Polynomial Automaticity, Context-Free Languages, and Fixed Points of Morphisms

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

This talk represents joint work with Ian Glaister.

Problem:

Given an object (finite string, infinite string, language, etc.), assign a measure of its complexity.

One Solution:

- Kolmogorov–(Chaitin–Solomonoff) complexity
  - See, for example, the recent book by Ming Li and Paul Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*
- "The complexity of an object is the length of the shortest program to produce it."
- very powerful idea
- lots of applications
- unfortunately not computable!
- is there a computable analogue?

Automaticity

- idea: like Kolmogorov complexity, but replace "Turing machine" with finite automaton
- goals:
  - measure complexity of languages
  - complexity measure is a function
  - regular languages should have $O(1)$ automaticity
  - languages "close" to regular should have "small" automaticity

Automaticity Defined

- $\Sigma^{\leq n} = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n$, the set of all strings in Σ^* of length $\leq n$.
- a language $L \subseteq \Sigma^*$ is an n'th order approximation to a language L' if $L \cap \Sigma^{\leq n} = L' \cap \Sigma^{\leq n}$.
- DFA = class of all deterministic (complete) finite automata over a finite alphabet Σ
- NFA = class of all nondeterministic finite automata over a finite alphabet Σ
- (deterministic) automaticity of a language L is the function which counts the number of states in the smallest DFA that accepts some n'th order approximation to L
- Formally, we define the (deterministic) automaticity of a language L to be the function $A_L(n) =$

  $$\min\{|M| \ : \ M \in \mathrm{DFA} \ \text{ and } L(M) \cap \Sigma^{\leq n} = L \cap \Sigma^{\leq n}\}.$$

- Similarly, we define the nondeterministic automaticity of a language L to be the function $N_L(n) =$

  $$\min\{|M| \ : \ M \in \mathrm{NFA} \ \text{ and } L(M) \cap \Sigma^{\leq n} = L \cap \Sigma^{\leq n}\}.$$

Previous Work

- Trakhtenbrot (1964)
- Grinberg & Korshunov (1966)
- Karp (1967)
- Breitbart (1970, 1971, 1973)
- Dwork and Stockmeyer (1989)
- Kaneps & Freivalds (1990)
- Shallit & Breitbart (1994)
- Pomerance, Robson, and Shallit (1995)

Basic Properties of Automaticity

1. $A_L(n) \leq A_L(n+1)$
2. $N_L(n) \leq N_L(n+1)$.
3. L is regular iff $A_L(n) = O(1)$.
4. L is regular iff $N_L(n) = O(1)$.
5. $A_L(n) = A_{\overline{L}}(n)$.
6. $A_L(n) \leq 2 + \Sigma_{w \in L \cap \Sigma^{\leq n}} |w|$.
7. $N_L(n) \leq 1 + \Sigma_{w \in L \cap \Sigma^{\leq n}} |w|$.

Definition. Two strings w, w' are called *n–dissimilar for L* if there exists a string v with $|wv|, |w'v| \leq n$ and either

(i) $wv \in L$, $w'v \notin L$; or

(ii) $wv \notin L$, $w'v \in L$.

Theorem. (Dwork & Stockmeyer; Kaneps & Freivalds) $A_L(n) =$ the maximum number of distinct pairwise n–dissimilar strings for L.

Polynomial Automaticity

Define the following three complexity classes:

- deterministic polynomial automaticity, or DPA

$$\mathrm{DPA} = \{L \subseteq \Sigma^* \ : \ \exists\, k \text{ such that } A_L(n) = O(n^k)\}.$$

- nondeterministic polynomial automaticity, or NPA

$$\mathrm{NPA} = \{L \subseteq \Sigma^* \ : \ \exists\, k \text{ such that } N_L(n) = O(n^k)\}.$$

- nondeterministic poly-log automaticity, or NPLA

$$\mathrm{NPLA} = \{L \subseteq \Sigma^* \ : \ \exists\, k \text{ such that } N_L(n) = O((\log n)^k)\}.$$

What are the closure properties of these classes, and how are they related?

A Hierarchy of Polynomial Automaticity

Theorem. For all integers $k \geq 0$ there exists a language L_k such that $A_{L_k} = \Theta(n^k)$.

