Journal de Théorie des Nombres
de Bordeaux **8** *(1996), 347–367*

Automaticity IV: Sequences, Sets, and Diversity

par JEFFREY SHALLIT

RÉSUMÉ. Dans cet article nous étudions la complexité de la description (i) de suites sur un alphabet fini et (ii) de sous-ensembles de l'ensemble $\mathbb{N}$ des entiers naturels.

Soit $(s(i))_{i\geq 0}$ une suite sur un alphabet fini Δ. Nous définissons la *k-automaticité* de s, notée $A_s^k(n)$, comme le plus petit nombre possible d'états d'un automate déterministe qui, pour tout i tel que $0 \leq i \leq n$, prend l'expression de i en base k comme entrée et calcule $s(i)$. Nous donnons des exemples de suites qui ont une grande automaticité dans toutes les bases k; par exemple nous montrons que la k-automaticité de la fonction caractéristique des nombres premiers satisfait $A_s^k(n) = \Omega(n^{1/43})$ pour tout $k \geq 2$, rendant ainsi quantitatif le théorème classique de Minsky et Papert suivant lequel l'ensemble des nombres premiers exprimés en base 2 n'est pas régulier.

Nous donnons des exemples de suites dont la k-automaticité est petite dans toute base, ainsi que des exemples de suites dont la k-automaticité est petite dans certaines bases et grande dans d'autres. Nous obtenons aussi des bornes pour l'automaticité de certaines suites qui sont des points fixes d'homomorphismes, par exemple les mots infinis de Fibonacci et de Thue-Morse.

Enfin nous définissons un concept voisin, appelé *diversité*, et nous donnons des exemples de suites ayant une diversité élevée.

ABSTRACT. This paper studies the descriptional complexity of (i) sequences over a finite alphabet; and (ii) subsets of $\mathbb{N}$ (the natural numbers).

If $(s(i))_{i\geq 0}$ is a sequence over a finite alphabet Δ, then we define the *k-automaticity* of s, $A_s^k(n)$, to be the smallest possible number of states in any deterministic finite automaton that, for all i with $0 \leq i \leq n$, takes i expressed in base k as input and computes $s(i)$. We give examples of sequences that have high automaticity in all bases k; for example, we show that the characteristic sequence of the primes has k-automaticity $A_s^k(n) = \Omega(n^{1/43})$ for all $k \geq 2$, thus making quantitative the classical theorem of Minsky and Papert that the set of primes expressed in base 2 is not regular.

Research supported in part by a grant from NSERC.
Manuscrit reçu le 5 avril 1996.

We give examples of sequences with low automaticity in all bases k, and low automaticity in some bases and high in others. We also obtain bounds on the automaticity of certain sequences that are fixed points of homomorphisms, such as the Fibonacci and Thue-Morse infinite words.

Finally, we define a related concept called *diversity* and give examples of sequences with high diversity.

1. Introduction and Definitions

In this paper, I study the descriptional complexity of (i) sequences over a finite alphabet; and (ii) subsets of $\mathbb{N}$ (the natural numbers).

In 1972, Cobham [5] introduced the notion of what is now called a *k-automatic sequence*. (In the literature, one can also find the terms *k-recognizable sequence* and *uniform tag sequence*.) Roughly speaking, a sequence $(s(i))_{i\geq 0}$ over a finite alphabet is k-automatic if and only if $s(i)$ is a finite-state function of the base-k representation of i.

However, most sequences are not k-automatic for any k. Instead of simply saying that a sequence is not k-automatic, we can measure quantitatively how "close" a sequence is to being k-automatic using the concept of *automaticity* studied in previous papers of the author and co-authors [26, 27, 20, 10]. In addition to its evident intrinsic interest, automaticity has proved useful in obtaining nontrivial lower bounds in computational complexity theory; see [7, 8, 16, 17].

More formally, define a deterministic finite automaton with output (DFAO) M to be a 6-tuple, $(Q, \Sigma, \delta, q_0, \Delta, \tau)$, where Q is a finite set of *states*, Σ is a finite *input alphabet*, q_0 is the *start state*, and Δ is a finite *output alphabet*. The map $\delta : Q \times \Sigma \to Q$ is called the *transition function*, and is extended in the obvious way to a map $\delta : Q \times \Sigma^* \to Q$. The map $\tau : Q \to \Delta$ is the *output function*. On input $w \in \Sigma^*$, the machine M outputs the single symbol $\tau(\delta(q_0, w))$. For more on these concepts, see, for example, [15].

