Automaticity and Rationality

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

Automaticity

- a computable measure of descriptional complexity for languages
- automaticity measures how closely a language can be approximated by regular languages
- goals:
  - complexity measure is a function
  - regular languages should have $O(1)$ automaticity
  - languages "close" to regular should have "small" automaticity

Automaticity Defined

- $\Sigma^{\leq n} = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^{\leq n}$, 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
- 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 automaticity of a language L to be the function

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

where by $|M|$ we mean the number of states in M.

Basic Properties of Automaticity

(a) $A_L(n) \leq A_L(n+1)$ for all $n \geq 0$.

(b) L is regular iff $A_L(n) = O(1)$.

(c) $A_L(n) = A_{\overline{L}}(n)$ for all languages L and $n \geq 0$.

(d) $A_L(n) \leq 2 + \Sigma_{w \in L \cap \Sigma^{\leq n}} |w|$ for all L and $n \geq 0$.

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. $A_L(n) =$ the maximum number of distinct pairwise n–dissimilar strings for L.

An Example

Consider the language

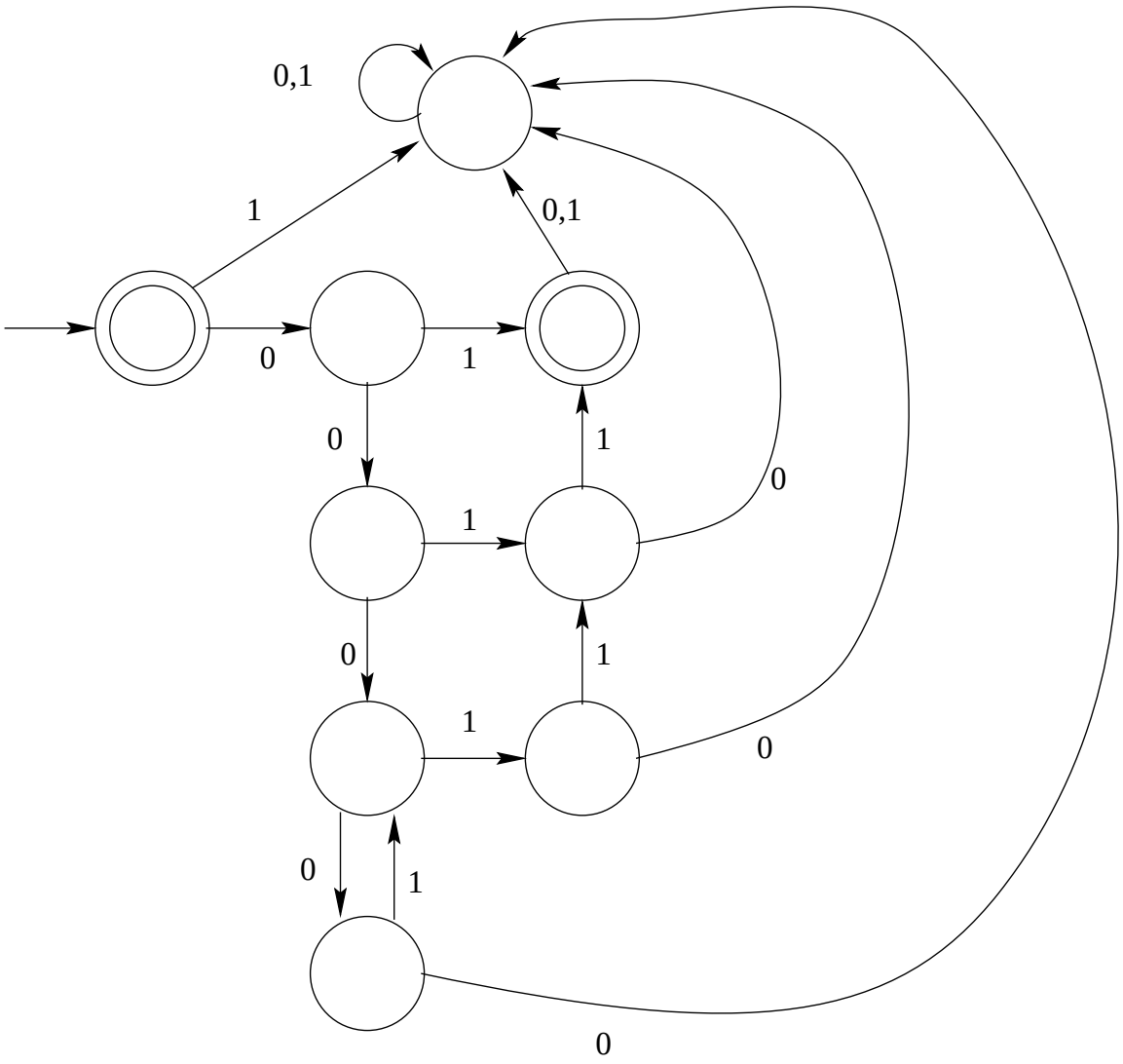
$$L = \{0^n 1^n \ : \ n \geq 0\}.$$

This language is clearly not regular. What is its automaticity?

Automaticity of $L = \{0^n 1^n \ : \ n \geq 0\}$

The automaticity of L is $A_L(n) = 2\lfloor n/2 \rfloor + 1$ for $n \geq 2$.

To see the upper bound, note that we can accept an n'th order approximation to L (for $n = 9$) with the following DFA:



Automaticity of $L = \{0^n 1^n \ : \ n \geq 0\}$

To get the lower bound for $n = 9$, note that we may take

$$\{\epsilon,\ 0,\ 00,\ 000,\ 0000,\ 1,\ 01,\ 001,\ 0001\}$$

as our set of n-dissimilar strings.

This argument easily generalizes to larger n.

Another Example

Consider

$$P = \{10, 11, 101, 111, 1011, 1101, 10001, 10011, \ldots\},$$

the set of primes expressed in base-2.

A classical theorem due to Minsky and Papert (1966) shows that P is not a regular language.

Theorem. The automaticity of P^R is $\Omega(2^{n/43})$.

(Here P^R denotes the reversal of the set P, i.e., the primes expressed with least significant digit first.)

The basic idea is to prove the following

