely many boundary points, then it can be defined in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$.

Proof:

- The set of boundary points of S is itself r-recognizable.
- This set of boundary points is finite, hence it only contains rational numbers.
- The set S can thus be expressed as a finite union of intervals with rational boundaries.

Strategy: Prove that the hypotheses

- $S \subseteq [0, 1]$,
- S weakly recognizable in two multiplicatively independent bases r and s, and
- S has infinitely many boundary points

lead to a contradiction.

Proof sketch:

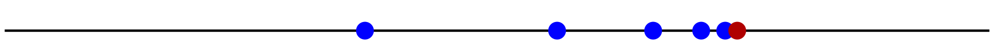
1. Build a (deterministic Muller) RVA $\mathcal{A}_r^B$ that recognizes the boundary points of S in base r.

2. By a pumping lemma, extract from $L(\mathcal{A}_r^B)$ a sequence of words

$$0 \star u\, t\, w^\omega,$$
$$0 \star u\, v\, t\, w^\omega,$$
$$0 \star u\, v^2\, t\, w^\omega,$$
$$0 \star u\, v^3\, t\, w^\omega,$$
$$\vdots$$

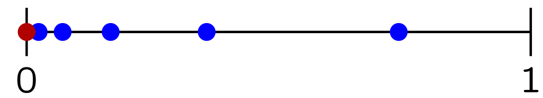
   that converges towards $0 \star uv^\omega$.

3. These words encode a sequence of distinct boundary points of S that converges towards some rational number.

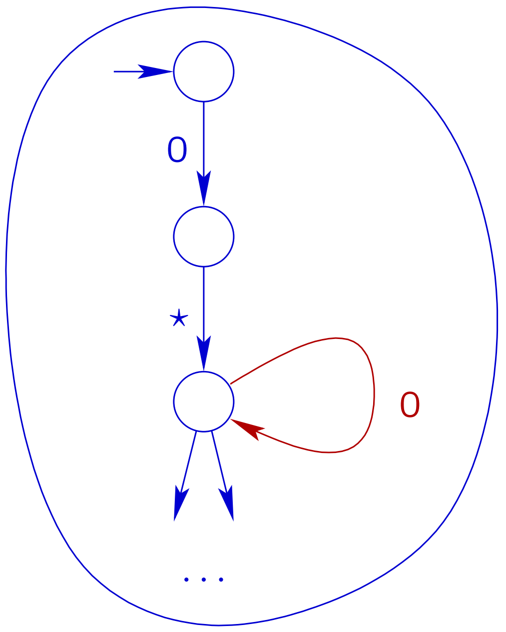


4. By scaling and translating S, one can replace it by a subset of $[0,1]$ that

- admits an infinite sequence of distinct boundary points converging towards 0,



- is recognized in two multiplicatively independent bases r and s by weak automata of the following form:



5. The corresponding algebraic property is that S becomes both r-product-stable and s-product-stable:

   Within $[0, 1]$:

$$\forall x : x \in S \Leftrightarrow rx \in S \Leftrightarrow sx \in S.$$

6. From this property, by Kronecker's approximation lemma, there exist boundary points of S in every interval of non-zero length.

7. Hence, S satisfies the dense oscillating property, and cannot be weakly recognizable.

   $\Rightarrow$ contradiction!

Summary (so far)

We have established the following result:

Theorem: A subset of $\mathbb{R}$ is simultaneously weakly recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$.

Question: Does this theorem extend to (strong) recognizability?

No!

Counterexample: Let $r > 1$ and $s > 1$ be multiplicatively independent bases that share the same prime factors f_1, f_2, ..., f_k.

The set

$$\left\{ \frac{n}{f_1^{i_1} f_2^{i_2} \cdots f_k^{i_k}} \mid n \in \mathbb{Z},\ i_1, i_2, \ldots, i_k \in \mathbb{N} \right\}$$

is

- recognizable in both bases r and s, and
- not definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$:
  - Otherwise, this set would be weakly recognizable in every base.
  - This would contradict the fact that the dense oscillating property is satisfied.