Let k be an integer ≥ 2 and define $\Sigma_k = \{0, 1, \ldots, k-1\}$. If $w \in \Sigma_k^*$, then by $[w]_k$ I mean w evaluated as a base-k integer, that is, if $w = w_1 w_2 \cdots w_r$, then $[w]_k = \sum_{1\leq i\leq r} w_{r-i+1} k^{i-1}$. If $n \geq 0$ is an integer, then by $(n)_k$ I mean the default base-k representation of n — that is, one not containing leading zeroes. Note that $(0)_k = \epsilon$, the empty string.

Suppose $(s(i))_{i\geq 0}$ is a sequence over the finite alphabet Δ. If there exists a DFAO M such that for all $i \geq 0$, we have $s(i) = \tau(\delta(q_0, w^R))$ for all $w \in \Sigma_k^*$ such that $[w]_k = i$, then the sequence $(s(i))_{i\geq 0}$ is said to be *k-automatic*. (Here w^R is the reverse of the string w.) Note that the slightly

awkward definition results from the problem of "leading zeroes" input, and our convention that the machine M reads the input number starting with the *least significant* digit.

Here is one alternate definition of k-automatic sequences. Define the *k-fiber of the sequence $(s(i))_{i\geq 0}$ at a* to be

$$\mathcal{F}_k(s,a) = \{(n)_k \ : \ s(n) = a\}.$$

Then $\mathcal{F}_k(s,a)$ is a regular set for all $a \in \Delta$ if and only if the sequence $(s(i))_{i\geq 0}$ is k-automatic.

Another alternate definition of k-automatic sequences can be given in terms of a set called the k-kernel. Let $(s(n))_{n\geq 0}$ be a sequence over a finite alphabet. The *k-kernel* of $(s(n))_{n\geq 0}$, which we denote by K_s^k, is defined as follows:

$$K_s^k = \{(s(k^i m + a))_{m\geq 0} \ : \ i \geq 0,\ 0 \leq a < k^i\}. \tag{1}$$

Eilenberg [9, Proposition 3.3, p. 107] proved that a sequence is k-automatic if and only if its k-kernel is finite.

Given a sequence $(s(i))_{i\geq 0}$, we can define its *k-automaticity* $A_s^k(n)$ as follows: $A_s^k(n)$ is the smallest possible number of states in any DFAO $M = (Q, \Sigma, \delta, q_0, \Delta, \tau)$ such that for all i with $0 \leq i \leq n$, we have $s(i) = \tau(\delta(q_0, w^R))$ for all $w \in \Sigma_k^*$ with $[w]_k = i$. We emphasize that the automaton is fed with the digits of i, starting with the *least significant digit*. This convention is actually important to specify, since it is known that there are languages of low automaticity whose reversal has high automaticity; see [10].

There is another way to define k-automaticity. Suppose we define the *n-truncated k-kernel* of the sequence s, as follows:

$$K_s^k(n) = \{(s(k^i m + a))_{0\leq m\leq (n-a)/k^i} \ : \ i \geq 0,\ 0 \leq a < k^i\}.$$

The n-truncated k-kernel consists of finite sequences. Call two such sequences $v, w \in K_s^k(n)$ *n-dissimilar* if there exists a position j for which both $v(j)$ and $w(j)$ are defined and $v(j) \neq w(j)$. (Note that under this definition, if v is a prefix of w, then v and w are similar.) Then $A_s^k(n)$ is defined to be the maximum number of pairwise n-dissimilar sequences in $K_s^k(n)$. It is not hard to see that this definition is identical to the previous one; see [27]. Note that the condition $m \leq (n-a)/k^i$ is equivalent to $k^i m + a \leq n$; in other words, the variable that is bounded by n is not m but the "true" variable $k^i m + a$.

The following basic results on automaticity are easy to prove [27]:

PROPOSITION 1. *Let $(s(i))_{i\geq 0}$ be a sequence over a finite alphabet Δ. Then*

(a) *$A_s^k(n) \leq A_s^k(n+1)$ for all $n \geq 0$;*

(b) $A_s^k(n) = O(1)$ *if and only if* s *is* k*-automatic;*
(c) *There exists an absolute constant* c *such that if* s *is not* k*-automatic, then* $A_s^k(n) \geq c \log_k n$ *for infinitely many* n.
(d) *For any sequence* s *we have* $A_s^k(n) = O(n/\log_k n)$.

As parts (b) and (c) of this theorem show, if a sequence is not k-automatic, then its k-automaticity must be greater than $c \log_k n$ infinitely often. This suggests studying sequences that are not k-automatic, but which are "as close as possible" to k-automatic. We say that a sequence $(s(i))_{i\geq 0}$ is *k-quasiautomatic* if $A_s^k(n) = O(\log n)$. We then have the following theorem, whose proof is easy and is omitted:

PROPOSITION 2. *A sequence* $(s(i))_{i\geq 0}$ *is* k*-quasiautomatic if and only if it is* k^e*-quasiautomatic for all* $e \geq 1$.