Lemma. Given integers r, a, b with $r \geq 2$, $1 \leq a, b < r$ with $\gcd(r, a) = \gcd(r, b) = 1$, and $a \neq b$, there exists $m = O(r^{165/4})$ such that $rm + a$ is prime and $rm + b$ is composite.

The proof of this lemma is an easy consequence of a deep theorem of Heath-Brown on the distribution of primes in arithmetic progressions ("Linnik's Theorem").

Taking $r = 2^n$, the lemma implies that there are at least $2^{n/43}$ n-dissimilar strings for the language P^R.

Previous Work

- Trakhtenbrot (1964)
  - called automaticity the "weight of a finite tree"
  - used a different model of automaton
  - proved upper and lower bounds

- Grinberg & Korshunov (1966)
  - improved bounds of Trakhtenbrot

- Karp (1967)
  - introduced (but did not name) the concept of automaticity exactly as we defined it
  - proved that if L is not a regular language, then $A_L(n) \geq (n+3)/2$ for infinitely many n
  - observed that the constants 2 and 3 in this theorem are best possible (but proof was flawed)

- Breitbart (1970, 1971, 1973)
  - studied the automaticity of the language $L_{k,p} = \{w \ : \ w$ is the base–p representation of a perfect k'th power $\}$.
  - gave examples of languages with high automaticity

- Paredaens & Vyncke (1977); Gabarró (1983); Balcázar, Díaz, & Gabarró (1985); Serna (1989)
  - studied similar notions

- Dwork and Stockmeyer (1989)
  - introduced "measure of nonregularity" based on n-dissimilar strings.
  - proved: if a language L is recognized by a two-way probabilistic finite automaton with probability $1/2 + \delta$ $(\delta > 0)$ in time $c^{n^{o(1)}}$, then L is regular.

- Kaneps & Freivalds (1990)
  - They defined $r_{\mathrm{sim}}(L, \Sigma^{\leq n})$ = same definition as Dwork & Stockmeyer for n–dissimilar strings
  - $r_{\mathrm{sim}}(L, \Sigma^{\leq n}) = A_L(n)$
  - independently proved Karp's theorem
  - proved: if a language L is recognized by a probabilistic TM with probability $1/2 + \delta$ for some $\delta > 0$ in $O(\log \log n)$, then L is regular

- Koskas and de Mathan (work in progress)
  - Showed how to apply automaticity to obtain irrationality measures in finite characteristic

Bounds on Automaticity

Theorem. Let $k = |\Sigma|$ and $C = 1/((\log_2 k)(k-1)^2)$. Then

$$A_L(n) \leq \frac{Ck^{n+2}}{n}(1 + o(1)).$$

Theorem. Let $k = |\Sigma|$. Then for "almost all" languages $L \subseteq \Sigma^*$, we have

$$A_L(n) > (1 - \epsilon)\frac{Ck^{n+1}}{n}$$

for any fixed $\epsilon > 0$ and all sufficiently large n.

