Formal Languages and Number Theory

Jeffrey Shallit
Department of Computer Science
University of Waterloo*
Waterloo, Ontario N2L 3G1
Canada
shallit@graceland.uwaterloo.ca
http://www.math.uwaterloo.ca/~shallit

* Currently on sabbatical at the University of Arizona

Number Theory and Formal Languages

Number Theory: the study of the properties of integers

Formal Languages: the study of the properties of strings

At their intersection:

(a) the study of the properties of integers based on their *representation* in some manner, such as representation in base k;

and

(b) the study of the properties of strings of digits based on the integers they represent.

Number Theory and Formal Languages

Classical example of a theorem of type (a): properties of integers based on their representation in base k:

- Kummer's 1852 theorem
- states that the highest power of a prime p which divides the binomial coefficient $\binom{n}{m}$ is equal to the number of "carries" when m is added to $n-m$ in base p.
- example: $n = 13, m = 8, p = 3$.
- then 8 is 22 in base 3, 5 is 12 in base 3, and adding them gives 111 with two carries in base 3, and we find $\binom{13}{8} = 3^2 \cdot 11 \cdot 13$.
- Easily proved using Legendre's theorem

$$\nu_p(n!) = \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \cdots$$

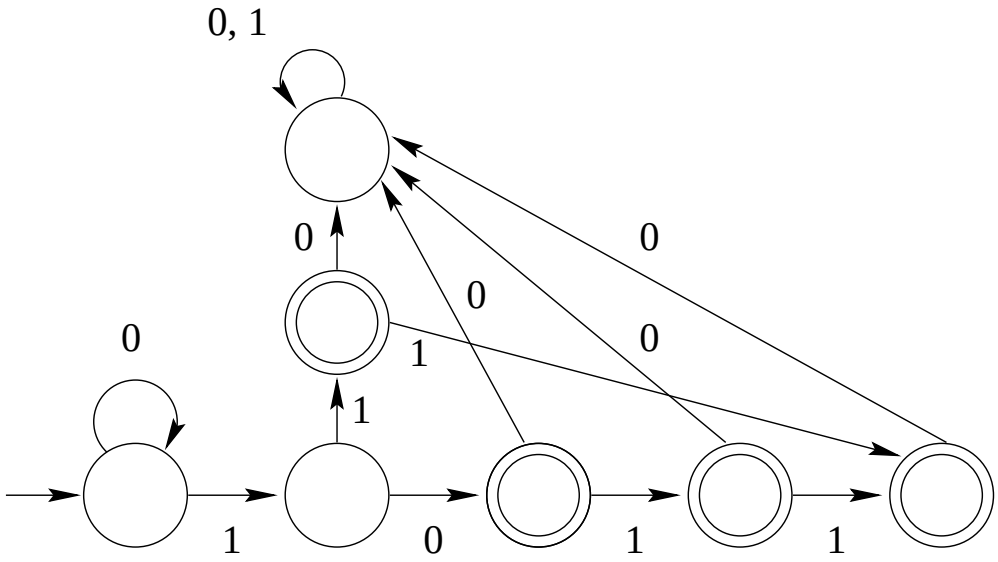
Number Theory and Formal Languages

Example of a theorem of type (b): properties of strings based on the integers they represent:

- Call a set of strings *sparse* if, as $n \to \infty$, it contains a vanishingly small fraction of all possible strings of length n.
- Can one find a sparse set S over $\{0, 1\}$ such that every string in $\{0, 1\}^*$ can be written as the concatenation of two strings from S?
- Solution (Enflo, Granville, Shallit, and Yu): Let S be the set of all strings of 0's and 1's such that the number of 1's is a sum of two squares.
- By Lagrange's theorem, every non-negative integer is the sum of *four* squares, so every string of 0's and 1's is the concatenation of two strings chosen from S. It can be shown, using a simple estimate in sieve theory, that S is sparse.

Finite Automata

- A *deterministic finite automaton* (DFA) is a simple model of a computer.