So far we have discussed the k-automaticity of sequences, but the same terminology can be used for sets of non-negative integers. We say a set $S \subseteq \mathbb{N}$ is k-automatic if its characteristic sequence $(\chi_S(n))_{n\geq 0}$ is k-automatic. Similarly, if S is a set, then by $A_S^k(n)$ we mean $A_{\chi_S}^k(n)$.

2. Classical sets with high automaticity in all bases

In this section, we examine two classical sets (the primes, the squarefree numbers) and show that their characteristic sequences have high k-automaticity (that is, $\Omega(n^\epsilon)$ for some $\epsilon > 0$) in all bases $k \geq 2$. (By $f = \Omega(g)$ we mean there exist positive constants c, n_0 such that $f(n) \geq cg(n)$ for all $n \geq n_0$.) For the primes, our results can be viewed as making quantitative the classical result of Minsky and Papert [19] that the primes expressed in base 2 cannot be accepted by a finite automaton.

Our method is based on the following useful lemma:

LEMMA 3. *Let* $(s(i))_{i\geq 0}$ *be a sequence over a finite alphabet* Δ*, and suppose that there exists a constant* d *such that for all* r, a, b *with* $r \geq 2$, $1 \leq a, b < r$, $a \neq b$, *and* $\gcd(r, a) = \gcd(r, b) = 1$, *there exists a non-negative integer* $m = O(r^d)$ *such that* $s(rm + a) \neq s(rm + b)$. *Then* $A_s^k(n) = \Omega(n^{1/(d+1)}/(k \log\log n))$ *for all* $k \geq 2$, *where the implied constant in the big-*Ω *does not depend on* k.

Proof. Since $m = O(r^d)$, there exists a constant c such that $m \leq cr^d - 1$ for all $r \geq 2$. Let $i = \lfloor (\log_k n - \log_k c)/(d+1) \rfloor$. Then

$$\frac{1}{k}\left(\frac{n}{c}\right)^{1/(d+1)} < k^i \leq \left(\frac{n}{c}\right)^{1/(d+1)}.$$

Put $r = k^i$. It follows that there exists $m \leq ck^{id} - 1$ such that $s(k^i m + a) \neq s(k^i m + b)$. However, $k^i m + a < (ck^{id} - 1)k^i + k^i = ck^{i(d+1)} \leq c \cdot (n/c) = n$,

and the same bound holds for $k^i m + b$. It follows that the two subsequences $(s(k^i t + a))_{t \geq 0}$ and $(s(k^i t + b))_{t \geq 0}$ are n-dissimilar. Since a, b were arbitrary integers relatively prime to r, we know that there are at least $\varphi(k^i)$ pairwise n-dissimilar sequences, where φ is Euler's phi-function.

By [21, Theorem 15], we know that $\varphi(n) \geq n/(5 \log\log n)$ for $n \geq 3$. Hence

$$\begin{aligned} \varphi(k^i) &\geq \frac{k^i}{5 \log\log k^i} \\ &\geq \frac{(1/k)(n/c)^{1/(d+1)}}{5 \log\log (n/c)^{1/(d+1)}}. \end{aligned}$$

Thus $A_s^k(n) = \Omega(n^{1/(d+1)}/(k \log\log n))$. ■

We first examine the automaticity of the characteristic sequence of the primes. We need the following lemmas.

LEMMA 4. *For all $x \geq 1$ we have $\prod_{x < p \leq 2x} p > e^{x/3}$, where the product is over primes only.*

Proof. Let $\vartheta(x) = \sum_{p \leq x} \log p$, where the sum is over primes only. We know that $\vartheta(x) < 1.000081x$ for $x > 0$ [22, p. 360], and $\vartheta(x) \geq .84x$ for $x \geq 101$ [21, Theorem 10]. It follows that $\sum_{x < p \leq 2x} \log p > 1.68x - 1.000081x > x/3$ for $x \geq 101/2$. Now it is easily verified by computer or hand calculation that $\sum_{x < p \leq 2x} \log p > x/3$ for $1 \leq x < 101/2$.

It follows that $\prod_{x < p \leq 2x} p > e^{x/3}$ for all $x \geq 1$. ■

LEMMA 5. *Given integers $k, l \geq 1$ with $\gcd(k, l) = 1$, there exists a prime $p = km + l$ with $m = O(\max(k, l)^{11/2})$. The constant in the big-O is independent of k and l.*

Proof. Choose $x = \max(1, l/k, 3 \log l)$. Then from the previous lemma we have $\prod_{x < p \leq 2x} p > l$, so there exists a prime $q \not| l$ with $x < q \leq 2x$. Now $q > l/k$, so $kq > l$, and $\gcd(kq, l) = 1$. Hence by Heath-Brown's version of Linnik's theorem [14], there exists a prime $p \equiv l \ (mod\ kq)$ with $p = O((kq)^{11/2})$. Since $q \leq 2x = 2 \max(1, l/k, 3 \log l)$, we have $p = O(\max(l^{11/2}, (k \log l)^{11/2}, k^{11/2}))$.

Hence $m = (p - l)/k = O(\max(l^{11/2} k^{-1}, k^{9/2} (\log l)^{11/2}), k^{9/2})$, and the result follows. ■

LEMMA 6. *Given integers r, a, b with $r \geq 2$, $\gcd(r, a) = \gcd(r, b) = 1$, $1 \leq a, b < r$, and $a \neq b$, there exists $m = O(r^{165/4})$ such that $rm + a$ is prime and $rm + b$ is composite.*

Proof. We use a trick suggested by papers of Hartmanis and Shank [12] and Allen [1].

By Heath-Brown's version of Linnik's theorem [14], there exists $m_0 = O(r^{9/2})$ such that $p = rm_0 + a$ is a prime. Define $q = rm_0 + b$. Then $q = O(r^{11/2})$. If q is composite, we're done, and $m_0 = O(r^{9/2})$. Otherwise, assume q is prime. Now, in Lemma 5, take $k = qr$ and $l = qr + p$. Then there exists $m_1 = O((qr+p)^{11/2}) = O(r^{143/4})$ such that $(qr)m_1 + (qr+p)$ is prime. However, $t = (qr)m_1 + (qr+q)$ is composite, since $q \mid t$ and $q < t$. Take $m = qm_1 + q + m_0$. Then $m = O(r^{165/4})$. ■

THEOREM 7. *The set P of prime numbers has k-automaticity $A_P^k(n) = \Omega(n^{1/43})$ for all integers $k \geq 2$.*

Proof. Combine Lemmas 3 and 6. ■

We note that the constant $1/43$ in Theorem 7 is not optimal. Indeed, the constant $11/2$ in Lemma 5 is almost certainly not optimal. Wagstaff [31] has provided a heuristic model that predicts that the least prime congruent to $l \pmod k$ is $O(\varphi(k)(\log k)(\log\varphi(k)))$. If this prediction were true, it would improve the constant $1/43$ in Theorem 7 to $1/(2+\epsilon)$.

We now turn to providing a lower bound on the k-automaticity of the squarefree numbers. Recall that a number n is said to be *squarefree* if $t^2 \nmid n$ for all integers $t > 1$.

LEMMA 8. *Let $(s_i)_{i\geq 0}$ be defined as follows:*

$$s_i = \begin{cases} 1, & \text{if } i \text{ is squarefree;} \\ 0, & \text{otherwise.} \end{cases}$$

Then for all $\epsilon > 0$, and r, a, b such that $r \geq 2$, $1 \leq a < r$, and $0 \leq b < r$ with $\gcd(a,r)$ squarefree and $a \neq b$, there exists an $m = O(r^{13/9+\epsilon})$ such that $rm + a$ is squarefree and $rm + b$ is not squarefree.

Proof. Let q be the least prime not dividing $r|b-a|$. Since $r|b-a| < r^2$, by the prime number theorem we have $q = O(\log r^2) = O(\log r)$. Now $rk + b \equiv 0 \pmod{q^2}$ if and only if $k = -br^{-1} \pmod{q^2}$. Let c be such that $0 \leq c < q^2$ and $c \equiv -br^{-1} \pmod{q^2}$. Consider the arithmetic progression

$$((rq^2)m + (rc + a))_{m\geq 0}.$$

We have $\gcd(rq^2, rc+a)$ is squarefree, because any prime divisor of rq^2 and $rc + a$ must be a divisor of r or q^2. But $t \mid r$ and $t \mid rc + a$ implies $t \mid a$, and we know $\gcd(r,a)$ is squarefree by hypothesis. On the other hand, $rc + a \equiv 0 \pmod q$ implies that $rc \equiv -a \pmod q$. But $rc \equiv -b \pmod q$, so $a \equiv b \pmod q$, a contradiction since $q \nmid a - b$. Hence $q^2 \nmid \gcd(rq^2, rc+a)$.

Then, by a result of Heath-Brown [13], there exists an $m_0 = O(r^{13/9+\epsilon})$ such that $(rq^2)m_0 + (rc+a)$ is squarefree. Take $m = q^2 m_0 + c$. Then $rm+a$ is squarefree, but $rm+b$ is divisible by q^2. ■

THEOREM 9. *The set S of squarefree numbers has k-automaticity $A_S^k(n) = \Omega(n^{2/5})$ for all $k \geq 2$.*

Proof. Apply Lemma 3 with $d = 13/9 + \epsilon$. ■

Again, the constant $2/5$ in Theorem 9 is not optimal.

3. A set with low automaticity in all bases

In this section I give an example of a sequence that is k-quasiautomatic for all $k \geq 2$.

THEOREM 10. *Define $a(1) = 1$, and $a(i+1) = a(i) + \prod_{2 \leq b \leq i+2} b^{1+\lfloor \log_b a(i) \rfloor}$ for $i \geq 1$. Then the set $A = \{a(i) : i \geq 1\}$ is not k-automatic, but is k-quasiautomatic for all $k \geq 2$.*

The sequence $(a(i))$ begins

$$1, 3, 39, 331815, 114126085737676800331815, \ldots$$

Proof. First, we note the following observation. Suppose there exists an infinite string $w = w_0 w_1 w_2 \cdots$ over $\Sigma_k = \{0, 1, \ldots, k-1\}$ such that all but finitely many members s of a set S have the "prefix property", that is, $(s)_k^R$ is a prefix of w. Then $A_S^k(n) = O(\log n)$. To see this, note that in this case we can write $S = S_1 \cup S_2$, where S_1 is finite and S_2 has the prefix property. To build an automaton that accepts all the base-k representations of elements of $S_2 \cap [0, n]$, we simply create a linear chain of nodes, with transitions between them labeled with the symbols of w. The accepting states correspond to the members of S_2, and of course we need a single dead state in addition to handle the other transitions. The resulting automaton has $\log_2 n + O(1)$ states.

Since S_1 is finite, we can accept it with a finite automaton. The result now follows because we can accept $S_1 \cup S_2$ using a direct product construction.

The construction of the sequence $(a(i))_{i \geq 1}$ should now be clear. For bases $k \geq 2$, the sequence has the property that $(a(i))_k^R$ is a prefix of $(a(i+1))_k^R$ provided $i \geq k-1$. Hence the observation of the previous two paragraphs applies, and the automaticity of A is $O(\log n)$ for all $k \geq 2$. Note, however, that the constant in the big-O depends on k.

To show that A is not k-automatic for any k, it suffices to show that $\lim_{i \to \infty} a(i)/k^i = \infty$. But this follows, since from the recurrence we have $a(i) \geq i!$. ■

4. Automaticity of fixed points of homomorphisms

Let φ be a homomorphism from Δ^* to Δ^*. If there is a symbol $a \in \Delta$ such that $\varphi(a) = ax$ for some $x \in \Delta^*$, then

$$y = ax\varphi(x)\varphi^2(x)\varphi^3(x)\cdots = \lim_{j\to\infty} \varphi^j(a)$$

is a fixed point of φ; that is, $\varphi(y) = y$. If further φ is *nonerasing* (i.e., $\varphi(b) \neq \epsilon$ for all $b \in \Delta$), then y is infinite. If $|\varphi(b)| = k$ for all $b \in \Delta$, then φ is said to be *k-uniform*. A 1-uniform homomorphism is called a *coding*. A well-known theorem of Cobham [5] states that $(s(i))_{i\geq 0}$ is the image (under a coding) of a fixed point of a k-uniform homomorphism if and only if $(s(i))_{i\geq 0}$ is k-automatic.

A natural problem is to determine the automaticity of fixed points of non-uniform homomorphisms. In particular, are there fixed points of homomorphisms which are quasi-automatic, but not automatic? This question was raised by the author in 1992 in the context of the fixed point $(t_n)_{n\geq 0}$ of the homomorphism $1 \to 121$; $2 \to 12221$. The sequence (t_n) and its relationship to the classical Thue-Morse sequence was studied by Allouche et al. [2]. Computation strongly suggests that (t_n) is 2-quasiautomatic. For example, $t_{16n+1} = t_{64n+1}$ for $0 \leq n \leq 1864134$, but not for $n = 1864135$. Although we are not yet able to prove the 2-quasiautomaticity of (t_n), it is possible to prove that it is not 2-automatic [24]. (This last result was, according to J.-P. Allouche (personal communication), also proved by M. Mkaouar.)

We now give three examples. First, we exhibit a homomorphism whose fixed point is 2-quasiautomatic, but not 2-automatic. Next, we give a homomorphism whose fixed point is 2-automatic, but not k-quasiautomatic for any odd k. Finally, we use some simple theorems of Diophantine approximation to exhibit a homomorphism whose fixed point is not k-quasiautomatic for any $k \geq 2$.

THEOREM 11. *Let $\varphi(c) = cba$, $\varphi(a) = aa$, and $\varphi(b) = b$. Let $(s_i)_{i\geq 0}$ be the fixed point of φ beginning with c. Let $X = \{2^j + j \ : \ j \geq 0\}$. Then*

(a) *$s_0 = c$ and $s_i = b$ if and only if $i \in X$;*
(b) *$(s_i)_{i\geq 0}$ is not 2-automatic.*
(c) *$(s_i)_{i\geq 0}$ is 2-quasiautomatic.*

Proof. Part (a) follows easily from the observation that

$$\varphi^r(c) = cbaba^2ba^4ba^8\cdots ba^{2^{r-1}}.$$

For part (b), it suffices to show that $L = \mathcal{F}_2(s,b)$ is not a regular set. It is easy to see that

$$L = \{1\,0^{n-\lfloor \log_2 n \rfloor - 1}(n)_2 \ : \ n \geq 1\} \ \cup \ \{1\}.$$

Now a routine argument using the pumping lemma [15] completes the proof.

Finally, for part (c), it suffices to construct an automaton with output with $O(\log n)$ states that generates the terms of the sequence (s_i) correctly for all $i \leq n$. We sketch the construction of such an automaton, leaving the details to the reader. Let $\Sigma^{\leq n} = \cup_{0 \leq i \leq n} \Sigma^i$. If L is a language, we say that L' is an *nth-order approximation to* L if $L \cap \Sigma^{\leq n} = L' \cap \Sigma^{\leq n}$. The basic idea of our construction is that it suffices to concentrate on $L^R = \mathcal{F}_2(s,b)^R$ and create an automaton accepting a $(1 + \lfloor \log_2 n \rfloor)$th order approximation to L^R. This is easy, since strings in L^R begin with a short sequence of bits which are followed by many zeroes and then a 1.

The state set consists of four parts. The first part is $A = \{q_w \ : \ w \in (0+1)^{\leq \lfloor \log_2 \log_2 n \rfloor + 1}\}$. This part of the automaton forms a binary tree that can handle all possible strings of length $\lfloor \log_2 \log_2 n \rfloor + 1$. The transitions between states in the first part are given by $\delta(q_w, e) = q_{we}$ for $|w| \leq \lfloor \log_2 \log_2 n \rfloor$ and $e \in \{0,1\}$. The output function for the states in A is given by $\tau(q_\epsilon) = c$, $\tau(q_w) = b$ if $[w^R]_2 \in X$, and $\tau(q_w) = a$ for all other w.

The second part of the automaton consists of a linear chain of states, $B = \{p_i \ : \ 0 \leq i < \lfloor \log_2 n \rfloor\}$. The transitions between states in the second part are given by $\delta(p_i, 0) = p_{i-1}$ for $2 \leq i < \lfloor \log_2 n \rfloor$, $\delta(p_1, 1) = p_0$, and $\delta(p_0, 0) = p_0$. The output function for these states is $\tau(p_i) = a$ for $i \neq 0$, and $\tau(p_0) = b$.

The third part of the state set is $C = \{p'_i \ : \ 0 \leq i < \lfloor \log_2 n \rfloor\}$, a copy of B. The output function for these states is $\tau(p'_i) = b$ for all i.

The fourth and final part consists of a single dead state d. We set $\delta(d, e) = d$ for $e \in \{0,1\}$, and $\tau(d) = a$.

The start state is q_ϵ. We leave to the reader the task of specifying the connections between the different groups of states, observing that transitions $\delta(q_w, 0)$ for $|w| = \lfloor \log_2 \log_2 n \rfloor + 1$ that are not self-loops go to a state in C if $[w^R]_2 \in X$, and otherwise go to a state in B. As an example, the machine in Figure 1 computes (s_i) correctly for all $i < 2^8$. The total number of states needed is $|A| + |B| + |C| + 1 \leq 6 \log_2 n$. ■

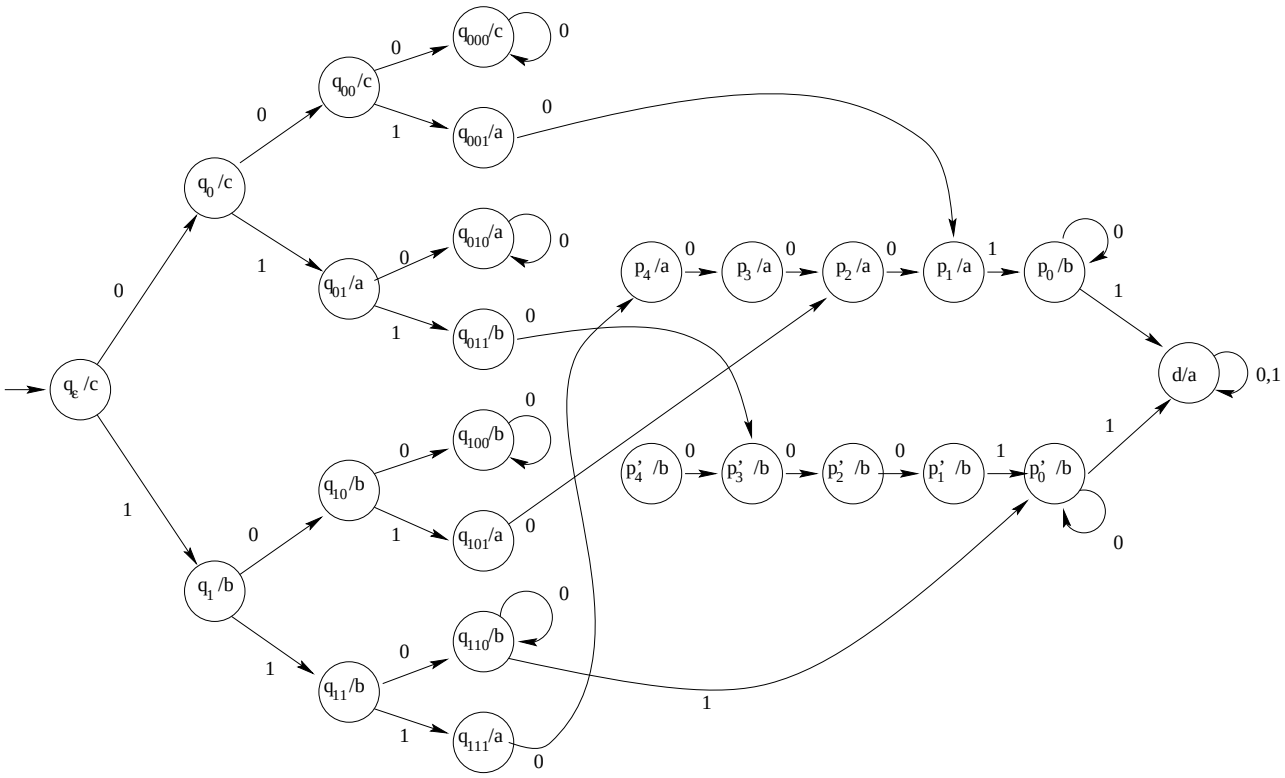

Figure 1: Automaton computing $s(i)$ for $0 \le i < 256$.
The input is the base-2 expansion of i, starting with the least significant bit. The output is $s(i)$. The states are labeled with the name of the state, followed by a slash, followed by the output associated with that state. All unmarked transitions go to the dead state, labeled d/a.

Next, we exhibit a homomorphism whose fixed point is 2-automatic, but has high k-automaticity for all odd k.

Define $\varphi(0) = 01$; $\varphi(1) = 00$, and consider the fixed point $(p(i))_{i\ge 0}$ starting with 0. It is easy to see that $p(i) = \nu_2(i+1) \bmod 2$, where $\nu_2(n)$ is the exponent of the highest power of 2 which divides n.

We first give two simple lemmas:

LEMMA 12. *Let $(s(i))_{i\ge 0}$ be a sequence over a finite alphabet Δ, and suppose that there exists a constant d such that for all r, a, b with $r \ge 2$, $0 \le a, b < r$, and $a \ne b$, there exists a non-negative integer $m = O(r^d)$ such that $s(rm + a) \ne s(rm + b)$. Then $A_s^k(n) = \Omega(n^{1/(d+1)}/k)$ for all $k \ge 2$,*

where the implied constant in the big-Ω does not depend on k.

Proof. Exactly the same as the proof of Lemma 3.

LEMMA 13. *Suppose r is odd and $1 \le a < b \le r$. Then there exists m such that $0 \le m < 4r$ and $\nu_2(rm+a) \not\equiv \nu_2(rm+b) \pmod 2$.*

Proof. Let $b - a = 2^c \cdot t$, where t is odd. Let $m \equiv (2^{c+1} - b)r^{-1} \pmod{2^{c+2}}$; the definition is meaningful since r is odd. Then $rm + b \equiv 2^{c+1} \pmod{2^{c+2}}$, so $\nu_2(rm+b) = c+1$. On the other hand,

$$\begin{aligned} rm + a &\equiv 2^{c+1} + a - b \pmod{2^{c+2}} \\ &\equiv 2^{c+1} - 2^c \cdot t \pmod{2^{c+2}} \\ &\equiv 2^c(2-t) \pmod{2^{c+2}}. \end{aligned}$$

Since t is odd, we have $\nu_2(rm+a) = c$. Now $2^c < r$, so $2^{c+2} < 4r$, and $0 \le m < 2^{c+2}$. ■

Now we can state and prove our theorem on the k-automaticity of $(p(i))_{i \ge 0}$.

THEOREM 14. *If $p(i) = \nu_2(i+1) \bmod 2$, then $(p(i))_{i \ge 0}$ is 2-automatic. If $k \ge 3$ is odd, then $A_p^k(n) = \Omega(n^{1/2}/k)$.*

Proof. The fact that p is 2-automatic follows from the fact that the defining homomorphism φ is 2-uniform; see [5].

To get the automaticity bound for odd k, simply combine Lemmas 12 and 13. ■

As a corollary, we can obtain a lower bound for the automaticity of the Thue-Morse sequence in all odd bases. Let $s_k(i)$ denote the sum of the digits of i when expressed in base k. Then the Thue-Morse sequence $(t(i))_{i \ge 0}$ is defined as follows: $t(i) = s_2(i) \bmod 2$.

It is easy to see that the Thue-Morse sequence is 2-automatic. However, we have the following

THEOREM 15. *Let $k \ge 3$ be an odd integer. Then $A_t^k(n) = \Omega(n^{1/4}/k^{1/2})$.*

Proof. Our proof is based on the following identity, which is well-known and easily proved by considering the base-2 expansion of $i+1$:

$$s_2(i+1) - s_2(i) = 1 - \nu_2(i+1).$$

Taking this modulo 2, we obtain

$$t(i+1) + t(i) + 1 \equiv p(i) \pmod 2, \tag{2}$$

where p is the function defined in Theorem 14.

Let $M_n = (Q, \Sigma_k, \delta, q_0, \Delta, \tau)$ be a DFAO computing $t(i)$ for all i with $0 \le i \le n$, and assume that M_n has $A_t^k(n)$ states. Now consider M_{n+1}, and create a slightly modified automaton $M' = (Q', \Sigma_k, \delta', q_0', \Delta, \tau')$ such that on input w with $[w^R]_k = i$, M' computes the shifted sequence $t(i+1)$ for $0 \le i \le n$. This can be achieved as follows: define $Q' = Q \times \{0, 1\}$, where the second component of every state denotes a carry to be propagated, and let $q_0' = [q_0, 1]$. Define

$$\delta'([q,1],a) = \begin{cases} [\delta(q, a+1), 0], \text{ if } 0 \le a < k-1, \\ [\delta(q,0), 1], \text{ if } a = k-1. \end{cases}$$

Also define $\tau'([q,0]) = \tau(q)$ and $\tau'([q,1]) = \tau(\delta(q,1))$. We leave it to the reader to verify that the construction does indeed compute the shifted sequence. Clearly $|Q'| = 2|Q|$.

We now implement equation (2) by forming the direct product of the automata M_n and M', and using an output function that computes the function $p(i)$ correctly for all i with $0 \le i \le n$. It follows that

$$A_p^k(n) \le 2A_t^k(n)A_t^k(n+1) \le 2(A_t^k(n+1))^2.$$

Since $A_p^k(n) = \Omega(n^{1/2}/k)$, the desired result follows. ■

We now turn to the third problem: finding a fixed point of a homomorphism of high automaticity in all bases. Our methods are based on the theory of Diophantine approximation [4] and *Sturmian words* (also called *characteristic words* or *Christoffel words*). For a survey on Sturmian words, see [3]. First, we introduce some notation.

If α is a real irrational number, we can expand it uniquely as an infinite continued fraction, $\alpha = [a_0, a_1, a_2, \ldots]$. The a_i are called the *partial quotients* of α. We say the partial quotients of α are *bounded by* B if $a_i \le B$ for all $i \ge 1$. (For a survey on bounded partial quotients, see [25].) We define $p_n/q_n = [a_0, a_1, \ldots, a_n]$, and call p_n/q_n the nth *convergent* to α. We define a_n', the nth *complete quotient*, to be $[a_n, a_{n+1}, \ldots]$. We define $\{\alpha\} = \alpha - \lfloor \alpha \rfloor$, the fractional part of α, and $\|\alpha\| = \min(\alpha - \lfloor \alpha \rfloor, \lceil \alpha \rceil - \alpha)$, the distance to the nearest integer.

We then have

LEMMA 16. *Let α be an irrational real number, $0 < \alpha < 1$, with partial quotients bounded by B. