# Number Theory and Formal Languages

Jeffrey Shallit
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
shallit@graceland.uwaterloo.ca
http://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 $13$ is $111$ in base $3$, $8$ is $22$ in base $3$, and adding them gives $210$ with two carries in base $3$, and we find $\binom{13}{8} = 3^2 \cdot 11 \cdot 13$.

# 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 $S$ of strings of $0$'s and $1$'s such that every string of $0$'s and $1$'s 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.

## Automatic Sequences

- *Automatic sequences* form the central concept at the intersection of number theory and formal languages
- Many classical sequences, such as the Thue-Morse sequence and the Rudin-Shapiro sequence, turn out to be automatic
- Roughly speaking, a sequence $(s_n)_{n \geq 0}$ is $k$-automatic if $s_n$ is a finite-state function of the base-$k$ expansion of $n$.

# <u>Basics of Finite Automata</u>

- a *deterministic finite automaton with output* (DFAO) is a simple model of a computer
- formally it is a 6-tuple: $M = (Q, \Sigma, \delta, q_0, \Delta, \tau)$ where:
  - $Q$ is a finite set of *states*;
  - the *size* of $M$ is $|M| := |Q|$, the number of states;
  - $\Sigma$ is a finite set of symbols, called the *input alphabet*;
  - $\delta : Q \times \Sigma \to Q$ is the *transition function*
  - $q_0 \in Q$ is the *initial state*;
  - $\Delta$ is the *output alphabet*;
  - $\tau : Q \to \Delta$ is the *output function*
- On input $w = w_1 w_2 \cdots w_n$, the machine enters states $q_0, \delta(q_0, w_1), \ldots , \delta(q_0, w_1 w_2 \cdots w_n)$ and outputs $\tau(\delta(q_0, w_1 w_2 \cdots w_n))$.
- A sequence $(s_n)_{n \geq 0}$ is $k$-automatic iff there exists a DFAO $M$ such that $s_n = \tau(\delta(q_0, w))$, where $w$ is the base-$k$ expansion of $n$.

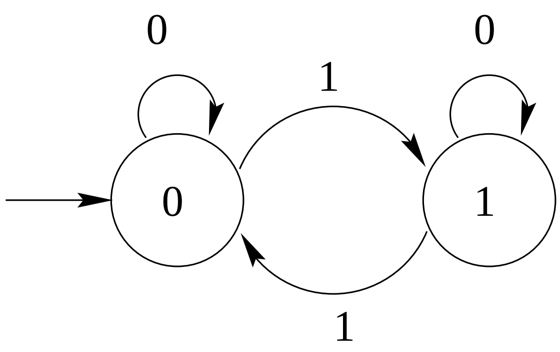
## Example of an Automatic Sequence

The Thue-Morse sequence

$$\begin{aligned} \mathbf{t} &= t_0 t_1 t_2 \cdots \\ &= 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1 \end{aligned}$$

counts the number of $1$'s (mod 2) in the binary expansion of $n$.

This sequence is generated by the following DFAO:




- outputs are written inside states
- start state has a single incoming arrow

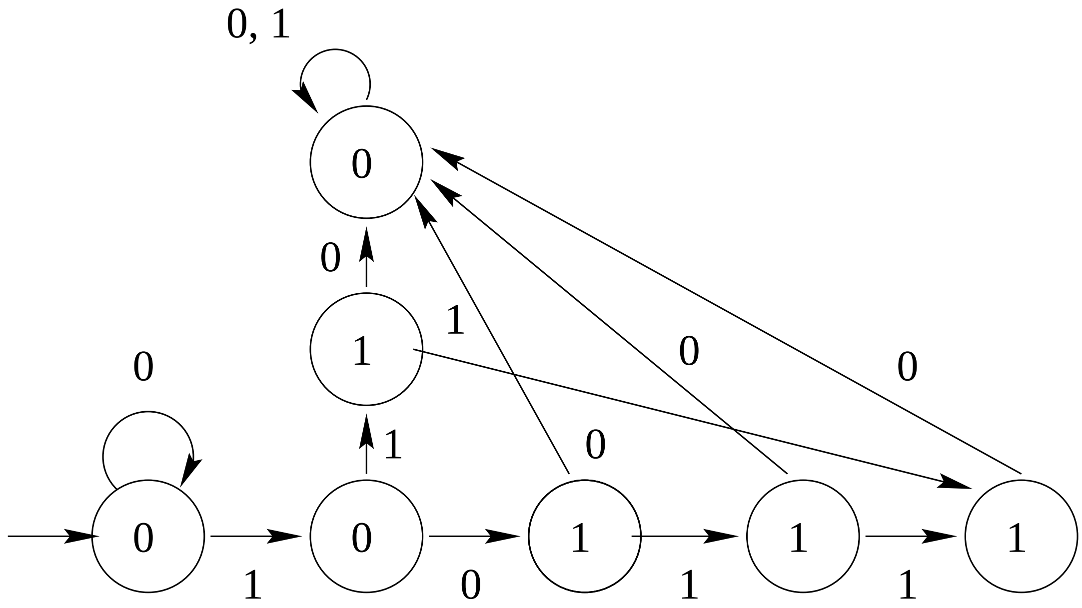
# Automaticity

- Most sequences are not $k$-automatic; for example, the characteristic sequence of $(2^n(2^n-1))_{n\geq 0}$ is not $2$-automatic.
- Can we somehow measure how "close" non-automatic sequences are to being automatic?