Transition diagram for automaton accepting the base-2 representations of the primes $p \leq 11$

Basics of Finite Automata

- Formally a DFA is a quintuple: $M = (Q, \Sigma, \delta, q_0, F)$ where:
  - Q is a finite set of *states*;
  - the *size* of M is $|M| := |Q|$, the number of states;
  - Σ is a finite set of symbols, called the *input alphabet*;
  - $q_0 \in Q$ is the *start state*;
  - $F \subseteq Q$ is the set of *final states*;
  - $\delta : Q \times \Sigma \to Q$ is the *transition function*, which is extended to $\delta : Q \times \Sigma^*$ in the obvious way
- The *language accepted by* M is denoted by $L(M)$ and is given by

$$\{w \in \Sigma^* \mid \delta(q_0, w) \in F\}.$$

- A language L is said to be *regular* if it is accepted by some DFA M.

State Complexity

The *state complexity* of a regular language L, $\mathrm{sc}(L)$, is the minimum number of states needed to accept it by a deterministic finite automaton (DFA).

The problem:
Given languages L, L' with state complexity n, n' respectively, what are good bounds on the state complexity of

- $L \cup L'$;
- LL';
- L^*, etc.?

State Complexity

For the state complexity of intersection, we have the following bound:

Proposition. *We have*

$$\mathrm{sc}(L \cap L') \leq \mathrm{sc}(L)\mathrm{sc}(L').$$

Proof. Let L be accepted by the DFA $(Q, \Sigma, \delta, q_0, F)$ and L' be accepted by the DFA $(Q', \Sigma, \delta', q_0', F')$. Then $L \cap L'$ can be accepted by a DFA

$$(Q'', \Sigma, \delta'', q_0'', F'')$$

where

- $Q'' := Q \times Q'$;
- $q_0'' := [q_0, q_0']$;
- $F'' := F \times F'$; and
- $\delta''([p, q], a) = [\delta(p, a), \delta(q, a)]$. ∎

State Complexity of Intersection

The upper bound of $\mathrm{sc}(L)\mathrm{sc}(L')$ can be achieved if L, L' are over an alphabet of size at least 2:

Proposition. (S. Yu.) *Define*

$$L := \{x \in \{a,b\}^* \ : \ |x|_a \equiv 0 \pmod{n};$$

$$L' := \{y \in \{a,b\}^* \ : \ |y|_b \equiv 0 \pmod{n'}.$$

Then

$$\mathrm{sc}(L \cap L') = nn'.$$

But what if L, L' are *unary*, that is, defined over an alphabet of one symbol?

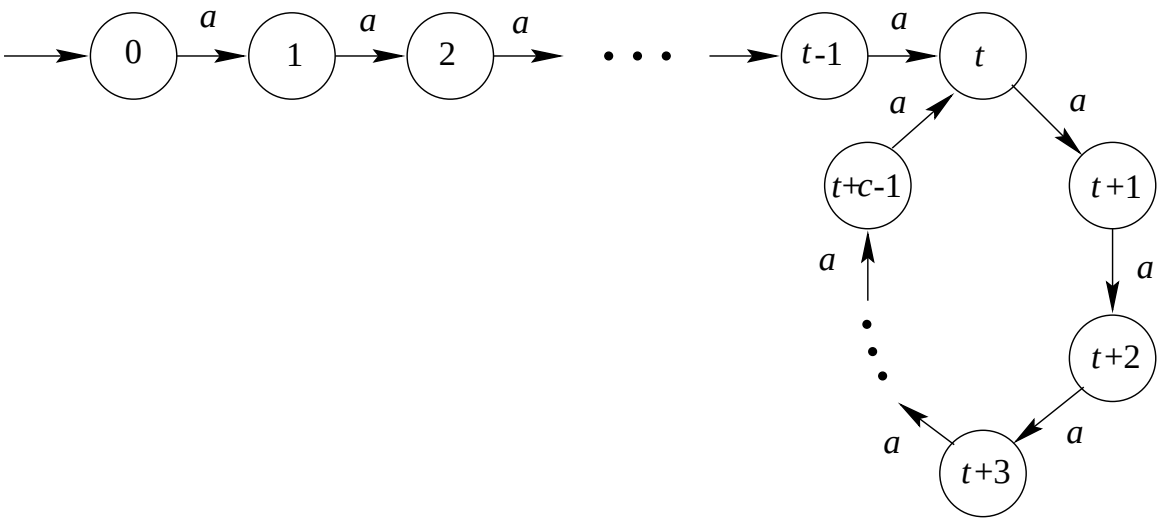
Clearly if $\gcd(n, n') = 1$ then the bound nn' can again be achieved, by taking $L = (a^n)^*$ and $L' = (a^{n'})^*$.