Next goal: Prove that

> a set $S \subseteq [0,1]$ is simultaneously recognizable in two bases r and s with different sets of prime factors iff it is definable in $\langle \mathbb{R}, +, \leq, 1 \rangle$.

Problem: With (strong) reachability, one cannot exploit anymore the dense oscillating property . . .

Sketch of the approach:

1. Assume, by contradiction, that S has infinitely many boundary points.
2. By similar techniques as for weak recognizability, replace S by a subset of $[0,1]$ that
   - is both r- and s-recognizable,
   - has infinitely many boundary points,
   - is product stable with respect to both r and s.

3. We now aim at exploiting Cobham's theorem in order to obtain additional properties. We replace S by the set

$$\{r^k x \mid x \in S \wedge k \in \mathbb{N}\}$$

that

- is both r- and s-recognizable,
- has infinitely many boundary points,
- is product stable in $\mathbb{R}_{\geq 0}$ with respect to both r and s.

4. Recall that S can be decomposed into a finite union

$$S = \bigcup_{i=1\ldots m} S^i_{\mathit{Int}} + S^i_{\mathit{Fract}},$$

where for each i, $S^i_{\mathit{Int}} \subseteq \mathbb{Z}$ and $S^i_{\mathit{Fract}} \subseteq [0,1]$ are both r- and s-recognizable.

5. By Cobham's theorem, each set S^i_{Int} is definable in $\langle \mathbb{Z}, +, \leq \rangle$. Thus, there exists $n \in \mathbb{N}_{>0}$ such that

$$\forall x \in \mathbb{N},\, x \geq n : x \in S^i_{Int} \Leftrightarrow x + n \in S^i_{Int},$$

which implies

$$\forall x \in \mathbb{R},\, x \geq n : x \in S \Leftrightarrow x + n \in S.$$

6. We replace S by the set $\frac{1}{n} S \setminus \{0\}$ that

- is both r- and s-recognizable,
- has infinitely many boundary points,
- is product stable in $\mathbb{R}_{\geq 0}$ with respect to both r and s,
- is 1-sum-stable in $\mathbb{R}_{>0}$: Within $\mathbb{R}_{>0}$:

$$x \in S \Leftrightarrow x + 1 \in S.$$

(Indeed, if $0 < x < 1$, for $k \in \mathbb{N}$ such that $r^k \geq 1$, we get

$$x \in S \Leftrightarrow r^k x \in S \Leftrightarrow r^k x + r^k \in S \Leftrightarrow x + 1 \in S.)$$

7. Let T be the set of all $t \in \mathbb{R}$ such that S is t-sum-stable in $\mathbb{R}_{>0}$. This set

   - is both r- and s-recognizable,
   - is product stable in $\mathbb{R}$ with respect to both r and s,
   - is closed under addition,
   - contains 1.

   Corollary: For every $k \in \mathbb{Z}$, we have $r^k \in T$ and $s^k \in T$.

   Intuitive meaning: Modifying finitely many digits in base-r or base-s encodings of a number does not affect whether it belongs or not to the set S.

   But . . . we have not yet exploited the property that r and s have different sets of prime factors!

8. Assume w.l.o.g. that one of the prime factors of s does not divide r. The rational numbers

$$\frac{1}{s}, \frac{1}{s^2}, \frac{1}{s^3}, \ldots$$

- are encoded in base r by words with increasingly large periods.
- belong to T, which is r-recognizable.

By a pumping lemma (and technical developments), one eventually obtains that replacing the period of a rational number encoding by another one does not affect whether it belongs or not to S.

9. Then, one has either $\mathbb{Q}_{>0} \subseteq S$ or $\mathbb{Q}_{>0} \cap S = \emptyset$, which finally gives

$$S = \emptyset \quad \text{or} \quad S = \mathbb{R}_{>0}.$$

$\Rightarrow$ contradiction!