# Automaticity

- Yes – with the concept of "automaticity"
- We define the $k$-automaticity of a sequence $(s_i)_{i\geq 0}$ to be the function which counts the number of states in the smallest DFAO $M$ which computes $s_i$ correctly for all $i \leq n$.
- We don't care how $M$ behaves on larger $n$.
- We write $A_s^k(n)$ for this function.

The following DFAO shows that $A_P^2(11) \leq 7$, where $P$ is the characteristic sequence of the prime numbers:



## Basic Properties of Automaticity

1. $A_s^k(n) \leq A_s^k(n+1)$.
2. The sequence $s$ is $k$-automatic iff $A_s^k(n) = O(1)$.
3. $A_s^k(n) = O(n/\log n)$
4. (Karp's theorem) There exists a constant $c$ such that if $s$ is not $k$-automatic, then $A_s^k(n) \geq c \log n$ infinitely often.

## An Example

Consider $P$, the characteristic sequence of the prime numbers

A classical theorem due to Minsky and Papert (1966) shows that $P$ is not $2$-automatic.

**Theorem.** For all $k \geq 2$, we have $A_P^k = \Omega(n^{1/43})$.

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").

## Cobham's Theorem

Cobham (1972) proved the following

**Theorem.** If a sequence is simultaneously $k$- and $l$-automatic, and $k$ and $l$ are multiplicatively independent, then the sequence is ultimately periodic.

- Can this be made more quantitative?
- In some cases, yes.

**Theorem.** Let $k \geq 3$ be an odd integer. Then $A_t^k(n) = \Omega(n^{1/4}k^{-1/2})$, where $t$ is the Thue-Morse sequence.

## Sturmian Sequences

Let $\alpha$ be a real irrational number with $0 < \alpha < 1$. For $i \geq 1$ define

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

The infinite word $\mathbf{s} = \mathbf{s}_\alpha = s_1 s_2 s_3 \cdots$ over $\{0, 1\}$ is sometimes called a *Sturmian* or *characteristic* word, and was studied by Bernoulli, Markov, Christoffel, H. J. S. Smith, etc.

**Example.** Let $\alpha = (\sqrt{5} - 1)/2$. Then

$$\mathbf{s}_\alpha = 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0 \cdots,$$

and $\mathbf{s}_\alpha$ is the fixed point of the homomorphism $1 \to 10$, $0 \to 1$.

# Sturmian Sequences and Automaticity

**Theorem.** Let $0 < \alpha < 1$ be an irrational real number with bounded partial quotients. Let $s_i = \lfloor (i+1)\alpha \rfloor - \lfloor i\alpha \rfloor$ for $i \geq 1$. Then for all $k \geq 2$, the $k$-automaticity of the sequence $(s_i)_{i \geq 1}$ is $\Omega(n^{1/4}/k)$.

The proof depends on the following two lemmas. The first is a version of the traditional inhomogeneous approximation theorem.

**Lemma.** Let $\alpha$ 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 $|p\alpha - \beta - q| \leq \frac{1}{N}$.

**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 $s_i = \lfloor (i+1)\alpha \rfloor - \lfloor i\alpha \rfloor$ 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 $s_{rm+c} \neq s_{rm+d}$.

# Diversity

- We say a sequence $(s(i))_{i\geq 0}$ is $k$-*diverse* if the $k$ subsequences $\{(s(ki+a))_{i\geq 0} \ : \ 0 \leq a < k\}$ are all distinct.
- A sequence is called *diverse* if it is $k$-diverse for all $k \geq 2$.
- If a sequence $s$ is diverse, then for all $n, a, b$ with

  $$0 \leq a, b < n \quad \text{and} \quad a \neq b,$$

  there exists $r$ such that

  $$s(nr+a) \neq s(nr+b).$$

  If there is a function $f$ such that $r = O(f(n))$, then $f$ is said to be a *diversity measure* for $s$.
- As we have already seen, Sturmian sequences corresponding to real numbers with bounded partial quotients have diversity measure $O(n^3)$.

**Theorem.** Almost all binary sequences have diversity measure $4\log_2 n$.

**Open Problem.** Exhibit an explicit example of a binary sequence with this diversity measure.

## For Further Reading


1. P. Enflo, A. Granville, J. Shallit and S. Yu, "On sparse languages $L$ such that $LL = \Sigma^*$", *Discrete Applied Mathematics* **52** (1994), 275–285.
2. J. Shallit and Y. Breitbart, Automaticity: Properties of a measure of descriptional complexity, *J. Comput. System Sci.* **53** (1996), 10–25.
3. J. Shallit, Automaticity IV: Sequences, Sets, and Diversity, *J. de Théorie des Nombres de Bordeaux* **8** (1996), 347–368.