But what if $\gcd(n, n') > 1$?

State Complexity of Intersection for Unary Languages

A connected unary DFA has the property that its transition diagram consists of

- a *tail* of $t \geq 0$ states and
- a *cycle* of $c \geq 1$ states.



Transition diagram of unary DFA
(accepting states not identified)

State Complexity of Intersection for Unary Languages

It is then not hard to prove that

Theorem. *Let M, M' be unary DFA's with tails of size t, t' and cycles of size c, c', respectively. If L, L' are the corresponding languages, we have*

$$\mathrm{sc}(L \cap L') \leq \max(t, t') + \mathrm{lcm}(c, c'). \qquad (1)$$

Furthermore, for all $t, t' \geq 0$ and $c, c' \geq 1$ there exist unary languages for which the bound (1) is achieved.

For example, if $t \geq t'$, take

$$\begin{aligned} L &= a^{t+c-1}(a^c)^*; \\ L' &= a^r(a^{c'})^*; \\ r &= t - 1 \bmod c'. \end{aligned}$$

Two New Number-Theoretic Functions

Thus, to estimate the worst-case behavior for the state complexity of intersection of unary languages with n and n' states, respectively, we must estimate the function

$$F(n,n') = \max_{\substack{1\le c\le n\\ 1\le c'\le n'}} \left(\max(n-c, n'-c') + \operatorname{lcm}(c,c') \right).$$

This in turn suggests studying the somewhat simpler and more natural function

$$G(n,n') = \max_{\substack{1\le c\le n\\ 1\le c'\le n'}} \operatorname{lcm}(c,c').$$

- The asymptotic behavior of F and G is still not known precisely
- There is a relation to JACOBSTHAL'S function $g(n)$, which is the *least integer r such that every set of r consecutive integers contains at least one integer relatively prime to n.*
- IWANIEC proved [1978] using the linear sieve that $g(n) = O((\log n)^2)$.
- It then follows that if $n \le n'$, we have

  $$F(n,n') \ge G(n,n') \ge nn' - c(\log n)^2 n$$

  for some constant c.

- Can also prove

Theorem. There exist a constant c_2 and infinitely many distinct pairs n, n' with $n' < n$ such that

$$G(n, n') \leq F(n, n') \leq nn' - c_2 \sqrt{\frac{\log n}{\log \log n}} n.$$

Enumeration of Unary DFA Languages

Theorem.

There are

$$2^n(n - \alpha + O(n2^{-n/2}))$$

distinct languages accepted by unary DFA's with n states, where

$$\alpha = \sum_{d \geq 2} \frac{\mu(d)}{1 - 2^{d-1}} \doteq 1.382714455402396285 47.$$

Proof. (Sketch)

The transition diagram of a connected unary DFA consists of a "tail" and a "loop". An n-state unary DFA is minimal iff

(a) its transition diagram is connected;

(b) the loop in its transition diagram is minimal;

(c) the last state in the tail is is final, and the last state in the loop is non-final, or vice versa.

Now count these.

Enumeration of Unary DFA Languages

This turns out to be

$$\sum_{1\le t\le n}\left(\sum_{d\,|\,t}\mu(d)2^{t/d}\right)2^{n-t} = 2^n\sum_{1\le t\le n}\sum_{d\,|\,t}\mu(d)2^{t/d-t}$$

$$= 2^n\left(n+\sum_{1\le t\le n}\sum_{\substack{d\,|\,t\\ d\neq 1}}\mu(d)2^{t/d-t}\right)$$