Summary

Theorem: A subset of $\mathbb{R}$ is simultaneously recognizable in two bases with different sets of prime factors iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$.

Corollary: Such a set is weakly recognizable in every base.

Next goal: Generalization to $\mathbb{R}^n$.

Extending Semenov's theorem to subsets of $\mathbb{R}^n$

Semenov's theorem: For every $n > 0$, a subset of $\mathbb{Z}^n$ is simultaneously recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{Z}, +, \leq \rangle$.

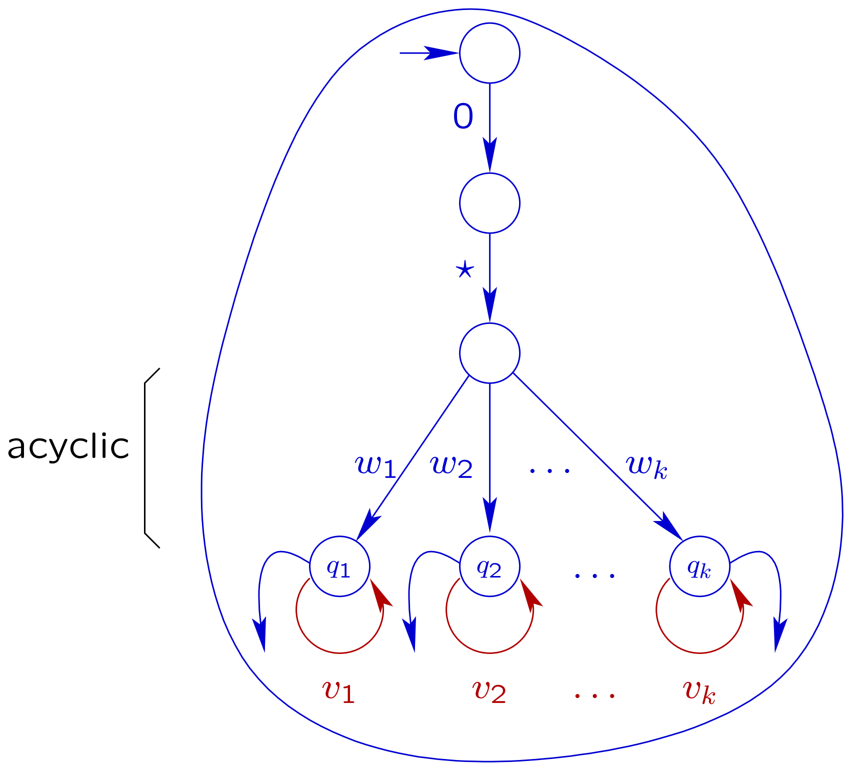
This work:

- A subset of $\mathbb{R}^n$ is simultaneously weakly recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ [B., Brusten, Leroux, 09].

- A subset of $\mathbb{R}^n$ is simultaneously recognizable in two bases with different sets of prime factors iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$ [B., Brusten, Leroux, 09].

Proof strategy:

1. Proceed by induction on n. The case $n = 1$ is already settled.
2. Assume $n > 1$. Reduce the problem to subsets of $[0, 1]^n$ (by a similar construction as for $n = 1$).
3. An RVA recognizing a set $S \subseteq [0, 1]^n$ contains the following structure:



4. This leads to a finite decomposition of S:

$$S = \bigcup_{w_i} \frac{1}{r^{|w_i|}}(S(q_i) + [w_i]_r),$$

where

- r is the representation base,
- $S(q_i)$ denotes the fractional set recognized from q_i, and
- $[w_i]_r$ denotes the natural number encoded by w_i.

$\Rightarrow$ It suffices to prove the theorems for each set $S(q_i)$.

