Number Theory and Formal Languages

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

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.

Languages

- If Σ is a finite set of symbols, then by Σ^* we mean the free monoid over Σ (set of all finite strings of symbols chosen from Σ);
- A *language* is a subset of Σ^*.
- Example:

$$\texttt{PAL} = \{x \in \{a,b\}^* \ : \ x = x^R\}$$

- The basic operations on languages are union, concatenation, and Kleene closure.
- concatenation:

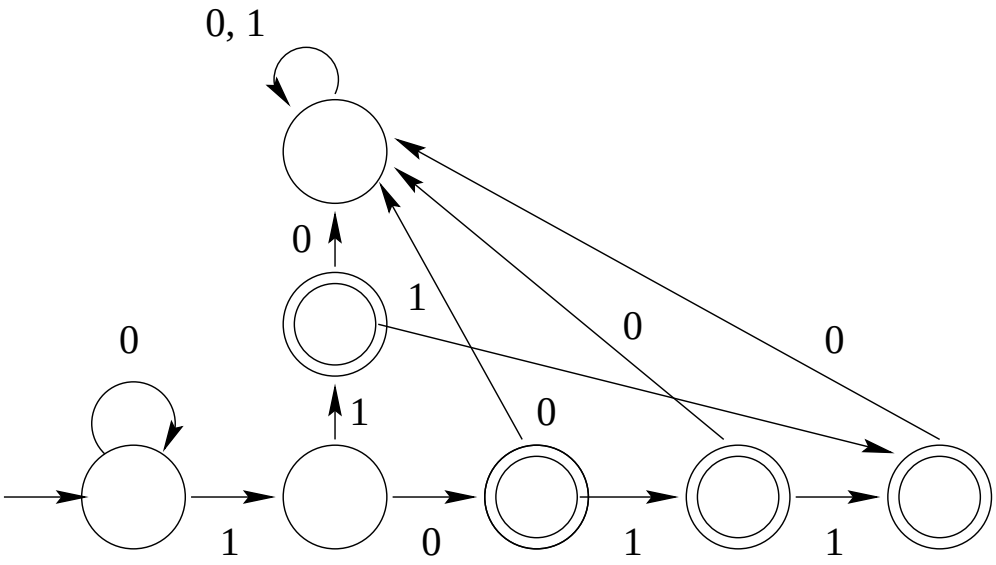
$$L_1 L_2 := \{xy \ : x \in L_1, \ y \in L_2\}.$$

- Kleene closure:

$$L^* := \bigcup_{i \geq 0} L^i.$$

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 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.

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*

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

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

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 $n \leq n'$, we have

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

for some constant c.

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.:

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

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 $A \to \gamma$, 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.

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:

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

This system has the following power series so-

lutions:

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

Example of Chomsky-Schützenberger Theorem

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,

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

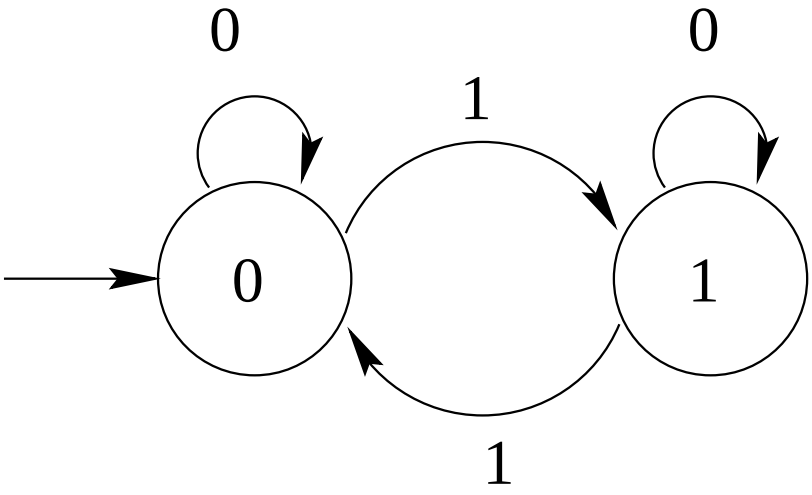
  and compute

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

- 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

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

- 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.