so it suffices to estimate

$$\sum_{1\le t\le n}\sum_{\substack{d\,|\,t\\ d\neq 1}}\mu(d)2^{t/d-t}.$$

Enumeration of Unary DFA Languages

We need to estimate

$$\sum_{1\le t\le n}\sum_{\substack{d\,|\,t\\ d\neq 1}}\mu(d)2^{t/d-t}.$$

Let $t=kd$ and reverse the order of summation. We find

$$\begin{aligned}
\sum_{1\le t\le n}\sum_{\substack{d\,|\,t\\ d\neq 1}}\mu(d)2^{t/d-t} &= \sum_{2\le d\le n}\mu(d)\sum_{1\le k\le \frac{n}{d}}2^{k-kd}\\
&= \sum_{2\le d\le n}\mu(d)\left(O(2^{n/d-n})+\sum_{k\ge 1}2^{k-kd}\right)\\
&= \sum_{2\le d\le n}\mu(d)\left(O(2^{n/d-n})+\frac{1}{2^{d-1}-1}\right)\\
&= \left(\sum_{2\le d\le n}\frac{\mu(d)}{2^{d-1}-1}\right)+O(n2^{-n/2})
\end{aligned}$$

and this gives the desired result. ∎

Context-Free Grammars

- A method for generating languages
- Modern mathematical formulation due to Chomsky [1956], although idea goes back to Indian philologist Panini, c. 400 b.c.e.
- Consists of a start symbol and *rewriting rules*, e.g.:

$$\begin{aligned} S &\rightarrow aSa \\ S &\rightarrow bSb \\ S &\rightarrow a \\ S &\rightarrow b \\ S &\rightarrow \epsilon \end{aligned}$$

which generates the palindromes over $\{a, b\}$.

Context-Free Grammars

- More formally, a context-free grammar (CFG) is a 4-tuple $G = (V, \Sigma, P, S)$ where
  - V is a finite set of variables
  - Σ is a finite alphabet
  - P is a set of production rules of the form $A \rightarrow \gamma$, where $A \in V$ and $\gamma \in (V \cup \Sigma)^*$
  - S is the start symbol

Context-Free Grammars

- We write $\alpha \Longrightarrow \beta$ if β can be obtained from α by the use of one production rule.
- We write $\Longrightarrow^*$ for the reflexive, transitive closure of $\Longrightarrow$.
- Then $L(G)$, the language generated by G is formally defined as

$$L(G) := \{x \in \Sigma^* \ : \ S \Longrightarrow^* x\}.$$

- Context-free grammars generate a class of languages, the context-free languages, which are a strict superset of the class of regular languages.

Descriptional Complexity of Context-Free Grammars

- Can measure the size of a context-free grammar as the number of symbols needed to write down its description.
- Suppose a CFG G generates a regular language. How big can the corresponding DFA be, in terms of the size of G?
  - If the CFG is over an alphabet with at least 2 symbols, the answer is, there is no recursive bound.
  - More precisely, MEYER and FISCHER proved [1971] that given any recursive function f, for arbitrarily large integers n there exists a CFG of size n describing a regular language L such that any DFA accepting L has at least $f(n)$ states.
  - But how about the unary case?
  - It is possible to show that there exists a constant such that any unary CFG of size n describing a regular language can be accepted by a DFA with at most $O(2^{cn^2})$ states.
  - But is this bound achievable?

An Example Exhibiting 2^{cn^2} Blowup

$$
\begin{aligned}
A_0 &\rightarrow a \\
A_{i+1} &\rightarrow A_i A_i \quad (i \geq 0)
\end{aligned}
$$

$$
\text{so } A_i \Longrightarrow^* \{a^{2^i}\}
$$

$$
B_i \rightarrow a A_i
$$

$$
\text{so } B_i \Longrightarrow^* \{a^{2^i+1}\}
$$

$$
\begin{aligned}
C_0 &\rightarrow a \\
C_{i+1} &\rightarrow a \mid C_i C_i \quad (i \geq 0)
\end{aligned}
$$

$$
\text{so } C_i \Longrightarrow^* \{a, a^2, a^3, \ldots, a^{2^i}\}
$$

$$
D_i \rightarrow D_i B_i \mid C_i \quad (i \geq 0)
$$

$$
\begin{aligned}
\text{so } D_i &\Longrightarrow^* \{a, a^2, a^3, \ldots, a^{2^i}\}\{a^{2^i+1}\}^* \\
&= \{a^j \ : \ j \not\equiv 0 \pmod{2^i+1}\}.
\end{aligned}
$$