Karp's Theorem

Theorem. Let $L \subseteq \Sigma^*$ be a nonregular language. Then

$$A_L(n) \geq \frac{n+3}{2}$$

for infinitely many n.

Theorem. The constants 3 and 2 in Karp's theorem are best possible, in the sense that the theorem would be false if 2 were replaced with any smaller number, or if 3 were replaced with any larger number.

Automaticity: The Unary Case

- $k = |\Sigma| = 1$
- different issues arise
- we have $A_L(n) \leq n + 1$, for all L and for all n.

Theorem. Let $L \subseteq 0^*$. Then

$$A_L(n) \leq n + 1 - \lfloor \log_2 n \rfloor$$

for infinitely many n.

Theorem. Let $L \subseteq 0^*$. Then for "almost all" L we have

$$A_L(n) > n - 2 \log_2 n - 2 \log_2 \log_2 n$$

for all sufficiently large n.

Can Karp's Theorem be Improved in the Unary Case?

Recall that Karp proved that if L is not regular, then $A_L(n) \geq (n+3)/2$ infinitely often. This implies that

$$\limsup_{n\to\infty} \frac{A_L(n)}{n} \geq \frac{1}{2}$$

for all nonregular L.

Theorem. (Cassaigne) If $L \subseteq 0^*$ is not regular, then

$$\limsup_{n\to\infty} \frac{A_L(n)}{n} \geq \frac{60 - 2\sqrt{10}}{89} \doteq .603095.$$

Nondeterministic Automaticity

Let NFA = class of all nondeterministic finite automata.

The function $N_L(n)$ is the nondeterministic automaticity of the language L, where

$$N_L(n) =$$

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

Theorem. Let $k = |\Sigma| \geq 2$ and $L \subseteq \Sigma^*$. Then $N_L(n) = O(k^{n/2})$.

Theorem. Let $k = |\Sigma| \geq 2$, and let $L \subseteq \Sigma^*$. Then for almost all L and every $\epsilon > 0$ we have

$$N_L(n) > (1 - \epsilon)\frac{k^{n/2}}{\sqrt{k-1}}$$

for all sufficiently large n.

Lower Bounds for Nondeterministic Automaticity

Theorem. Suppose $L \subseteq \Sigma^*$. If L is not regular, then $N_L(n) \geq \log_2((n+3)/2)$ for infinitely many n.

This lower bound is best possible, up to a constant, since the complement $\overline{L}$ of the Stearns-Hartmanis-Lewis language

$$L = \{2\ (1)_2^R\ 2\ (2)_2^R\ 2\ (3)_2^R\ 2\ (4)_2^R\ 2 \cdots\ 2\ (n)_2^R : n \geq 1\}$$

has nondeterministic automaticity $O(\log n)$. Here $(k)_2$ is the representation of k in base-2, and w^R denotes the reversal of the string w.

