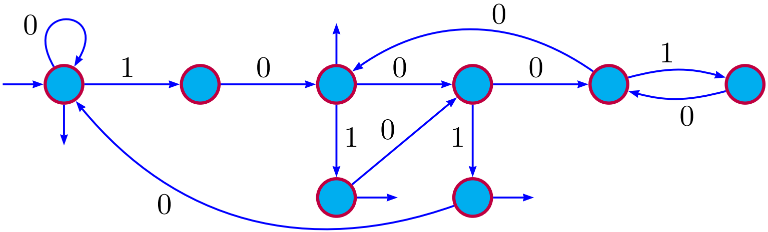
Complexité syntaxique et numérations

Émilie Charlier

JMC 2012, Rouen, 13 juin

An example first



13	8	5	3	2	1	
				1	0	2
			1	0	1	4
		1	0	0	1	6
	1	0	0	0	0	8
	1	0	0	1	0	10
	1	0	1	0	1	12
						⋮

The set $2\mathbb{N}$ of even integers is *F-recognizable* or *F-automatic*, i.e., the language $\mathrm{rep}_F(2\mathbb{N}) = \{\varepsilon, 10, 101, 1001, 10000, \ldots\}$ is accepted by some finite automaton.

Remark (in terms of the Chomsky hierarchy)

With respect to the Zeckendorf system, *any* F-recognizable set can be considered as a "*particularly simple*" set of integers.

We get a similar definition for other numeration systems.

Numeration systems

- A *numeration system* (NS) is an increasing sequence of integers $U = (U_n)_{n\geq 0}$ such that
  - $U_0 = 1$ and
  - $C_U := \sup_{n\geq 0} \lceil U_{n+1}/U_n \rceil < +\infty$.
- U is *linear* if it satisfies a linear recurrence relation over $\mathbb{Z}$.
- Let $n \in \mathbb{N}$. A word $w = w_{\ell-1} \cdots w_0$ over $\mathbb{N}$ represents n if

$$\sum_{i=0}^{\ell-1} w_i \, U_i = n.$$

- In this case, we write $\mathrm{val}_U(w) = n$.

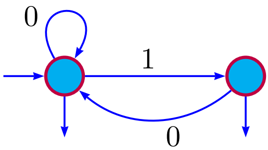
Greedy representations

- A representation $w = w_{\ell-1} \cdots w_0$ of an integer is *greedy* if

$$\forall j, \sum_{i=0}^{j-1} w_i U_i < U_j.$$

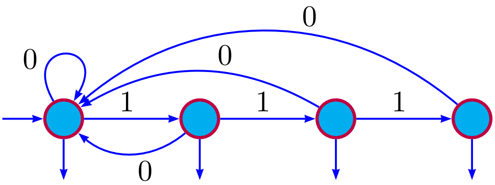
- In that case, $w \in \{0, 1, \ldots, C_U - 1\}^*$.
- $\text{rep}_U(n)$ is the greedy representation of n with $w_{\ell-1} \neq 0$.
- $X \subseteq \mathbb{N}$ is U-recognizable $\stackrel{\Delta}{\Leftrightarrow}$ $\text{rep}_U(X)$ is accepted by a finite automaton.
- $\text{rep}_U(\mathbb{N})$ is the numeration language.

Zeckendorf (or Fibonacci) numeration system



- $F_{n+2} = F_{n+1} + F_n$
- $F_0 = 1,\ F_1 = 2$
- $\mathcal{A}_F$ accepts all words that do not contain 11.

The ℓ-bonacci numeration system



- $U_{n+\ell} = U_{n+\ell-1} + U_{n+\ell-2} + \cdots + U_n$
- $U_i = 2^i$, $i \in \{0, \ldots, \ell - 1\}$
- $\mathcal{A}_U$ accepts all words that do not contain 1^ℓ.

U-recognizability of arithmetic progressions

Theorem

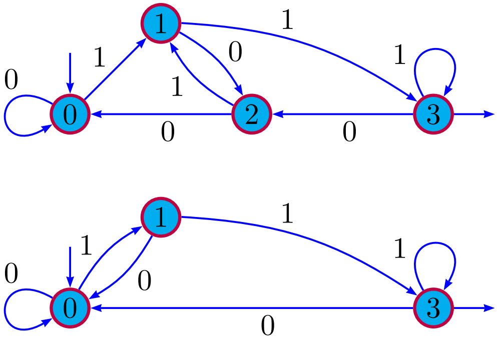
Let $U = (U_i)_{i\geq 0}$ be a NS such that $\mathbb{N}$ is U-recognizable. Then $m\,\mathbb{N}+r$ is U-recognizable for all $m, r \in \mathbb{N}$, and, given a DFA accepting $\mathrm{rep}_U(\mathbb{N})$, a DFA accepting $\mathrm{rep}_U(m\,\mathbb{N}+r)$ can be obtained effectively. Consequently, any ultimately periodic set is U-recognizable.

Theorem

Let U be a PNS. If $\mathbb{N}$ is U-recognizable, then U is linear, i.e., it satisfies a linear recurrence relation over $\mathbb{Z}$.

Motivations

What is the “best automaton” we can get?



DFAs accepting the binary representations of $4\mathbb{N} + 3$.

Question

The general algorithm doesn't provide a minimal automaton. What is the state complexity of $\mathrm{rep}_U(m\mathbb{N} + r)$?

Related questions

Suppose that $\mathrm{rep}_U(\mathbb{N})$ is regular and let $\mathcal{A}_U$ be the trim minimal automaton recognizing $\mathbb{N}$.

- What does $\mathcal{A}_U$ look like?

Suppose we are interested in a certain property $\mathcal{P}$ on sets of integers (like being ultimately periodic) that is U-recognizable.

- Describe the state complexity of a set $X \in \mathcal{P}$ w.r.t U.
- Describe the *syntactic complexity* of a set $X \in \mathcal{P}$ w.r.t U.

Honkala's decision procedure 1986

Given any finite automaton recognizing a set X of integers written in base b, it is decidable whether X is ultimately periodic.

Main ideas for an automata-resolution of this problem:

- If $X \subseteq \mathbb{N}$ is ultimately periodic, then the state complexity of the associated minimal DFA should grow with the period and preperiod of X.
- Analyse the inner structure of DFAs accepting the U-representations of $m\,\mathbb{N}+r$.

Information we are looking for

Consider a NS U such that $\mathbb{N}$ is U-recognizable.

How many states does the trim minimal automaton $\mathcal{A}_{U,m}$ recognizing $m\,\mathbb{N}$ contain?

1. Give upper/lower bounds?
2. Study special cases, e.g., Zeckendorf numeration system.

All these questions could be reformulated using the syntactic monoid instead of the minimal automaton.

State complexity

A general upper bound

Theorem (Krieger et al. 2009, Angrand-Sakarovitch 2010)

Let $m, r \in \mathbb{N}$ *with* $m \geq 2$ *and* $r < m$.
If $\mathrm{rep}_U(\mathbb{N})$ *is accepted by a* n*-state DFA, then the minimal automaton of* $\mathrm{rep}_U(m\mathbb{N} + r)$ *has at most* $n\, m^n$ *states.*

NB: This result remains true for the larger class of *abstract numeration systems*.

An exact result for the integer bases

Theorem (Alexeev 2004)

Let $b, m \geq 2$. Let N, M be such that $b^N < m \leq b^{N+1}$ and
$(m,1) < (m,b) < \cdots < (m,b^M) = (m,b^{M+1})$.
The minimal automaton recognizing $m\mathbb{N}$ in base b has exactly

$$\frac{m}{(m,b^{N+1})} + \sum_{t=0}^{\min\{N,M-1\}} \frac{b^t}{(m,b^t)} \text{ states.}$$

In particular, if m and b are coprime, then this number is just m.
Further, if $m = b^n$, then this number is $n+1$.

A lower bound

Theorem (C-Rampersad-Rigo-Waxweiler 2011)

Let U be any numeration system (not necessarily linear). The number of states of $\mathcal{A}_{U,m}$ is at least $|\operatorname{rep}_U(m)|$.

The Hankel matrix

- Let $U = (U_n)_{n\geq 0}$ be a linear numeration system.
- Let $k = k_{U,m}$ be the length of the shortest linear recurrence relation satisfied by $(U_i \bmod m)_{i\geq 0}$.
- For $t \geq 1$ define

$$H_t := \begin{pmatrix} U_0 & U_1 & \cdots & U_{t-1} \\ U_1 & U_2 & \cdots & U_t \\ \vdots & \vdots & \ddots & \vdots \\ U_{t-1} & U_t & \cdots & U_{2t-2} \end{pmatrix}.$$

- For $m \geq 2$, $k_{U,m}$ is also the largest t such that $\det H_t \not\equiv 0 \pmod{m}$.

A system of linear congruences

- Let $S_{U,m}$ denote the number of k-tuples $\mathbf{b}$ in $\{0, \ldots, m-1\}^k$ such that the system

$$H_k \, \mathbf{x} \equiv \mathbf{b} \pmod{m}$$

has at least one solution $\mathbf{x} = (x_1, \ldots, x_k)$.

- $S_{U,m} \leq m^k$.

Calculating $S_{U,m}$

- $U_{n+2} = 2U_{n+1} + U_n$, $(U_0, U_1) = (1, 3)$
- $(U_n)_{n \geq 0} = 1, 3, 7, 17, 41, 99, 239, \ldots$
- Consider the system

$$\begin{cases} 1\,x_1 + 3\,x_2 & \equiv & b_1 & \pmod 4 \\ 3\,x_1 + 7\,x_2 & \equiv & b_2 & \pmod 4 \end{cases}$$

- $2x_1 \equiv b_2 - b_1 \pmod 4$
- For each value of b_1 there are at most 2 values for b_2.
- Hence $S_{U,4} = 8$.

General state complexity result

Theorem

Let $m \geq 2$ and let $U = (U_n)_{n\geq 0}$ be a numeration system s.t.

(a) $\mathbb{N}$ is U-recognizable;

(b) $\mathcal{A}_U$ has a single strongly connected component $\mathcal{C}_U$;

(c) for all states p, q in $\mathcal{C}_U$ with $p \neq q$, there exists a word x_{pq} such that $p \cdot x_{pq} \in \mathcal{C}_U$ and $q \cdot x_{pq} \notin \mathcal{C}_U$, or vice-versa;

(d) $(U_n \bmod m)_{n\geq 0}$ is purely periodic.

Then the number of states of $\mathcal{A}_{U,m}$ from which infinitely many words are accepted is

$$|\mathcal{C}_U|\, S_{U,m}.$$

Idea of the proof

Let L be a language over the alphabet Σ.
The Myhill-Nerode equivalence relation for L: $u \sim_L v$ means that for all $y \in \Sigma^*$, $uy \in L \Leftrightarrow vy \in L$.

The number of states of $\mathcal{A}_{U,m}$ from which infinitely many words are accepted is the number of sets $u^{-1} 0^* \operatorname{rep}_U(m\,\mathbb{N})$ where u is s.t. $q_0 \cdot u$ belongs to $\mathcal{C}_U$ (where q_0 is the initial state of $\mathcal{A}_U$).

For all $u, v \in A_U^*$ s.t. $q_0 \cdot u$ and $q_0 \cdot v$ belong to $\mathcal{C}_U$, we have $u \sim_{0^* \operatorname{rep}_U(m\mathbb{N})} v$ iff

$$\begin{cases} u \sim_{0^* \operatorname{rep}_U(\mathbb{N})} v \quad \text{and} \\ \forall i \in \{0, \ldots, k-1\},\ \operatorname{val}_U(u0^i) \equiv \operatorname{val}_U(v0^i) \pmod{m} \end{cases}$$

Result for strongly connected automata

Corollary

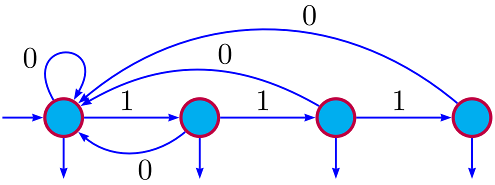
If U satisfies the conditions of the previous theorem and $\mathcal{A}_U$ is strongly connected, then the number of states of $\mathcal{A}_{U,m}$ is

$$|\mathcal{A}_U|\, S_{U,m}.$$

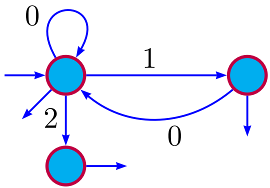
Further, we get an automatic procedure to obtain directly the minimal automaton $\mathcal{A}_{U,m}$ of $0^* \operatorname{rep}_U(m\mathbb{N})$.

Bertrand numeration systems

- *Bertrand numeration system*: w is in $\operatorname{rep}_U(\mathbb{N})$ if and only if $w0$ is in $\operatorname{rep}_U(\mathbb{N})$.
- E.g., the ℓ-bonacci system is Bertrand.



A non-Bertrand system



- $U_{n+2} = U_{n+1} + U_n,\ U_0 = 1,\ U_1 = 3$
- $(U_n)_{n\geq 0} = 1, 3, 4, 7, 11, 18, 29, 47, \ldots$
- 2 is a greedy representation but 20 is not.

Theorem (Bertrand)

A numeration system U is Bertrand iff there is a $\beta > 1$ s.t.

$$0^* \operatorname{rep}_U(\mathbb{N}) = L(\beta).$$

In that case, if $d_\beta^*(1) = (t_i)_{i \geq 1}$, then

$$U_n = t_1 U_{n-1} + \cdots + t_n U_0 + 1.$$

- If β is a Parry number, the system is linear and we have a finite trim minimal automaton $\mathcal{A}_\beta$ accepting $L(\beta)$.
- Consequently, $\operatorname{rep}_U(\mathbb{N})$ is regular and $\mathcal{A}_U = \mathcal{A}_\beta$.

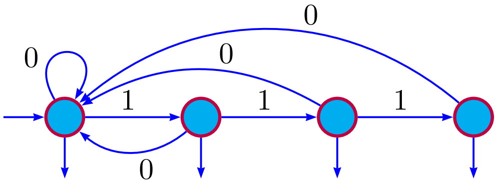
Applying our state complexity result to the Bertrand systems

Proposition

Let U be the Bertrand numeration system associated with a non-integral Parry number $\beta > 1$. The set $\mathbb{N}$ is U-recognizable and the trim minimal automaton $\mathcal{A}_U$ of $0^* \operatorname{rep}_U(\mathbb{N})$ fulfills the hypotheses of the theorem.

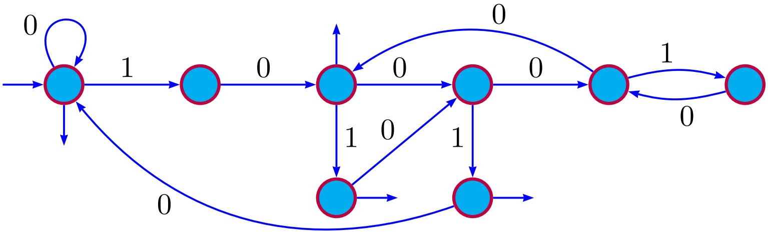
Consequently the previous state complexity result applies to the class of Bertrand numeration systems.

Result for the ℓ-bonacci system



Corollary

For U the ℓ-bonacci system, the number of states of $\mathcal{A}_{U,m}$ is ℓm^{ℓ}.



13	8	5	3	2	1	
				1	0	2
			1	0	1	4
		1	0	0	1	6
	1	0	0	0	0	8
	1	0	0	1	0	10
	1	0	1	0	1	12
						⋮

Further work for state complexity

- Analyze the structure of $\mathcal{A}_U$ for systems with no dominant root.
- Remove the assumption that $(U_n \bmod m)_{n\geq 0}$ is purely periodic in the state complexity result.
- Look at any arithmetic progressions $X = m\,\mathbb{N} + r$.

Transition to syntactic complexity

Transition to syntactic complexity

Let $N_U(m) \in \{1, \dots, m\}$ denote the number of values that are taken infinitely often by the sequence $(U_i \bmod m)_{i \geq 0}$.

Example (Zeckendorf system)

$(F_i \bmod 4) = (1, 2, 3, 1, 0, 1, 1, 2, 3, \dots)$, so $N_F(4) = 4$.

$(F_i \bmod 11) = (1, 2, 3, 5, 8, 2, 10, 1, 0, 1, 1, 2, 3, \dots)$, so $N_F(11) = 7$.

Theorem (C-Rigo 2008)

Let $U = (U_i)_{i \geq 0}$ be a NS s.t. $\lim_{i \to +\infty} U_{i+1} - U_i = +\infty$.

If $X \subseteq \mathbb{N}$ is an ultimately periodic U-recognizable set of period p, then any DFA accepting $\mathrm{rep}_U(X)$ has at least $N_U(p)$ states.

- If $N_U(m) \to +\infty$ as $m \to +\infty$, then we obtain a decision procedure to the periodicity problem.
- If U satisfies

$$U_{i+k} = a_1 U_{i+k-1} + \cdots + a_k U_i,\ i \geq 0, \quad \text{with} \quad a_k = \pm 1,$$

then $\lim_{m \to +\infty} N_U(m) = +\infty$.
- Works for the Zeckendorf system.
- Not true for integer base b: $N(b^n) = 1$ for all $n \geq 0$.

- The formula for the state complexity of $m\,\mathbb{N}$ for the Zeckendorf system is much simpler than the formula for integer base b systems.
- In this point of view, state complexity is not completely satisfying.
- Hope: Find a complexity that would handle all these systems in a kind of uniform way.

Syntactic complexity

Syntactic complexity

- Let L be a language over the alphabet Σ.
- Myhill-Nerode equivalence relation for L: $u \sim_L v$ means that for all $y \in \Sigma^*$, $uy \in L \Leftrightarrow vy \in L$.
- Leads to the minimal automaton of L: $|\mathcal{A}_L| = |\Sigma^*/{\sim_L}|$ is the state complexity of L.
- Syntactic congruence for L: $u \equiv_L v$ means that for all $x, y \in \Sigma^*$, $xuy \in L \Leftrightarrow xvy \in L$.
- Leads to the syntactic monoid of L: $(\Sigma^*/{\equiv_L}, \circ)$ where $[u] \circ [v] = [uv]$.
- $|\Sigma^*/{\equiv_L}|$ is the syntactic complexity of L.

Theorem

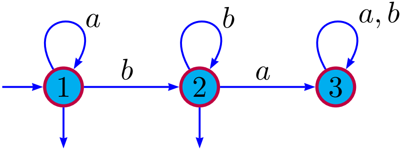
A language L *is regular iff* $\Sigma^*/{\equiv_L}$ *is finite.*

Theorem

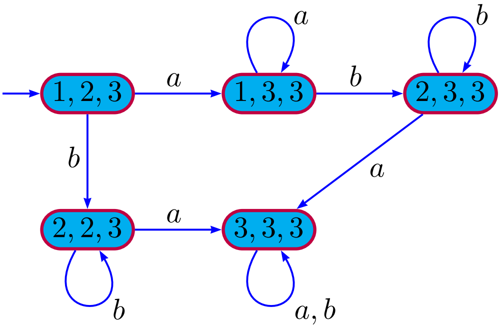
Let L *be a language over* Σ. *Two words* $u, v \in \Sigma^*$ *are s.t.* $u \equiv_L v$ *iff they perform the same transformation on the set of states of the minimal automaton* $\mathcal{A}_L$: $q \cdot u = q \cdot v$ *for all states* q.

An example: $L = a^*b^*$

Minimal automaton:

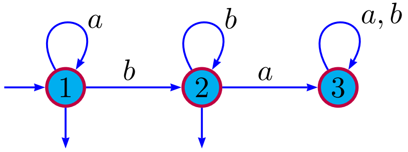


Representation of the syntactic monoid:

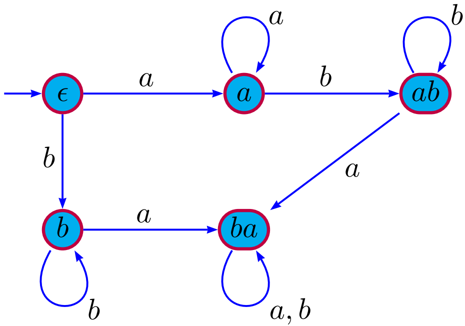


An example: $L = a^*b^*$

Minimal automaton:



Representation of the syntactic monoid:



Syntactic complexity for integer bases

The syntactic complexity of $X \subseteq \mathbb{N}$ is the syntactic complexity of the language $0^* \operatorname{rep}_U(X)$.

For x, y coprime, $\operatorname{ord}_y(x) = \min\{j \in \mathbb{N}_0 \colon x^j \equiv 1 \pmod{y}\}$.

Theorem (Rigo-Vandomme 2011)

- *Let $m, b \geq 2$ be coprime integers.*
  If $X \subseteq \mathbb{N}$ is periodic of minimal period m, then the syntactic complexity of X is equal to $m \operatorname{ord}_m(b)$.

Main idea: For all $u, v \in A_U^*$, we have $u \equiv_{0^* \operatorname{rep}_b(X)} v$ iff

$$\begin{cases} |u| \equiv |v| \pmod{\operatorname{ord}_m(b)} \quad \text{and} \\ \operatorname{val}_b(u) \equiv \operatorname{val}_b(v) \pmod{m} \end{cases}$$

Theorem (continued)

- *Let* $b \geq 2$ *and* $m = b^n$ *with* $n \geq 1$.
  (a) *The syntactic complexity of* $m\mathbb{N}$ *is equal to* $2n + 1$.
  (b) *If* $X \subseteq \mathbb{N}$ *is periodic of minimal period* m, *then the syntactic complexity of* X *is* $\geq n + 1$.
- *Let* $b \geq 2$ *and* $m = b^n q$ *with* $n \geq 1$ *and* $(b, q) = 1$. *Then the syntactic complexity of* $m\mathbb{N}$ *is equal to* $(n + 1)\, q \operatorname{ord}_q(b) + n$.

A general lower bound for the integer base case

Theorem (Lacroix-Rampersad-Rigo-Vandomme 2012)

Let $b \geq 2$ *and* $m = db^n q$ *with* $n \geq 1$ *and* $(b, q) = 1$ *and where* n *and* q *are chosen to be maximal.*

If $X \subseteq \mathbb{N}$ *is periodic of minimal period* m*, then the syntactic complexity of* X *is*

$$\geq \max\left(q\,\mathrm{ord}_q(b), \frac{\gamma + 1}{q\,\mathrm{ord}_q(b)}\right),$$

where $\gamma \to +\infty$ *as* n *or* $d \to +\infty$*.*

Zeckendorf numeration system and further work

Theorem

The syntactic complexity of $m\,\mathbb{N}$ *is*

$$4m^2 p_F(m) + 2$$

where $p_F(m)$ *is the minimal period of* $(F_i \mod m)_{i\geq 0}$.

So far, we can show that this result extends to the Bertrand systems s.t. $(U_n \mod m)_{n\geq 0}$ is purely periodic.

Further work and conclusion

Further work for syntactic complexity:

- Try to estimate the syntactic complexity of periodic sets for a larger class of numeration systems.

Conclusion

Syntactic complexity seems to allow us to handle integer bases and the Zeckendorf system at once.