An Example Exhibiting 2^{cn^2} Blowup

And finally, let

$$S_i \quad \rightarrow \quad \epsilon \mid D_0 \mid D_1 \mid D_2 \mid \cdots \mid D_i$$

so $S_i \stackrel{*}{\Longrightarrow} \{\epsilon\} \cup$
$\{a^k \ : \ k \not\equiv 0 \pmod{\text{lcm}(2^0+1, 2^1+1, \ldots, 2^i+1)}\}.$

Now let $G_n = (V_n, \{a\}, P_n, S_n)$, where

$$V_n = \{A_i, B_i, C_i, D_i, S_i \ : \ 0 \leq i \leq n\}$$

and P_n is the set of $O(n)$ productions given above involving these variables.

It is clear that $L(G_n)$ is regular.

The shortest string not generated by G_n is of length

$$\text{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1)$$

and so any DFA accepting $L(G_n)$ must have at least this many states.

It remains to estimate

$$\text{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1)$$

An Example Exhibiting 2^{cn^2} Blowup

We use the following theorem of BÉZIVIN [1989]:

Theorem. *Let a, b be integers with $b \neq 0$ and $\gcd(a,b) = 1$. Let α, β be zeroes of the polynomial $X^2 - aX - b$. For $m \geq 2$ define*

$$u_m(n) = \frac{\alpha^{mn} - \beta^{mn}}{\alpha^n - \beta^n}.$$

Then

$$\lim_{n\to\infty} \frac{\log(u_m(1)u_m(2)\cdots u_m(n))}{\log \operatorname{lcm}(u_m(1), u_m(2), \ldots, u_m(n))} = \frac{(m-1)L(m)\pi^2}{6H(m)},$$

where

$$L(m) = \prod_{p \,|\, m} \left(1 - \frac{1}{p^2}\right)$$

and

$$H(m) = \sum_{\substack{d \,|\, m \\ d>1}} \frac{\varphi(d)\varphi(m/d)d}{m}.$$

Now take $a = 3, b = -2, m = 2$. Then $\alpha = 2$ and $\beta = 1$, and we obtain

$$\lim_{n\to\infty} \frac{\log((2^0+1)(2^1+1)\cdots(2^n+1))}{\log \operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1))} = \frac{\pi^2}{8}.$$

An Example Exhibiting 2^{cn^2} Blowup

On the other hand, it is easy to see that

$$\lim_{n\to\infty} \frac{(2^0+1)(2^1+1)\cdots(2^n+1)}{2^0 \cdot 2^1 \cdots 2^n} = c_1,$$

where $c_1 \doteq 4.768$, so it follows that

$$\log((2^0+1)(2^1+1)\cdots(2^n+1)) \sim \log c_1 + \frac{n(n+1)}{2}\log 2.$$

Putting this together with the Bézivin result, we get

$$\log \operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1) \sim \frac{4\log 2}{\pi^2} n^2.$$

The Primitive Words Problem

- Let Σ be a finite alphabet with at least two letters. A word $w \in \Sigma^*$ is said to be *primitive* if it cannot be expressed in the form x^k with $k \geq 2$.
- Open problem in formal languages: is the language P of primitive words context-free?
- Answer is almost certainly no, but nobody knows how to prove this.
- PETERSEN [1996] proved the weaker result that P is not unambiguously context-free.
- He did this using the CHOMSKY-SCHÜTZENBERGER theorem, which states that if L is a context-free language having an unambiguous grammar, and $a_n := |L \cap \Sigma^n|$, then $\sum_{n \geq 0} a_n X^n$ is a formal power series in $\mathbb{Z}[[X]]$ which is algebraic over $\mathbb{Q}(X)$.
- A remarkably simple proof was found by ALLOUCHE using the theory of automatic sequences.

Example of Chomsky-Schützenberger Theorem

Consider the unambiguous grammar

$$\begin{aligned} S &\to M \\ S &\to U \\ M &\to 0M1M \\ M &\to \epsilon \\ U &\to 0S \\ U &\to 0M1U \end{aligned}$$

which represents strings of "if-then-else" clauses.

Then this has the following commutative image:

$$\begin{aligned} S &= M + U \\ M &= x^2M^2 + 1 \\ U &= Sx + x^2MU \end{aligned}$$

Example of Chomsky-Schützenberger Theorem

This system has the following power series solutions:

$$\begin{aligned} M &= 1 + x^2 + 2x^4 + 5x^6 + 14x^8 + 42x^{10} + \cdots \\ U &= x + x^2 + 3x^3 + 4x^4 + 10x^5 + 15x^6 + 35x^7 + \cdots \\ S &= 1 + x + 2x^2 + 3x^3 + 6x^4 + 10x^5 + 20x^6 + \cdots \end{aligned}$$

By the Chomsky-Schützenberger theorem, each variable satisfies an algebraic equation over $\mathbb{Q}(x)$.

For example, we have

$$x(2x-1)S^2 + (2x-1)S + 1 = 0$$

Automata as Computers of Sequences

- We can generalize our notion of automaton to provide an output, not simply accept/reject.
- Formally, we define a *deterministic finite automaton with output* (DFAO) as a sextuple: $(Q, \Sigma, \delta, q_0, \Delta, \tau)$, where Δ is the finite *output alphabet* and $\tau : Q \to \Delta$ is the *output mapping*.
- Next, we decide on a integer base $k \geq 2$ and represent n as a string of symbols over the alphabet $\Sigma = \{0, 1, 2, \ldots, k-1\}$.
- To compute f_n, given an automaton M, express n in base-k, say,

  $$a_r a_{r-1} \cdots a_1 a_0,$$

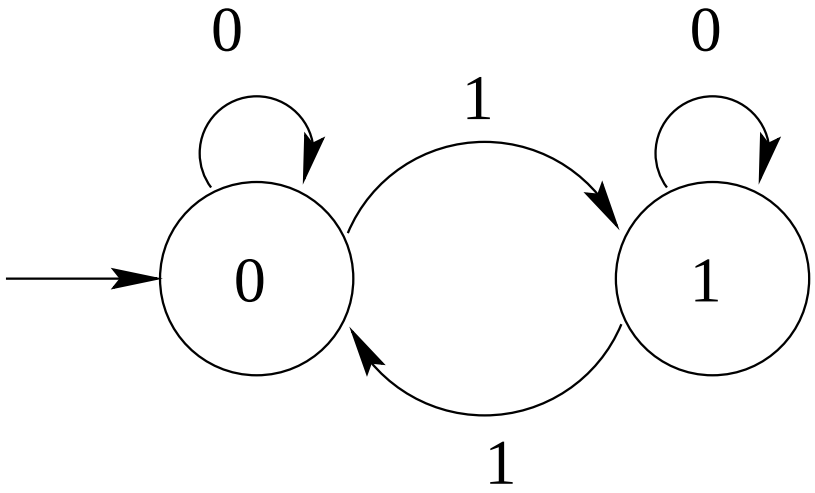
  and compute

  $$f_n = \tau(\delta(q_0, a_r a_{r-1} \cdots a_1 a_0)).$$

- Any sequence that can be computed in this way is said to be k-automatic.

Example: The Thue-Morse sequence

- The THUE-MORSE sequence $(t_n)_{n\geq 0}$ is defined as follows: t_n is the parity of the number of 1's in the binary expansion of n.
- $(t_n)_{n\geq 0} = 0\,1\,1\,0\,1\,0\,0\,1\,1\,0\,0\,1\,0\,1\,1\,0\cdots$
- We have $t_0 = 0$; $t_{2n} = t_n$, and $t_{2n+1} = 1 - t_n$ for $n \geq 0$.
- THUE (c. 1906) studied this sequence because it is *cubefree*: it contains no subword of the form www, where w is a nonempty word.
- It is computed by the following DFAO:



Robustness of the Notion of Automatic Sequence

- the order in which the base-k digits are fed into the automaton in does not matter (provided it is fixed for all n);
- other representations also work (such as expansion in base-$(-k)$);
- automatic sequences are closed under many operations, such as shift, periodic deletion, q-block compression, and q-block substitution.
- if a symbol in an automatic sequence occurs with well-defined frequency r, then r is rational.

The Theorem of Christol

Theorem. (Christol [1980]). Let $(u_n)_{n\geq 0}$ be a sequence over

$$\Sigma = \{0, 1, \ldots, p-1\},$$

where p is a prime. Then the formal power series $U(X) = \sum_{n\geq 0} u_n X^n$ is algebraic over $GF(p)[X]$ if and only if $(u_n)_{n\geq 0}$ is p-automatic.

Example.

Let, as before, $(t_n)_{n\geq 0}$ denote the Thue-Morse sequence, i.e., $t_n =$ sum of the bits in the binary expansion of n, mod 2. Then $t_{2n} \equiv t_n$ and $t_{2n+1} \equiv t_n + 1$. If we set $A(X) = \sum_{n\geq 0} t_n X^n$, then

$$\begin{aligned} A(X) &= \sum_{n\geq 0} t_{2n} X^{2n} + \sum_{n\geq 0} t_{2n+1} X^{2n+1} \\ &= \sum_{n\geq 0} t_n X^{2n} + X \sum_{n\geq 0} t_n X^{2n} + X \sum_{n\geq 0} X^{2n} \\ &= A(X^2) + X A(X^2) + X/(1-X^2) \\ &= A(X)^2(1+X) + X/(1+X)^2. \end{aligned}$$

Hence $(1+X)^3 A^2 + (1+X)^2 A + X = 0$.

Back to Primitive Words

- Let $\psi_k(n)$ be the number of primitive words of length n over a k-letter alphabet.
- Then it is easy to see (using MÖBIUS inversion) that

$$\psi_k(n) = \sum_{d \mid n} \mu(d) k^{n/d}.$$

- If P_k were unambiguously context-free then by the CHOMSKY-SCHÜTZENBERGER theorem

$$R(X) = \sum_{n \geq 1} \psi_k(n) X^n$$

would be algebraic over $\mathbb{Q}(X)$.

- Then

$$R'(X) = \sum_{n \geq 1} \frac{\psi_k(n)}{k} X^n$$

would also be algebraic over $\mathbb{Q}(X)$.

- Let p be a prime dividing k. Then it is not hard to see that

$$R'_p(X) = \sum_{n \geq 1} (\frac{\psi_k(n)}{k} \bmod p) X^n$$

would also be algebraic over $GF(p)(X)$.

- But

$$\begin{aligned}\frac{\psi_k(n)}{k} &= \sum_{d\,|\,n} \mu(d) k^{n/d-1} \\ &= \mu(n) + \sum_{\substack{d\,|\,n \\ d\neq n}} \mu(d) k^{n/d-1} \\ &\equiv \mu(n) \pmod{p}.\end{aligned}$$

- It follows that

$$R'_p(X) = \sum_{n\geq 1} \mu(n) X^n$$

  and so the sequence $(\mu(n) \bmod p)_{n\geq 0}$ must be p-automatic.

- But then $(\mu(n)^2 \bmod p)_{n\geq 0}$ would be p-automatic.

- However, $\mu(n)^2 \equiv 1 \pmod{p}$ if and only if n is squarefree.

- By a classical theorem, the density of the squarefree numbers exists and is equal to $6/\pi^2$, an irrational number.

- But the density of symbols in automatic sequences (if it exists) must be rational, a contradiction.

- It follows that $R(X)$ is not algebraic over $\mathbb{Q}(X)$ and so P_k is not unambiguously context-free.

For Further Reading

1. G. Christol, Ensembles presque périodiques k-reconnaissables, *Theoret. Comput. Sci.* **9** (1979), 141–145.

2. M. Domaratzki, D. Kisman, and J. Shallit, On the number of distinct languages accepted by finite automata with n states, to appear, *J. Autom. Lang. Combinat.*.

3. P. Enflo, A. Granville, J. Shallit and S. Yu, On sparse languages L such that $LL = \Sigma^*$, *Disc. Appl. Math.* **52** (1994), 275–285.

4. G. Pighizzini and J. Shallit, Unary language operations, state complexity, and Jacobsthal's function, to appear, *Int. J. Foundations of Computer Science*.