Another Example of a Language with Low Nondeterministic Automaticity

Theorem. Define

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

Then L is nonregular and

$$N_L(n) = O((\log n)^2/(\log \log n)).$$

Proof. We need 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

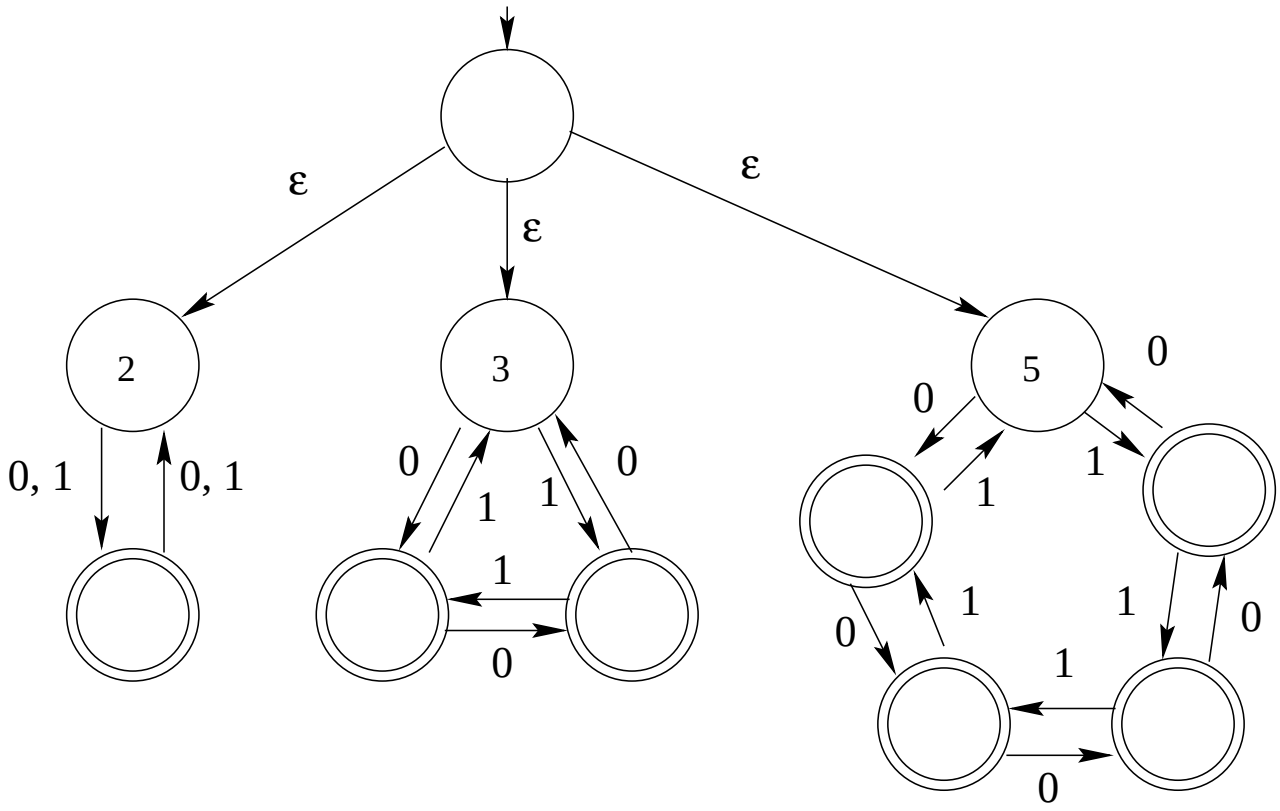
- "guess" the correct prime $p \leq 4.4 \log n$;
- "verify" that $|w|_0 \not\equiv |w|_1 \pmod{p}$.

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



Lower Bounds for Nondeterm. Automaticity: The Unary Case

Theorem. There exists a constant c (which does not depend on L) such that if $L \subseteq 0^*$ is not regular, then

$$N_L(n) \geq c(\log n)^2/(\log \log n)$$

infinitely often.

Open Question: Is this lower bound best possible?

Carl Pomerance has shown that for all monotonically increasing functions f, there exists a language $L = L(f)$ such that

$$N_L(n) = O(f(n)(\log n)^2/(\log \log n)),$$

thus showing the lower bound is essentially tight.

Low Nondeterministic Automaticity: A Unary Example

Theorem. Define $L = \{0^n : n \geq 1$ and the least positive integer not dividing n is not a power of $2\}$. Then L is nonregular and

$$N_L(n) = O((\log n)^3/(\log \log n)).$$

Proof. The construction depends on the following two facts:

> If $0^n \in L$, then there exists a prime power p^k, $p \geq 3$, $k \geq 1$, $p^k \leq 5 \log n$, such that
>
> $$n \not\equiv 0 \pmod{p^k}$$
> $$n \equiv 0 \pmod{2^s}$$
>
> with $2^s < p^k < 2^{s+1}$.
>
> Further, if such a prime power p^k exists, then $0^n \in L$.

An NFA accepting an n-th order approximation to L can now be constructed as follows:

Low Nondeterministic Automaticity: A Unary Example

- "guess" the correct odd prime power $p^k \leq 5 \log n$;
- verify that, on input 0^r, we have
  - $* \; r \not\equiv 0 \pmod{p^k}$
  - $* \; r \equiv 0 \pmod{2^s}$, with $2^s < p^k < 2^{s+1}$.