Proof. Let

$$L_k = \{0^{a_1}\ 1\ 0^{a_2}\ 1\ \cdots\ 0^{a_k}\ 1 \quad 0^{a_1}\ 1\ 0^{a_2}\ 1\ \cdots\ 0^{a_k}\ 1\ :$$
$$a_1, a_2, \ldots, a_k \geq 0\}.$$

Let $n' = \lfloor n/2 \rfloor$. Then

$$\binom{n'}{k} \leq A_{L_k}(n) \leq \binom{n'}{k} + 2\binom{n'+1}{k} - 1.$$

Closure Properties of DPA

Theorem. The class DPA is closed under union, intersection, complement, and inverse homomorphism.

Proof. Simply adapt the usual constructions.

Theorem. The class DPA is not closed under concatenation.

Proof. Let $L_1 = (0+1)^*$, $L_2 = \{1(0+1)^{2^k} : k \geq 0\}$, and $L = L_1 L_2$. Then $A_{L_1}(n) = O(1)$, and $A_{L_2}(n) = O(n)$, but it can be shown that $A_L(n) \geq 2^{n/3}$ for infinitely many n.

Theorem. The class DPA is not closed under Kleene closure.

Proof. Define $L = \{b(a+b)^{k^2-1} : k \geq 2\}$. Then $L \in \mathrm{DPA}$, but it can be proved that

$$A_{L^*}(n) \geq n^{n^{1/8}/8}$$

for all n sufficiently large.

Lower Bounds for Nondeterministic Automaticity

Theorem. Let L be a language, and suppose there exists a set of pairs $P = \{(x_i, w_i) \ : \ 1 \leq i \leq m\}$ such that

(a) $x_i w_i \in L \ \cup \ \Sigma^{\leq n}$; and

(b) $x_j w_i \notin L$ for $1 \leq i, j \leq m$, $i \neq j$, and $|x_j w_i| \leq n$.

Then $N_L(n) \geq m$.

Proof. Let $M = (Q, \Sigma, \delta, q_0, F)$ be any nondeterministic finite automaton accepting an nth order approximation to L. Now $x_i w_i \in L$, and $|x_i w_i| \leq n$. Since M accepts all strings of length $\leq n$ in L, we have $\delta(q_0, x_i w_i) \cap F \neq \emptyset$. Hence there exists at least one state $q \in \delta(q_0, x_i)$ such that $p \in \delta(q, w_i)$, where $p \in F$.

