Journal of Automata, Languages and Combinatorics **5** (2000) 3, 255–268

AUTOMATICITY AND RATIONALITY[1]

JEFFREY SHALLIT[2]

Department of Computer Science, University of Waterloo
Waterloo, Ontario N2L 3G1 Canada
e-mail: shallit@graceland.uwaterloo.ca

ABSTRACT

Automaticity is a measure of descriptional complexity for formal languages L, and measures how closely L can be approximated by regular languages. I survey some of the known results and open problems on automaticity. I also discuss a measure which I call "rationality", and explain how it generalizes the well-known concept of linear complexity.

Keywords: Regular language, finite automata, formal power series, rational series, linear complexity, linear span, linear recurrence, linear feedback shift register sequence, automaticity, rationality, continued fraction.

1. Introduction

Let L be a formal language, that is, a subset of Σ^*, where Σ is a finite alphabet. We say L is *regular* if L is accepted by some finite automaton. Of course, not every language is regular, but we can approximate any language arbitrarily closely with regular languages. We say a language L' is an *n'th order approximation* to a language L if L and L' agree on all strings of length $\leq n$, that is, if $L \cap \Sigma^{\leq n} = L' \cap \Sigma^{\leq n}$, where by $\Sigma^{\leq n}$ we mean $(\Sigma \cup \{\epsilon\})^n$. The *automaticity function* $A_L(n)$ is defined to be the least number of states in any deterministic finite automaton (DFA) accepting an n'th order approximation to L.

Example 1 Let $|w|_a$ denote the number of occurrences of the letter a in the string w. Consider the language $L = \{w \in \{0,1\}^* \mid |w|_0 = |w|_1\}$. Then it can be shown that $A_L(n) = n + 1$ for $n \geq 0$.

Similarly, we define the *nondeterministic automaticity function* $N_L(n)$ to be the least number of states in any nondeterministic finite automaton (NFA) accepting an n'th order approximation to L.

[1] Full version of an invited lecture presented at the First International Workshop on *Descriptional Complexity of Automata, Grammars and Related Structures* held in Magdeburg, Germany, July 20 – 23, 1999.

[2] Supported in part by a grant from NSERC.

Example 2 Let L be the language of Example 1. Then it can be shown that $N_{\overline{L}}(n) = O((\log n)^2/(\log\log n))$.

Automaticity is a measure of descriptional complexity for languages, and was first studied (in a slightly different form) by Trakhtenbrot [29]. Dwork and Stockmeyer [4, 5] used it (under the name "nonregularity") to prove that if a two-way probabilistic finite automaton M recognizes a nonregular language with probability $\frac{1}{2} + \delta$ for some fixed $\delta > 0$, then there exists a constant b such that M uses at least 2^{n^b} expected time for infinitely many n. Similar results were obtained by Kaneps and Freivalds [11, 12].

More recently, the author and co-authors [27, 28, 20, 8, 26] studied the problem of approximating L by regular languages, proved new results and unified old results in the area. One of the basic results is Karp's theorem [13, 28]:

Theorem 1 *If L is a nonregular language, then $A_L(n) \geq (n+3)/2$ infinitely often. Furthermore, this bound is best possible, in the sense that the theorem is not true if the* "2" *in the denominator is replaced by any smaller positive real number, or if the* "3" *in the numerator is replaced by any larger real number.*

Deterministic automaticity can also be defined more generally for formal power series. A *formal power series f in non-commuting variables* is a map from $\Sigma^* \to \Delta$, where Δ is a set, possibly of infinite cardinality. The image of a word $w \in \Sigma^*$ under f is denoted (f, w), and we write $f = \sum_{w\in\Sigma^*}(f,w)w$.

A deterministic finite automaton with output (DFAO) is a 6-tuple $M = (Q, \Sigma, \Delta, \delta, q_0, \tau)$, where Q is a finite nonempty set of states, Σ is a finite nonempty set called the input alphabet, Δ is a finite nonempty set called the output alphabet, q_0 is the initial state, $\delta\colon Q \times \Sigma \to Q$ is the transition function, and $\tau\colon Q \to \Delta$ is the output map. We extend the domain of δ to $Q \times \Sigma^*$ in the usual way. We say a DFAO M computes a finite-state function $g_M\colon \Sigma^* \to \Delta$ if $g(w) = \tau(\delta(q_0, w))$ for all $w \in \Sigma^*$.

For any formal series $f : \Sigma^* \to \Delta$, the automaticity function $A_f(n)$ is defined to be the smallest number of states in any DFAO M such that $(f, w) = (g_M, w)$ for all strings w of length $\leq n$. Thus $A_L(n) = A_{\chi(L)}(n)$, where χ is the characteristic function.

Automaticity is computable, and there exists a simple deterministic algorithm for computing $A_f(n)$, given the values of f on all strings of length $\leq n$.