5. Each set $S(q_i)$ is r^l-product-stable, with $l = |v_i|$, with respect to some pivot $\vec{z}_i \in [0,1]^n \cap \mathbb{Q}^n$:

$$\forall \vec{x} \in [0,1]^n - \vec{z}_i : \vec{x} \in S(q_i) - \vec{z}_i \Leftrightarrow \frac{1}{r^l}\vec{x} \in S(q_i) - \vec{z}_i.$$

6. If S is s-recognizable, then there exists $l' > 0$ such that each $S(q_i)$ is also $s^{l'}$-product-stable.

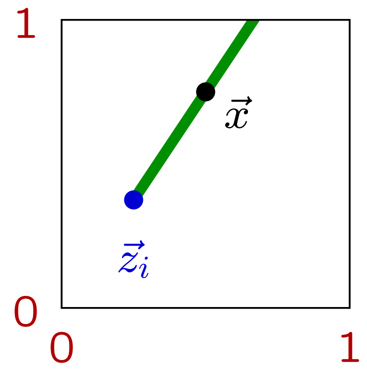
7. Assume that either

- r and s are multiplicatively independent and S is both r- and s-weakly recognizable, or
- r and s have different sets of prime factors and S is both r- and s-recognizable.

By reasoning similarly to the case $n = 1$, one obtains that for any i and $\vec{x} \in [0,1]^n \cap \mathbb{Q}^n$ such that $\vec{x} \neq \vec{z}_i$, the set

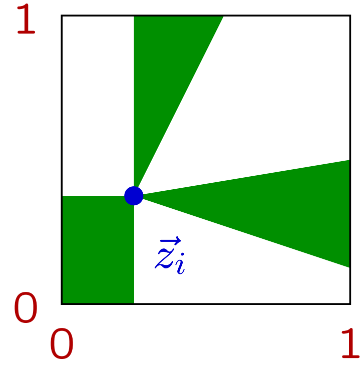
$$\vec{z}_i + \lambda(\vec{x} - \vec{z}_i) \in [0,1]^n \mid \lambda \in \mathbb{R}_{>0}$$

is either a subset of $S(q_i)$ or disjoint from $S(q_i)$.

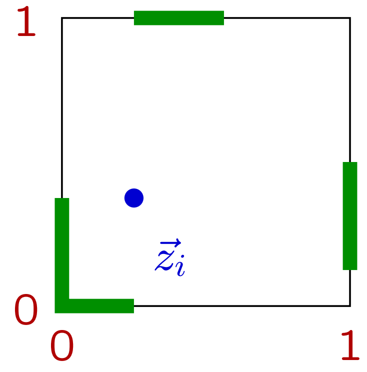


8. Thus, each set $S(q_i)$ has a conical structure centered on $\vec{z}_i$:

$$\forall \vec{x} \in [0,1]^n, \lambda \in]0,1] : \vec{x} \in S(q_i) \Leftrightarrow \lambda(\vec{x} - \vec{z}_i) + \vec{z}_i \in S(q_i).$$



9. Such a set is definable in $\langle \mathbb{R}, +, \leq, 1 \rangle$ iff its intersection with the boundaries of $[0,1]^n$ is definable in the same theory,



which follows from the $(n-1)$-dimensional case.

Summary of results

- A subset of $\mathbb{R}^n$ is simultaneously weakly recognizable in two multiplicatively independent bases iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$
- A subset of $\mathbb{R}^n$ is simultaneously recognizable in two bases with different sets of prime factors iff it is definable in $\langle \mathbb{R}, \mathbb{Z}, +, \leq \rangle$.

Conclusions

- Generalizations of Cobham's and Semenov's theorems to mixed real & integer linear arithmetic.
- We considered both recognizability and weak recognizability properties.
- One surprise: For (strong) recognizability, multiplicative independence has to be strengthened.
- Weak deterministic automata turns out to be sufficient for all base-independent applications.
- A byproduct of the proof technique is a detailed documentation of the internal structure of RVA, which can be exploited to develop efficient data structures.