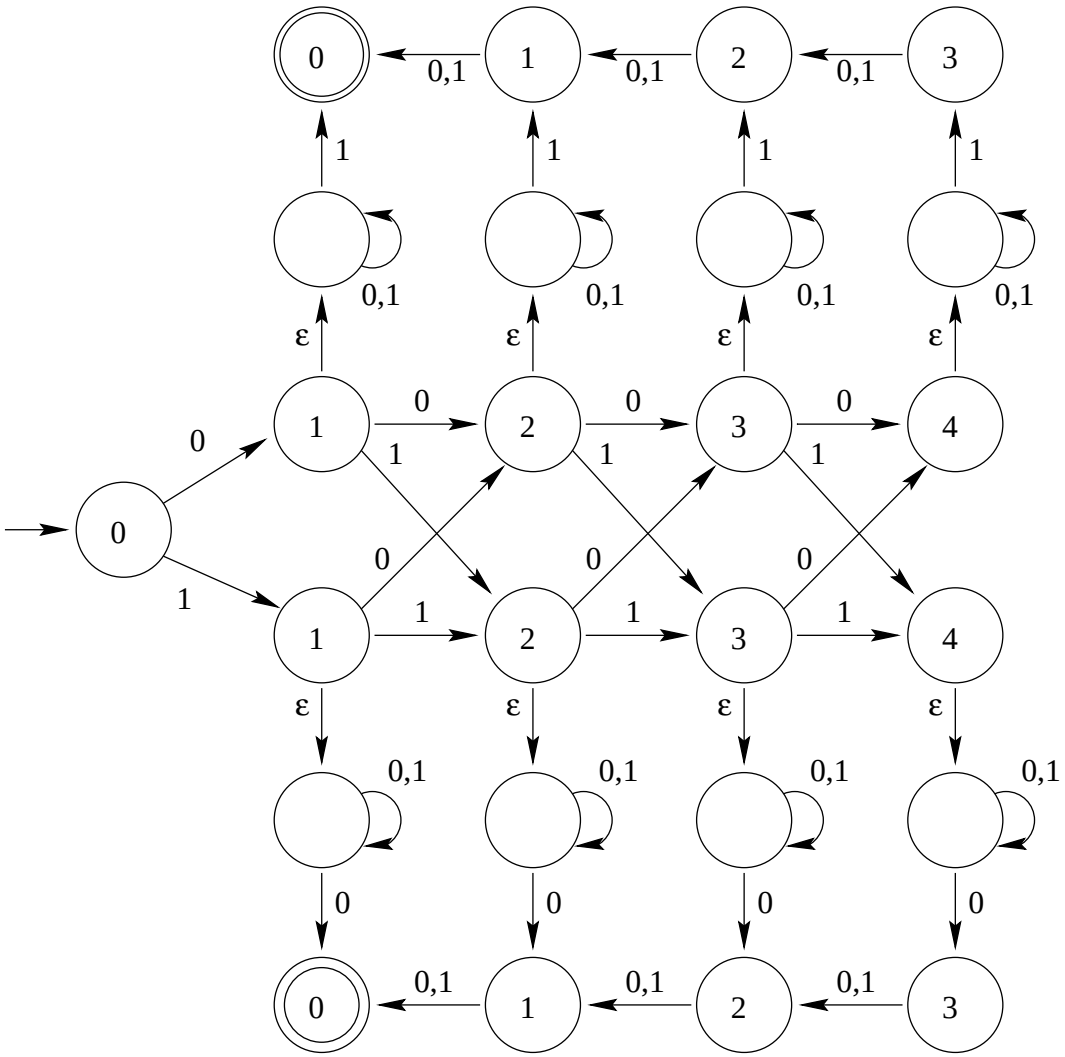
However, for every *other* string $w_j, (j \neq i)$, we must have $q \notin \delta(q_0, x_j)$. For if $q \in \delta(q_0, x_j)$, we would have $p \in \delta(q_0, x_j w_i)$ and so $x_j w_i \in L$, a contradiction, since $|x_j w_i| \leq n$.

Hence every set $\delta(q_0, x_i)$ contains a state q which does not appear in any other set $\delta(q_0, x_j)$ with $j \neq i$. It follows that there must be at least m different states in M.

An Application: NPA not Closed Under Complement

Consider the set $L = \{w \in (0+1)^* \ : \ w \neq w^R\}$, where w^R is the reverse of the string w.

Then $N_L(n) = \Theta(n)$. For the upper bound, note that we can "guess" the position where w differs from w^R and then verify it. For example, the construction for $n = 9$ is given below:



An Application: NPA not Closed Under Complement

However, the complement

$$\overline{L} = \{w \in (0+1)^* \ : \ w = w^R\}$$

is not in NPA. To see this, note that the set

$$S_n = (0+1)^{\lfloor n/2 \rfloor}$$

forms a uniformly n-dissimilar string set for $\overline{L}$, with the "witness" for w being w^R. Hence

$$N_{\overline{L}}(n) \geq 2^{\lfloor n/2 \rfloor},$$

and $\overline{L}$ is not in NPA.

Another Application: NPLA not Closed Under Complement

Let $L = \{w \in (0+1)^* \ : \ |w|_0 \neq |w|_1\}$.

To see that $L \in \mathrm{NPLA}$, we use the following fact from number theory:

> Let $n \geq 2$ and suppose $0 \leq i, j < n$. Then $i \neq j$ iff there exists a prime $p \leq 4.4 \log n$ such that $i \not\equiv j \pmod{p}$.