The following theorems give upper bounds on deterministic and nondeterministic automaticity [28]:

Theorem 2 *Let $|\Sigma| = k$, $|\Delta| = l$, and let $f\colon \Sigma^* \to \Delta$. If $2 \leq k$, $l < \infty$, then*

$$A_f(n) \leq (Ck^{n+2}/n)(1 + o(1)),$$

where $C = (\log_k l)/(k-1)^2$.

Theorem 3 *Let $k = |\Sigma| \geq 2$ and let $L \subseteq \Sigma^*$. Then*

$$N_L(n) \leq \begin{cases} \dfrac{2(k^{n/2+1}-1)}{k-1} & \textit{if } n \textit{ even,} \\ \dfrac{k^{(n+1)/2}+k^{(n+3)/2}-2}{k-1} & \textit{if } n \textit{ odd,} \end{cases}$$
$$= O(k^{n/2}).$$

It is an interesting and challenging problem to compute $A_L(n)$ and $N_L(n)$ for specific languages. We do not even know a closed form for $\max_{L\subseteq\Sigma^*} A_L(n)$ when $|\Sigma| \geq 2$. For example, consider the case where $L = \{x \in \{0,1\}^* \mid [x^R]_2 \text{ is a prime}\}$, where by $[z]_b$ we mean the integer represented by the string z considered in base b, and $z \to z^R$ is the reversal function. Then we have [26]:

Theorem 4 *We have $A_L(n) = \Omega(2^{n/43})$.*

There are some natural descriptional complexity classes associated with A_L and N_L. We define the class of deterministic polynomial automaticity (DPA) as follows:

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

We define the class of nondeterministic poly-log automaticity (NPLA) as follows:

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

In [8], the inclusion NPLA $\subseteq$ DPA was stated as an open problem. We resolve this open problem by proving NPLA $\not\subseteq$ DPA:

Theorem 5 *Let*

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

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

Proof. If $x \in L$ and $n := |x|$, then $n = 2t^2 + 2t + 2^t + 3$. Thus $t \approx \log_2 n$. It is not hard to show, using the methods discussed in [8], that $A_L(n) \geq 2^{c(\log n)^2}$. (To see this, note that by considering strings of the form $w_1\#w_2\#\ldots\#w_t$, we find there are approximately $2^{t^2} = 2^{c(\log n)^2}$ pairwise n-dissimilar strings.) Since $A_L(n) = A_{\overline{L}}(n)$, we see that $\overline{L} \notin$ 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

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

of length $\leq n$ we can check

- if $|w_1| \neq s$ using $O(\log n)$ states;
- if $s \neq t$ using $O(\log n)$ states;

- if $|w_1| \neq |w_i|$ or $|w_1| \neq |x_j|$ using $O(\log n)$ states;
- if $s, t, |w_i|, |x_j| \leq \log_2 n$ using $O(\log n)$ states;
- if $w_i \neq x_i$ using $O((\log n)^2)$ states (guessing i and the position where $w_i \neq x_i$, 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.

It follows that $\overline{L} \in$ NPLA. □

2. Formal Power Series and Rationality

Now we consider a "generalization"[3] of automaticity to formal power series with coefficients chosen from a field K. The role of automata is now replaced by rational functions; see, for example, [24, 1]. The main goal of this section and the next is not so much to prove new results as to illustrate how known results from different fields are connected. These connections are particularly poignant for me since, in the unary case, extremal examples are associated with continued fractions I discovered about twenty-five years ago.

More formally, let Σ be an alphabet of cardinality k, e.g., $\Sigma = \{0, 1, \ldots, k-1\}$ or $\Sigma = \{x_0, x_1, \ldots, x_{k-1}\}$. Let K be a field, and let f be a formal power series with coefficients in K. The series f is said to be *rational* if it can be obtained by a finite number of applications of the operations + (sum or union), $\cdot$ (concatenation or product), and * (Kleene star or quasi-inverse). For example, if $f = (2x_0+2x_1)^* x_1 (x_0+x_1)^*$, then the coefficient of $x_{i_1} x_{i_2} \ldots x_{i_r}$ in f is $[i_r i_{r-1} \ldots i_2 i_1]_2$.

A formal series f is *recognizable* if there exists a matrix-valued homomorphism $\mu : \Sigma^* \to K^{m \times m}$ and row and column vectors $\lambda \in K^{1 \times m}$, $\gamma \in K^{m \times 1}$ such that $(f, w) = \lambda \mu(w) \gamma$ for all $w \in \Sigma^*$. We call (λ, μ, γ) a *linear representation* for f. The dimension of the representation (λ, μ, γ) is defined to be m. By the Kleene-Schützenberger Theorem [1, Theorem I.6.1], we know that a series is rational if and only if it is recognizable. The *rank* of a rational series is defined to be the minimum possible dimension of any linear representation of f.