This construction uses at most $O((\log n)^3/(\log \log n))$ states.

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

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

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 $L_1 = (0+1)^*$ and $L_2 = \{1(0+1)^{2^k}\ :\ k \geq 0\}$. 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.

An Open Question Resolved

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

- There is no reason to believe it to be so (since in principle we might need as many as $2^{O((\log n)^k)}$ states to simulate a nondeterministic machine with $O((\log n)^k)$ states)
- But until recently no example was known.

Theorem. Let

$$L = \{w_1\#w_2\#\cdots\#w_t\#\#w_1\#w_2\#\cdots\#w_t\#\#\#0^{2^t} \ : \ w_i \in \{0,1\}^* \text{ and } |w_1| = |w_2| = \cdots = |w_t| = t\}.$$

Then $A_{\overline{L}}(n) = 2^{\Omega((\log n)^2)}$, so $\overline{L} \notin \mathrm{DPA}$, but $N_{\overline{L}}(n) = O((\log n)^2)$, so $\overline{L} \in \mathrm{NPLA}$.

Proof.

If $x \in L$ and $n := |x|$, then $n = 2t^2 + 2t + 2^t + 3$. Thus $t \approx \log_2 n$. It follows that there are approximately $2^{t^2} = 2^{c(\log n)^2}$ pairwise n-dissimilar strings (take the strings of the form $w_1\#w_2\#\cdots\#w_t$). Hence $A_{\overline{L}} \geq 2^{c(\log n)^2}$.

Proof that NPLA $\not\subseteq$ DPA

On the other hand, we can accept an n'th order approximation to $\overline{L}$ using $O((\log n)^2)$ states. We use nondeterminism to check if any of the equality conditions are violated. More precisely, given a string of the form

$$w_1\#w_2\#\cdots\#w_s\#\#x_1\#x_2\#\cdots\#x_t\#\#\#0^a$$

we can check

- If $|w_1| \neq s$ using $O(\log n)$ states;
- If $|w_1| \neq |w_i|$ or $|w_1| \neq |x_j|$ using $O(\log n)$ states;
- If $w_i \neq x_j$ using $O((\log n)^2)$ states (guessing i and the position where $w_i \neq x_j$, and checking);
- If $a \neq 2^t$ by checking inequality modulo primes $\leq 4.4 \log n$, which can be done using $O((\log n)^2/(\log\log n))$ states.

Rationality

- Instead of languages, we consider formal power series in noncommuting variables, of the form $f = \sum_{w \in \Sigma^*} (f, w) w$
- Instead of finite automata, we consider rational formal power series
  - A formal series is rational if it can be generated by a finite number of applications of $+$ (sum), $\cdot$ (product), or * (quasi-inverse).
  - Example:

    $$(x_0+x_1)^* x_1 (2x_0+2x_1)^* = x_1+x_0x_1+2x_1x_0+3x_1^2+x_0^2x_1+$$

    $$2x_0x_1x_0+3x_0x_1^2+4x_1x_0^2+5x_1x_0x_1+6x_1^2x_0+7x_1^3+\cdots$$

  - By the Kleene-Schützenberger Theorem, a series is rational iff it is recognizable
  - By recognizable, we mean that there exist a matrix-valued homomomorphism μ, and row and column vectors λ, γ, such that $(f, w) = \lambda\ \mu(w)\ \gamma$.
- Rationality measures how closely a formal power series may be approximated by rational functions

Rationality Formally Defined

- If f is a rational series, then it has a linear representation (λ, μ, γ).
- If $\mu(a)$ is an $m \times m$ matrix, then the dimension of the representation (λ, μ, γ) is defined to be m.
- The minimum possible dimension of any linear representation of f is called the rank of f
- The rationality measure $R_f(n)$ is defined to be the minimum possible rank of any rational series g such that $(f, w) = (g, w)$ for all w with $|w| \leq n$.
- Originally introduced by Hespel (1984)

Basic Properties of Rationality

(a) $R_f(n) \leq R_f(n+1)$ for $n \geq 0$.

(b) $R_f(n) \leq A_f(n)$ for $n \geq 0$.

(c) If L is a language, then $R_{\overline{L}}(n) \leq R_L(n)+1$ for $n \geq 0$.

(d) $R_{f+g}(n) \leq R_f(n) + R_g(n)$ for $n \geq 0$.

(e) $R_f(n) = O(1)$ if and only if f is rational.

Upper Bounds on Rationality

Theorem. Let $|\Sigma| = k$, and let $f : \Sigma^* \to K$ be a formal series. Then

$$R_f(n) \leq \begin{cases} n+1, & \text{if } k = 1; \\ \dfrac{2(k^{(n+1)/2} - 1)}{k-1}, & \text{if } k > 1 \text{ and } n \text{ odd}; \\ \dfrac{k^{n/2} + k^{(n+2)/2} - 2}{k-1}, & \text{if } k > 1 \text{ and } n \text{ even}. \end{cases}$$

Expected Value of Rationality

- We give a lower bound for $|\Sigma| = k \geq 2$ and over any finite field.
- Our model is that each coefficient (f, w) is chosen from $\mathrm{GF}(q)$ randomly and uniformly with probability $1/q$.

Theorem. We have

$$E[R_f(n)] \geq \begin{cases} 2^{(n+1)/2} - 1 - O(1), & \text{if } n \text{ is odd and } k = 2; \\ \dfrac{k^{(n+1)/2} - 1}{k - 1} - o(1), & \text{if } n \text{ is odd and } k > 2; \\ k^{n/2} - o(1), & \text{if } n \text{ is even.} \end{cases}$$

An Analogue of Karp's Theorem

Theorem. If f is not rational, then

$$R_f(n) \geq \frac{n+2}{2}$$

for infinitely many n.

Furthermore, this result is best possible.

Rationality in the Unary Case

- If the alphabet Σ is of size 1, then rational functions over Σ correspond to rational functions as usually defined, i.e., as the quotient of two polynomials
- Furthermore, in this case rationality essentially coincides with the well-known measure known as "linear complexity"
- We say that a sequence $s = (s_i)_{i \geq 0}$ satisfies a linear recurrence of order k if there exist constants $a_0, a_1, \ldots, a_k$ with $a_k \neq 0$, such that $\Sigma_{0 \leq j \leq k} a_j s_{i+j} = 0$ for all $i \geq 0$.
- The linear complexity (or linear span) of s, $\mathcal{L}_s(n)$, is defined to be the least k such there exists a sequence $t = (t_i)_{i \geq 0}$ which satisfies a linear recurrence of order k, and further $s_i = t_i$ for $0 \leq i < n$.
- If $f(X) = \Sigma_{i \geq 0} s_i X^i$, then $R_f(n) = \mathcal{L}_s(n+1)$.

Linear Complexity and Continued Fractions

Theorem. (Niederreiter, 1988) A sequence $s = (s_i)_{i \geq 0} \ldots$ satisfies $\mathcal{L}_s(n) = \lfloor (n+1)/2 \rfloor$ for $n \geq 1$ iff the formal series $\sum_{i \geq 0} s_i T^{-(i+1)}$ has a continued fraction expansion $[0,\ a_1,\ a_2,\ a_3, \ldots]$ with $\deg(a_j) = 1$ for all $j \geq 1$.

Example(S, 1979)

$\sum_{i \geq 0} T^{-2^i}$ has continued fraction expansion

$$[0,\ T-1,\ T+2,\ T,\ T,\ T-2,\ T,\ T+2,\ T,\ T-2,\ \ldots]$$

where all the partial quotients (except a_0) have degree 1.

It follows that the sequence

$$(1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, \ldots)$$

has linear complexity $\lfloor (n+1)/2 \rfloor$ and hence the formal series

$$1 + X + X^3 + X^7 + X^{15} + X^{31} + \cdots$$

has rationality $\lfloor (n+2)/2 \rfloor$ for all $n \geq 0$.

Open Problems

1. Estimate $N_L(n)$ where L consists of the base-2 representations of the prime numbers

2. (Norton) Does

   $$\mathcal{L}_f(n) = \lfloor (n+1)/2 \rfloor$$

   for all sufficiently large n, where

   $$f = \Sigma_{i \geq 0} p_{i+1} X^i,$$

   and p_i is the ith prime?

3. Does there exist a unary language L with

   $$N_L(n) = O((\log n)^2/(\log \log n))?$$

4. Give a good estimate for

   $$F(n) := \max_{L \subseteq \{0,1\}^*} A_L(n)$$

   or compute F exactly for $1 \leq n \leq 20$.