Thus, to nondeterministically accept some nth order approximation to L, we can

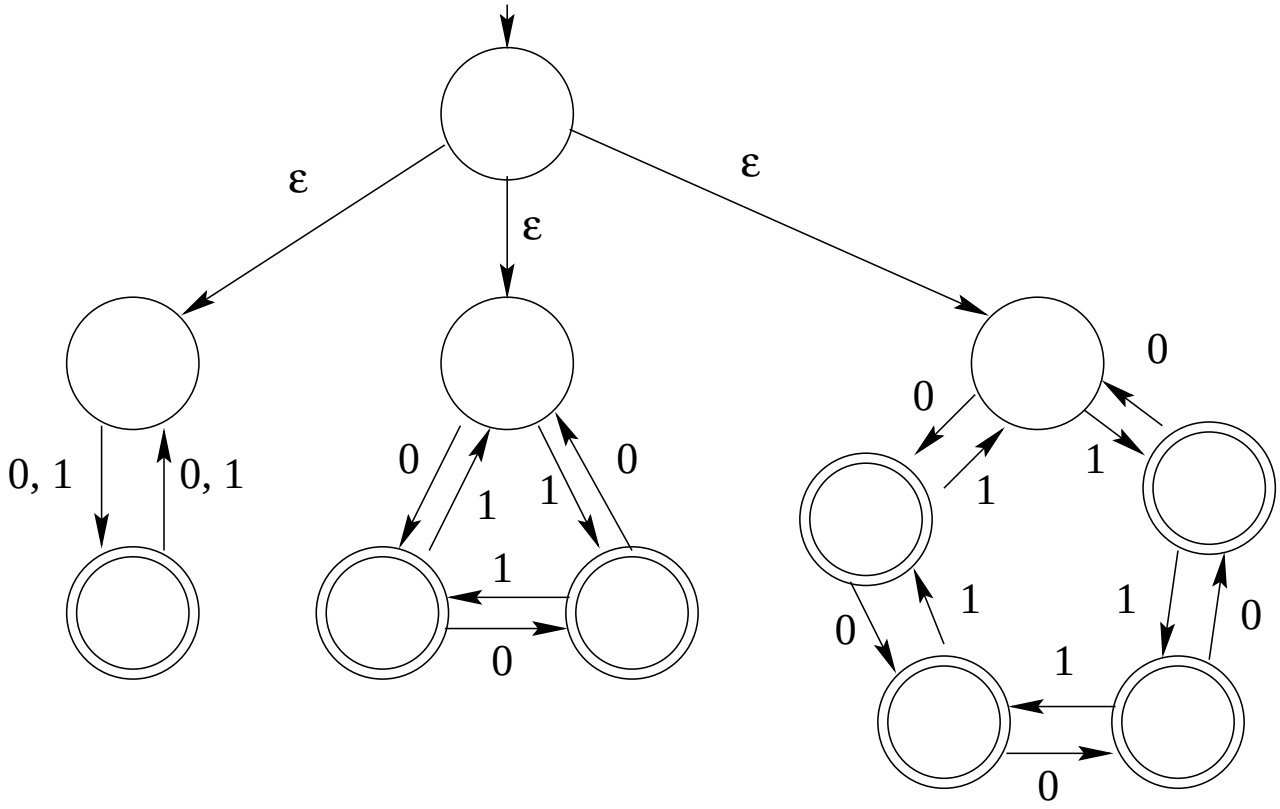
- "guess" the correct prime $p \leq 4.4 \log n$;
- "verify" that $|w|_0 \not\equiv |w|_1 \pmod{p}$.

This construction uses at most

$$1 + \sum_{p \leq 4.4 \log n} p = O((\log n)^2/(\log \log n))$$

states.

Example of Construction



However, $\overline{L} = \{w \in (0+1)^* \ : \ |w|_0 = |w|_1\}$ is not in NPLA. To see this, observe that if it were, then so would

$$L' = \overline{L} \ \cap \ 0^*1^* = \{0^i 1^i \ : \ i \geq 0\}.$$

But L' is not in NPLA. To see this, note that the set

$$\{\epsilon, 0, 00, \ldots, 0^{\lfloor n/2 \rfloor}\}$$

forms a set of uniformly n-dissimilar strings for L'. The "witness" for 0^i is 1^i. Hence $N_{L'}(n) \geq \lfloor n/2 \rfloor + 1$.

An Open Question

Open Question. Is $\mathrm{NPLA} \subseteq \mathrm{DPA}$?