Given a formal series $f : \Sigma^* \to K$, we define the *rationality measure* $R_f(n)$ to be the minimum possible rank of any recognizable (= rational) series g such that $(f, w) = (g, w)$ for all w with $|w| \leq n$. This measure of descriptional complexity (indexed slightly differently) was originally introduced by HESPEL [9]. For a language L, we define $R_L(n) = R_{\chi(L)}(n)$.

First, we prove that the rationality measure $R_f(n)$ is computable. To do so, we introduce (as is usual when dealing with rational series) the notion of *Hankel matrix.* Given a formal power series $f : \Sigma^* \to K$, its associated Hankel matrix H_f is defined to be an infinite matrix with rows and columns indexed by elements of Σ^*, such that the entry in the row indexed by x and column indexed by y is (f, xy).

[3] The reason why "generalization" is in quotes is that in order for it to be a true generalization, we would have to consider semirings instead of fields.

The *rank* of a (possibly infinite) matrix H may be defined as follows: it is 0 if all entries of H are 0; it is t if (a) there exists in H a $t \times t$ submatrix with nonzero determinant and (b) every $(t+1) \times (t+1)$ submatrix has zero determinant; and otherwise the rank is infinite. The following result is well-known (e. g., [6]):

Theorem 6 *The rank of a rational series f is equal to the rank of its associated Hankel matrix H_f.*

Given a (not necessarily rational) formal power series f, we now define its associated *truncated Hankel matrix* $H_f^{(n)}$ as follows: it is a square matrix with $1+k+k^2+\cdots+k^n$ rows and columns, indexed by the elements of $\Sigma^{\leq n}$. The entry in row x and column y is (f, xy), provided that $|xy| \leq n$; otherwise it is a unique indeterminate $z_{x,y}$. The *minrank* of a matrix with field entries and indeterminates is defined to be the least possible rank over all possible substitutions of field elements for the indeterminates. The following theorem was essentially stated by Hespel and Jacob [10]:

Theorem 7 *The minrank of $H_f^{(n)}$ is equal to $R_f(n)$.*

Proof. The procedure given in [2] computes the minrank of an $r \times r$ matrix, provided (a) every variable occurs exactly once and (b) for each row i there exists an integer k_i such that all the field elements in row i appear in columns 1 through k_i, and all the indeterminates appear in columns $k_i + 1$ through r. Evidently the matrix defined above is of this form. The method is to consider each row in turn, and decide if the columns 1 through k_i (i. e., field elements and not indeterminates) linearly depend on previous rows and columns 1 through k_i. If so, the row is discarded. If not, the rank is incremented by one. The rank t computed by this procedure is clearly a lower bound on $R_f(n)$, since no matter what values are chosen for the indeterminates, the resulting Hankel matrix will have rank at least t.

On the other hand, this procedure expresses each row as a linear combination of some set of t rows of $H_f^{(n)}$. From this information we can easily compute a linear representation that is consistent with the provided data on strings of length $\leq n$. This shows that $R_f(n) \leq t$. □

Example 3 Let f be the formal power series over $\{0,1\}^*$ such that $(f,x) = 1$ if $[x]_2 = 0$, and otherwise $(f,x) = p_r$, the r'th prime, if $[x]_2 = r$. Then the associated truncated Hankel matrix is given in Figure 1.

Here the blank entries represent indeterminates. Clearly this truncated matrix has rank 3, and a basis can be formed from the rows r_ϵ, r_1, and r_{10}, corresponding to the respective prefixes.

We can now construct a linear representation of rank 3. To do so, we first need to compute $\mu(0)$, $\mu(1)$ such that

$$\mu(0)\begin{bmatrix} r_\epsilon \\ r_1 \\ r_{10} \end{bmatrix} = \begin{bmatrix} r_0 \\ r_{10} \\ r_{100} \end{bmatrix} \qquad \text{and} \qquad \mu(1)\begin{bmatrix} r_\epsilon \\ r_1 \\ r_{10} \end{bmatrix} = \begin{bmatrix} r_1 \\ r_{11} \\ r_{101} \end{bmatrix}.$$

	ϵ	0	1	00	01	10	11	000	001	010	011	100	101	110	111
ϵ	1	1	2	1	2	3	5	1	2	3	5	7	11	13	17
0	1	1	2	1	2	3	5								
1	2	3	5	7	11	13	17								
00	1	1	2												
01	2	3	5												
10	3	7	11												
11	5	13	17												
000	1														
001	2														
010	3														
011	5														
100	7														
101	11														
110	13														
111	17														

Figure 1: The truncated Hankel matrix $H_f^{(3)}$ in Example 3

There are many choices. One possibility is to choose

$$\mu(0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 2 & 1 & 1 \end{bmatrix} \qquad \text{and} \qquad \mu(1) = \begin{bmatrix} 0 & 1 & 0 \\ -16 & 12 & -1 \\ 2 & 0 & 3 \end{bmatrix}.$$

Finally, we set

$$\lambda = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \gamma = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

and the corresponding rational function approximates f to order 3.