Transcendence in Finite Characteristic

Define for $n \geq 1$

$$\zeta_q(n) = \sum_{\substack{P \text{ monic} \\ P \in GF(q)[X]}} \frac{1}{P^n}$$

Thus, for example,

$$\begin{aligned}\zeta_2(1) &= \frac{1}{1} + \frac{1}{X} + \frac{1}{X+1} + \frac{1}{X^2} + \frac{1}{X^2+1} + \cdots \\ &= 1 + X^{-2} + X^{-3} + X^{-4} + X^{-5} + X^{-9} + \cdots \\ &\in GF(2)[[X^{-1}]].\end{aligned}$$

This function ζ_q, now called the CARLITZ zeta function, has many properties similar to those of the RIEMANN zeta function. For example, it admits the following EULER product:

$$\zeta_q(n) = \prod_{\substack{P \text{ irreducible} \\ P \in GF(q)[X]}} \frac{1}{1 - \frac{1}{P^n}}.$$

Transcendence in Finite Characteristic

CARLITZ also showed that if $q-1 \mid n$, then $\zeta_q(n) = \pi_q^n \cdot r$ where r is a rational function and

$$\pi_q := \prod_{k \geq 1} \left(1 - \frac{X^{q^k} - X}{X^{q^{k+1}} - X} \right).$$

WADE proved that π_q is transcendental. Here is another proof, due to ALLOUCHE, using automatic sequences and CHRISTOL'S theorem.

Taking the logarithmic derivative, we get

$$\begin{aligned} \frac{\pi_q'}{\pi_q} &= \sum_{k \geq 1} \frac{1}{X^{q^{k+1}} - X} \\ &= \left(\sum_{k \geq 1} \frac{1}{X^{q^k} - X} \right) - \frac{1}{X^q - X}. \end{aligned}$$

Transcendence in Finite Characteristic

Now suppose that π_q is algebraic over $GF(q)(X)$. Then so is the formal derivative π_q'. Hence so is π_q'/π_q. But then so is

$$\sum_{k\geq 1} \frac{1}{Xq^k - X} = \sum_{k\geq 1} \frac{1}{[k]},$$

the so-called "bracket series" introduced by WADE, who defined $[k]{:=}X^{q^k} - X$.

Thus to prove π_q transcendental, it suffices to show that

$$\sum_{k\geq 1} \frac{1}{Xq^k - X}$$

is transcendental.

We now have

$$\sum_{k\geq 1}\frac{1}{Xq^k-X}=\sum_{k\geq 1}\frac{1}{Xq^k\left(1-\left(\frac{1}{X}\right)^{q^k-1}\right)}$$

$$=\sum_{k\geq 1}\frac{1}{Xq^k}\sum_{n\geq 0}\left(\frac{1}{X}\right)^{n(q^k-1)}$$

$$=\frac{1}{X}\sum_{k\geq 1}\frac{1}{Xq^{k-1}}\sum_{n\geq 0}\left(\frac{1}{X}\right)^{n(q^k-1)}$$

$$=\frac{1}{X}\sum_{\substack{k\geq 1\\ n\geq 0}}\left(\frac{1}{X}\right)^{(n+1)(q^k-1)}$$

$$=\frac{1}{X}\sum_{\substack{k\geq 1\\ n\geq 1}}\left(\frac{1}{X}\right)^{n(q^k-1)}.$$

Transcendence in Finite Characteristic

Hence

$$\sum_{k\geq 1}\frac{1}{Xq^k-X}=\frac{1}{X}\sum_{m\geq 1}\left(\frac{1}{X}\right)^m\sum_{\substack{k,n\geq 1\\ n(q^k-1)=m}}1$$

$$=\frac{1}{X}\sum_{k\geq 1}\left(\frac{1}{X}\right)^m\sum_{\substack{k\geq 1\\ q^k-1\,|\,m}}1$$

$$=\frac{1}{X}\sum_{m\geq 1}\left(\frac{1}{X}\right)^m c(m),$$

where

$$c(m):=\sum_{\substack{k\geq 1\\ q^k-1\,|\,m}}1.$$

Now, by Christol's theorem, in order to show that $\sum_{k\geq 1}\frac{1}{Xq^k-X}$ is transcendental over $GF(q)(X)$, it suffices to prove that $(c(m) \bmod p)_{m\geq 1}$ is not q-automatic, where $q=p^e$ for some e.

Transcendence in Finite Characteristic

If the sequence $(c(m) \bmod p)_{m \geq 1}$ were q-automatic, then the subsequence $(c(q^n-1) \bmod p)_{n \geq 0}$ would be ultimately periodic. But

$$\begin{aligned} c(q^n - 1) &= \sum_{\substack{k \geq 1 \\ q^k - 1 \mid q^n - 1}} 1 \\ &= \sum_{\substack{k \geq 1 \\ k \mid n}} 1 \\ &= d(n), \end{aligned}$$

where $d(n)$ is the number of positive integral divisors of n.

It now suffices to show that $(d(n) \bmod p)_{n \geq 1}$ is not ultimately periodic. This can be done by a simple argument using DIRICHLET'S theorem. This contradiction completes the proof. ∎

For Further Reading

1. J.-P. Allouche, Note on the transcendence of a generating function, in A. Laurincikas and E. Manstavicius, eds., *Proc. Palanga Conference for the 75th Birthday of Prof. Kubilius*, New Trends in Probability and Statistics **4** (1997), 461–465.

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

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. J. Shallit, State complexity and Jacobsthal's function, in *Proc. CIAA 2000*, to appear.