Automaticity of Context-Free Languages

- Breitbart & Shallit gave an example of a context-free language L_s with the minimum possible deterministic automaticity, $A_{L_s}(n) = \lfloor (n+3)/2 \rfloor$.
- What is the *maximum* possible automaticity for a context-free language?

Theorem. For all real $\epsilon > 0$, there exists a CFL of deterministic automaticity $\Omega(2^{n(1-\epsilon)})$.

Proof. Let

$$L_r = \{w\ 0\ 1^a\ 0^b\ :\ w \in (0+1)^*, 1 \le a \le r, b \ge 0,$$
$$w_{-(rb+a)} = 1\}.$$

(Here w_{-i} is the ith symbol from the right in w.) Then it can be shown that

$$A_{L_r}(n) = \Omega(2^{\lfloor rn/(r+1) \rfloor - r}).$$

Automaticity of Sequences

- let $(s_i)_{i \geq 0}$ be a sequence over a finite alphabet
- let $k \geq 2$ be an integer
- define the k-automaticity of the sequence s to be the function $A_s^k(n)$ which counts the number of states in the smallest DFA which correctly computes s_i for $0 \leq i \leq n$.
- our notion of "compute" is that the input i is expressed in base k, and then the digits of the representation are fed into the automaton starting with the least significant digit.
- The output is $\tau(q)$, where q is the last state encountered.
- general problem: can one produce sequences of low and high k-automaticity?
- in particular, can one produce fixed points of homomorphisms of low and high k-automaticity?

Automaticity of Sequences

Theorem. Let $\varphi(c) = cba$, $\varphi(a) = aa$, and $\varphi(b) = b$. Let $(r_i)_{i \geq 0}$ be the fixed point of φ starting with c. Then $A_r^2(n) = \Theta(\log n)$.

Theorem. Let $\varphi(a) = ab$, $\varphi(b) = a$. Let $(s_i)_{i \geq 0}$ be the fixed point of φ starting with a. Then $A_s^k(n) = \Omega(n^{1/5})$ for all $k \geq 2$.

This last result actually applies to *any* Sturmian sequence based on θ, where θ is a real number with bounded partial quotients.

Theorem. Let $\varphi(0) = 01$, $\varphi(1) = 10$. Let $(t_i)_{i \geq 0}$ be the fixed point of φ starting with 0 (the "Thue-Morse" sequence). Then $A_s^k(n) = \Omega(n^{1/4}/\sqrt{k})$ for all odd $k \geq 3$.

The Proof for Sturmian Sequences

Lemma. Let α be an irrational real number, $0 < \alpha < 1$, with partial quotients bounded by B. Let $0 \leq \beta < 1$ be a real number. Then for all $N \geq 1$ there exist integers p, q with $0 \leq p, |q| \leq (B+2)N^2$ such that $|p\alpha - \beta - q| \leq 1/N$.

Lemma. Let $0 < \alpha < 1$ be an irrational real number with partial quotients bounded by B. Define the Sturmian word $s_1 s_2 s_3 \cdots$ by

$$s_i = \lfloor (i+1)\alpha \rfloor - \lfloor \alpha \rfloor$$

for $i \geq 1$. Let $r \geq 2$ be an integer. Then for all integers c, d with $0 \leq c, d < r$, $c \neq d$, there exists an integer m with $0 \leq m \leq 4(B+2)^3 r^3$ such that $s_{rm+c} \neq s_{rm+d}$.

For Further Reading

1. J. Shallit and Y. Breitbart, Automaticity: Properties of a measure of descriptional complexity, in *STACS '94*, Lecture Notes in Comp. Sci. # 775, pp. 619–630. Revised version, to appear, *J. Comput. System Sci.*

2. C. Pomerance and J. M. Robson and J. Shallit, Automaticity II: Descriptional complexity in the unary case, to appear, *Theoret. Comput. Sci.*

3. I. Glaister and J. Shallit, Automaticity III: Polynomial automaticity and context-free languages. To appear, *Computational Complexity*.

4. J. Shallit, Automaticity IV: Sequences, Sets, and Diversity, to appear, *J. de Théorie des Nombres de Bordeaux*.

Copies of these papers are available from:

`http://math.uwaterloo.ca/~shallit`