Now that we see that the rationality measure $R_f(n)$ is computable, we turn to the properties of this measure. First, we prove a useful lemma, which is analogous to MOORE's result for the case of finite automata [28, Lemma 4].

Lemma 8 *Let f be a rational series of rank m, and let g be a rational series of rank n. If $(f, w) = (g, w)$ for all w with $|w| \leq m + n - 1$, then $f = g$.*

Proof. Let f have linear representation (λ, μ, γ), and let g have linear representation $(\lambda', \mu', \gamma')$. Then we can form a linear representation for $h = f - g$ of dimension $m + n$ as follows:

$$\lambda'' = \begin{bmatrix} \lambda & -\lambda' \end{bmatrix}; \qquad \mu''(a) = \begin{bmatrix} \mu(a) & 0 \\ 0 & \mu'(a) \end{bmatrix}; \qquad \text{and} \qquad \gamma'' = \begin{bmatrix} \gamma \\ \gamma' \end{bmatrix}.$$

Now by construction $(h, w) = 0$ for all w of length $\leq m + n - 1$. Also, h is of rank $\leq m + n$. Then, by [1, Corollary II.3.6], we have $h = 0$. Hence $f = g$. □

Theorem 9 *Let $f, g \colon \Sigma^* \to K$ be formal series. Then*

(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.*

Proof. (a) If g is an $(n+1)$'th order approximation to f, then g is also an n'th order approximation to f.

(b) If $A_f(n) = q$, then there exists a DFAO $M = (Q, \Sigma, \Delta, \delta, q_0, \tau)$ computing an n'th order approximation to f, with $|Q| = q$. Let

$$\lambda = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}, \qquad \mu(a)_{i,j} = (m_{i,j})_{0 \leq i,j < q}$$

where

$$m_{i,j} = \begin{cases} 1 & \text{if } \delta(i,a) = j; \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\gamma = \begin{bmatrix} \tau(0) \\ \tau(1) \\ \vdots \\ \tau(q-1) \end{bmatrix}.$$

Then it is easy to see that (λ, μ, γ) is a linear representation for g_M, of dimension q. Hence $R_f(n) \leq q$.

(c) Let (λ, μ, γ) be a linear representation of dimension $R_L(n)$ that represents an n'th order approximation to L (or more precisely, $\chi(L)$). We can form a linear representation $(\lambda', \mu', \gamma')$ of dimension $1 + R_L(n)$ that represents an n'th order approximation to $\overline{L}$ as follows:

$$\mu'(a) = \left[\begin{array}{ccc|c} & & & 0 \\ & \mu(a) & & \vdots \\ & & & 0 \\ 0 & \ldots & 0 & 1 \end{array}\right]$$

for $a \in \Sigma$; and

$$\lambda' = \begin{bmatrix} -\lambda & 1 \end{bmatrix}; \qquad \gamma' = \begin{bmatrix} \gamma \\ 1 \end{bmatrix}.$$

(d) Left to the reader.

(e) Suppose f is rational. Then it is of rank d, for some finite d. Clearly $R_f(n) \le d$ for all n.

For the converse, suppose $R_f(n) = O(1)$. Then, since $R_f(n)$ increases monotonically by (a), we know that there exist n_0, d such that $R_f(n) = d$ for all $n \ge n_0$. Note that this means that for each $n \ge n_0$ there exists a linear representation $(\lambda_n, \mu_n, \gamma_n)$ of dimension d such that $(f, w) = \lambda_n \mu_n(w) \gamma_n$ for all w with $|w| \le n$. A priori, it is at least conceivable that each of these representations gives a different rational function. Let us show that for n sufficiently large, each of these linear representations represents the same rational function. Let $n, n' \ge \max(n_0, 2d-1)$, and consider the rational functions $f_n, f_{n'}$ represented by $(\lambda_n, \mu_n, \gamma_n)$ and $(\lambda_{n'}, \mu_{n'}, \gamma_{n'})$. Then f_n and $f_{n'}$ are each of rank d, but agree on all strings of length $\le 2d-1$. By Lemma 8, they must be identical. It follows that $f = f_n = f_{n'}$, and so f is rational. □

We now prove the analogue of KARP's theorem mentioned previously.

Theorem 10 *If f is not rational, then $R_f(n) \ge (n+2)/2$ for infinitely many n.*

Proof. Suppose f is not rational, but (contrary to what we want to prove), there is an n_0 such that $R_f(n) \le (n+1)/2$ for all $n > n_0$. By Theorem 9 (e), we know $R_f(n) \to \infty$ as $n \to \infty$. Hence there exists $r > n_0$ with $R_f(r+1) > R_f(r)$. Let g be a rational function of rank $s = R_f(r+1)$ which forms an $(r+1)$'th order approximation to f. Similarly, let h be a rational function of rank $s' = R_f(r)$ which forms an r'th order approximation to f. Since $R_f(r+1) > R_f(r)$, there exists a word x of length $r+1$ with $(g, x) \ne (h, x)$, and $(g, w) = (h, w)$ for all strings w of length $\le r$. Then, by Lemma 8, we have $r + 1 \le s + s' - 1 \le (r+2)/2 + (r+1)/2 - 1$. Hence we obtain $r + 1 \le r + 1/2$, a contradiction, and the result follows. □

Theorem 11 *Theorem 10 is best possible over $\mathbb{Q}$ and any finite field, in the sense that the 2 in the denominator cannot be replaced by any smaller positive real number, and the 2 in the numerator cannot be replaced by any larger real number.*

Proof. In Section 3, we observe that the formal series $S = \sum_{i \ge 0} X^{2^i - 1}$ satisfies $R_S = \lfloor (n+2)/2 \rfloor$ for all $n \ge 0$. Furthermore, this holds over $\mathbb{Q}$ and any finite field.

Now suppose that f is not rational, but we have $R_f(n) \ge (n+c)/d$ infinitely often, where $0 < d < 2$. (Note that c could conceivably be negative.) Choose k large enough so that $d \le 2 - 1/k$. Then for all $n > 4k - 2kc$ we have

$$2k(n+c) > 2kn + 4k - n - 2 = (n+2)(2k-1).$$

It follows that for all $n > 4k - 2kc$ we have

$$\frac{k(n+c)}{2k-1} > \frac{n+2}{2};$$

hence for infinitely many n and all nonrational f we have

$$R_f(n) \ge \frac{n+c}{d} \ge \frac{n+c}{2-1/k} > \frac{n+2}{2},$$

a contradiction when $f = S$.

Similarly, suppose that for all nonrational f we have $A_f(n) \geq (n+c)/2$ infinitely often, where $c > 2$. But then for all n we have

$$R_S(n) = \lfloor (n+2)/2 \rfloor \leq (n+2)/2 < (n+c)/2,$$

a contradiction. □

We now obtain upper bounds for $R_f(n)$.

Theorem 12 *Let $|\Sigma| = k$, and let $f\colon \Sigma^* \to K$ be a formal series. Then*

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

Furthermore, this bound is best possible.

Proof. For $k = 1$, the associated truncated Hankel matrix is of dimension $(n+1) \times (n+1)$, and so clearly $R_f(n) \leq n+1$. This bound is achieved if $(f, 0^i) = 0$ for $0 \leq i < n$ and $(f, 0^n) = 1$.

For $k \geq 2$, see Hespel [9]. □

It is easy to see that no series f can attain the maximum possible rationality for all n. However, the following example shows that (1) can be achieved up to a constant factor.

Example 4 Let $|\Sigma| = k$, and define a series f by $(f, w) = 1$ if w is a palindrome (equals its reversal), and 0 otherwise. Then it is not hard to prove that $R_f(n) = \frac{k^{\lfloor (n+2)/2 \rfloor}-1}{k-1}$ for $n \geq 0$.

What is the expected value of the rationality $E[R_f(n)]$ of a "randomly-chosen" rational series? In the unary case, this was computed for $K = GF(2)$ by Rueppel [23]. Here 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 $GF(q)$ randomly and uniformly with probability $1/q$.

Theorem 13 *We have*

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

Proof. Gerth [7] observed that the number of $s \times t$ matrices over $GF(q)$ with rank r is given by

$$\prod_{0 \leq i \leq r-1} (q^t - q^i) \frac{q^{s-i}-1}{q^{i+1}-1}.$$

Suppose n is even, and consider the truncated Hankel matrix associated with f. If we consider the rows labelled with strings of length $n/2$, and the columns labelled with strings of length $\leq n/2$, then these elements are all independent, and give a randomly chosen $k^{n/2} \times \frac{k^{(n/2)+1}-1}{k-1}$ matrix. The expected rank $E[R_n]$ of this matrix is a lower bound on $E[R_f(n)]$. By GERTH's result it can be seen that $E[R_n] = k^{n/2} - o(1)$.

Now suppose n is odd. We consider the truncated Hankel matrix with rows labelled with strings of length $(n+1)/2$ and columns labelled with strings of length $\leq (n-1)/2$. These elements are independent, and give a randomly chosen $k^{(n+1)/2} \times \frac{k^{(n+1)/2}-1}{k-1}$ matrix. If $k = 2$, then $\lim_{n\to\infty} E[R_n] = 2^{(n+1)/2} - 1 - c_q$, where c_q is a constant that depends on q (e.g., $c_2 \doteq .460501$, $c_3 \doteq .162089$, etc.). If $k \geq 2$, then $E[R_n] = \frac{k^{(n+1)/2}-1}{k-1} - o(1)$. □

There are also examples of formal series with polynomial rationality of all degrees:

Theorem 14 *For all integers $r \geq 0$ there exists a formal series $f = f_r$ such that $R_f(n) = \Theta(n^r)$.*

Proof. For $r = 0$ this is clear. For $r \geq 1$, let $f = f_r$ be the characteristic series of

$$L_r = \{0^{a_1}\,1\,0^{a_2}\,1\,\ldots\,1\,0^{a_r}\,1\,0^{a_1}\,1\,0^{a_2}\,1\,\ldots\,1\,0^{a_r}\,1 \mid a_1, \ldots, a_k \geq 0\}.$$

Then the proof of [8] can easily be modified to show that $R_f(n) = \Theta(n^r)$. □

3. Rationality in the Unary Case and Linear Complexity

In this section we examine the rationality measure in the unary case, where $|\Sigma| = 1$. It turns out that in this case, the measure of rationality essentially coincides with the well-known concept of linear complexity.

Let $s = (s_i)_{i\geq 0}$ be a sequence over a field K. We say that s satisfies a *linear recurrence of order k* (or is a *linear feedback shift register*) if there exist constants $a_0, a_1, \ldots, a_k$ with $a_k \neq 0$, such that $\sum_{0\leq j\leq k} a_j s_{i+j} = 0$ 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 that 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 $0 \leq i < n$. Linear complexity – particularly when K is a finite field, such as *GF*(2) – has been actively studied in combinatorics and cryptography [22, 18]. In systems and control theory, it is known as the *minimum partial realization* problem [14].

Theorem 15 *Suppose $f(X) = \sum_{i\geq 0} s_i X^i$. Then $R_f(n) = \mathcal{L}_s(n+1)$.*

Proof. Suppose $R_f(n) = t$. Then there exists a $t \times t$ matrix M, and row and column vectors λ, γ such that $(f, 0^i) = \lambda M^i \gamma$. Then, by the Cayley-Hamilton Theorem, M satisfies its own characteristic equation, so there exist constants $a_0, a_1, \ldots, a_{t-1}$ such that

$$M^t + a_{t-1}M^{t-1} + \cdots + a_1 M + a_0 I = 0,$$

where I is the $t \times t$ identity matrix. Multiplying by M^i, we get

$$M^{i+t} + a_{t-1}M^{i+t-1} + \cdots + a_1 M^{i+1} + a_0 M^i = 0$$

for $0 \le i \le n - t$. Hence

$$\lambda M^{i+t}\gamma + a_{t-1}\lambda M^{i+t-1}\gamma + \cdots + a_1 \lambda M^{i+1}\gamma + a_0 \lambda M^i \gamma = 0$$

for $0 \le i \le n - t$, and so we get

$$a_t s_{i+t} + a_{t-1} s_{i+t-1} + \cdots + a_0 s_i = 0$$

for $0 \le i \le n - t$, where $a_t = 1$. It follows that $\mathcal{L}_s(n+1) \le t$.

Now suppose $\mathcal{L}_s(n+1) = t$. Then there exist constants $a_0, a_1, \ldots, a_t$, with $a_t \ne 0$, such that $\sum_{0 \le j \le t} a_j s_{i+j} = 0$ for $0 \le i \le n - t$. Let

$$M = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -\frac{a_0}{a_t} & -\frac{a_1}{a_t} & -\frac{a_2}{a_t} & \ldots & -\frac{a_{t-1}}{a_t} \end{bmatrix}.$$

Then clearly

$$\begin{bmatrix} s_{i+1} \\ s_{i+2} \\ \vdots \\ s_{i+t} \end{bmatrix} = M \begin{bmatrix} s_i \\ s_{i+1} \\ \vdots \\ s_{i+t-1} \end{bmatrix}$$

for $0 \le i \le n - t$. It follows that

$$\begin{bmatrix} s_i \\ s_{i+1} \\ \vdots \\ s_{i+t-1} \end{bmatrix} = M^i \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{t-1} \end{bmatrix}$$

for $0 \le i \le n - t + 1$. Hence, if we define

$$\lambda := \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}; \qquad \gamma := \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{t-1} \end{bmatrix},$$

then $s_i = \lambda M^i \gamma$ for $0 \le i \le n$. It follows that $R_f(n) \le t$. □

One of the basic tools of linear complexity is the Berlekamp-Massey algorithm [15]. This gives an efficient algorithm for computing $\mathcal{L}_s(n)$ in the unary case, as well as the following useful observation: if $\mathcal{L}_s(n+1) > \mathcal{L}_s(n)$, then $\mathcal{L}_s(n+1) = n + 1 - \mathcal{L}_s(n)$. As A. KLAPPER has kindly pointed out to me, this gives an alternative proof of Theorem 10 in the unary case.

There is no known bound for $R(n)$ for the primes similar to that in Theorem 4. However, for the unary case there is the following conjecture, due to G. NORTON [19]:

Conjecture 16 *Let $p = \sum_{i\geq 0} p_{i+1}X^i$, where p_n denotes the n'th prime ($p_1 = 2$). Then $R_p(n) = \lfloor (n+2)/2 \rfloor$ for all sufficiently large n.*

We remark that $R_p(n) = \lfloor (n+2)/2 \rfloor$ for $0 \leq n \leq 24$, except for $n = 6, 7$.

We now turn to examples of unary power series $f \in K[[X]]$ such that $R_f(n) = \lfloor (n+2)/2 \rfloor$ for all $n \geq 0$. It is relatively easy to find such examples when $K = \mathbb{Q}$. For example, one can choose $f := \sum_{i\geq 0} c_i X^i$ such that the c_i grow sufficiently quickly to avoid any linear recurrence, say $c_i = (i+1)!$. It is not difficult to show that

$$\begin{vmatrix} 1! & 2! & 3! & \ldots & r! \\ 2! & 3! & 4! & \ldots & (r+1)! \\ 3! & 4! & 5! & \ldots & (r+2)! \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r! & (r+1)! & (r+2)! & \ldots & (2r-1)! \end{vmatrix} = r! \left(\prod_{1\leq j\leq r-1} j! \right)^2 > 0.$$

However, this example does not easily map to finite fields, since the series f mod p is actually rational for all primes p. We would like to find a series with coefficients in $\{0, 1, -1\}$.

One such power series is $g = \sum_{i\geq 0} X^{2^i - 1}$, which corresponds to the sequence $s = (1, 1, 0, 1, 0, 0, 0, 1, \ldots)$. RUEPPEL conjectured that $\mathcal{L}_s(n) = \lfloor (n+1)/2 \rfloor$, and this was first proven by DAI [3]. In the linear complexity literature, such sequences are said to have "perfect staircase profile" or PSP.

We recall that if K is a field, then formal series in $K[[1/T]]$ have a unique continued fraction expansion $[a_0(T), a_1(T), \ldots]$, where the *partial quotients* a_i are polynomials. Many examples of sequences with PSP can be constructed using the following beautiful connection between linear complexity and continued fractions, due to NIEDERREITER [17].

Theorem 17 *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 $j \geq 1$.*

In 1979 I had showed [25] that $\sum_{i\geq 0} T^{-2^i}$ has continued fraction expansion

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

where all the partial quotients (except a_0) have degree 1. Combining this with Theorem 17, we get a proof of RUEPPEL's conjecture, and this completes the proof of Theorem 11.

MORII and KASAHARA [16] pointed out that the sequence $s' = (1, 0, 1, 0, 0, 0, 1, 0, \ldots)$ which is derived from RUEPPEL's sequence s by a shift, also has PSP. This can also be obtained by combining Theorem 17 with a result due to VAN DER POORTEN and myself [21] that $\sum_{i\geq 1} \pm T^{1-2^i}$ has a continued fraction expansion where all the partial quotients (except a_0) are $\pm T$.

Another example of sequence with PSP (although not with coefficients in $\{0, 1, -1\}$) is $(1, 0, 1, 0, 2, 0, 5, 0, 14, \ldots)$, which corresponds to the power series $C =$

$\sum_{i\geq 0} c_i T^{-(2i+1)}$, where $c_i = \frac{\binom{2i}{i}}{i+1}$ is the i'th *Catalan number.* The continued fraction expansion of C is easily shown to be $[0, T, -T, T, -T, \ldots]$. When this is reduced modulo 2, we obtain the Morii-Kasahara sequence.

Acknowledgements

I wish to thank Andrew Klapper for his expertise and generous advice. I also thank the referee for a careful reading of this paper.

References

[1] J. Berstel, C. Reutenauer, *Rational Series and Their Languages.* Vol. 12 of EATCS Monographs on Theoretical Computer Science, Springer-Verlag, 1988.

[2] J. F. Buss, G. S. Frandsen, J. Shallit, The computational complexity of some problems of linear algebra. *J. Comput. System Sci.* **58** (1999), 572–596.

[3] Z.-D. Dai, Proof of Rueppel's linear complexity conjecture. *IEEE Trans. Inform. Theory* **32** (1986), 440–443.

[4] C. Dwork, L. Stockmeyer, On the power of 2-way probabilistic finite state automata. In: *Proc. 30th Ann. Symp. Found. Comput. Sci.* IEEE Press, 1989, 480–485.

[5] C. Dwork, L. Stockmeyer, A time complexity gap for two-way probabilistic finite-state automata. *SIAM J. Comput.* **19** (1990), 1011–1023.

[6] M. Fliess, Matrices de Hankel. *J. Math. Pures Appl.* **53** (1974), 197–224. Erratum, **54** (1975), 481.

[7] F. Gerth III, Limit probabilities for coranks of matrices over $GF(q)$. *Lin. Mult. Alg.* **19** (1986), 79–93.

[8] I. Glaister, J. Shallit, Automaticity III: Polynomial automaticity and context-free languages. *Computational Complexity* **7** (1998), 371–387.

[9] C. Hespel, Approximation de séries formelles par des séries rationnelles. *RAIRO Inform. Théor. App.* **18** (1984), 241–258.

[10] C. Hespel, G. Jacob, Approximation of nonlinear dynamic systems by rational series. *Theoret. Comput. Sci.* **79** (1991), 151–162.

[11] J. Kaneps, R. Freivalds, Minimal nontrivial space complexity of probabilistic one-way Turing machines. In: B. Rovan (ed.), *Proc. 15th Symposium, Mathematical Foundations of Computer Science 1990.* LNCS **452**, Springer-Verlag, 1990, 355–361.

[12] J. Kaneps, R. Freivalds, Running time to recognize nonregular languages by 2-way probabilistic automata. In: J. Leach Albert, B. Monien, M. Rodríguez Artalejo (eds.), *Proc. 18th Int'l Conf. on Automata, Languages, and Programming (ICALP).* LNCS **510**, Springer-Verlag, 1991, 174–185.

[13] R. M. KARP, Some bounds on the storage requirements of sequential machines and Turing machines. *J. Assoc. Comput. Mach.* **14** (1967), 478–489.

[14] M. KUIJPER, J. C. WILLEMS, On constructing a shortest linear recurrence relation. *IEEE Trans. Automatic Control* **42** (1997), 1554–1558.

[15] J. L. MASSEY, Shift-register synthesis and BCH decoding. *IEEE Trans. Inform. Theory* **15** (1969), 122–127.

[16] M. MORII, M. KASAHARA, Perfect staircase profile of linear complexity for finite sequences. *Inform. Process. Lett.* **44** (1992), 85–89.

[17] H. NIEDERREITER, Sequences with almost perfect linear complexity profile. In: D. CHAUM, W. L. PRICE (eds.), *Advances in Cryptology – EUROCRYPT '87 Proceedings.* LNCS **304**, Springer-Verlag, 1988, 37–51.

[18] H. NIEDERREITER, Cryptology – the mathematical theory of data security. In: T. MITSUI, K. NAGASAKA, T. KANO (eds.), *Prospects of Mathematical Science.* World Scientific, 1988, 189–209.

[19] G. NORTON, On the minimal realizations of a finite sequence. *J. Symbolic Comput.* **20** (1995), 93–115.

[20] C. POMERANCE, J. M. ROBSON, J. SHALLIT, Automaticity II: Descriptional complexity in the unary case. *Theoret. Comput. Sci.* **180** (1997), 181–201.

[21] A. J. VAN DER POORTEN, J. O. SHALLIT, Folded continued fractions. *J. Number Theory* **40** (1992), 237–250.

[22] R. A. RUEPPEL, *Analysis and Design of Stream Ciphers.* Springer-Verlag, 1986.

[23] R. A. RUEPPEL, Linear complexity and random sequences. In: F. PICHLER (ed.), *Advances in Cryptology – EUROCRYPT '85 Proceedings.* LNCS **219**, Springer-Verlag, 1986, 167–188.

[24] A. SALOMAA, M. SOITTOLA, *Automata-Theoretic Aspects of Formal Power Series.* Springer-Verlag, 1978.

[25] J. SHALLIT, Simple continued fractions for some irrational numbers. *J. Number Theory* **11** (1979), 209–217.

[26] J. SHALLIT, Automaticity IV: Sequences, sets, and diversity. *J. Théorie Nombres Bordeaux* **8** (1996), 347–367.

[27] J. SHALLIT, Y. BREITBART, Automaticity: properties of a measure of descriptional complexity. In: P. ENJALBERT et al. (eds.), *STACS 94, Proc. 11th Symp. Theoretical Aspects of Comp. Sci.* LNCS **775**, Springer-Verlag, 1994, 619–630.

[28] J. SHALLIT, Y. BREITBART, Automaticity I: Properties of a measure of descriptional complexity. *J. Comput. System Sci.* **53** (1996), 10–25.

[29] B. A. TRAKHTENBROT, On an estimate for the weight of a finite tree. *Sibirskii Matematicheskii Zhurnal* **5** (1964), 186–191. In Russian.

(Received: October 12, 1999; revised: May 15, 2000)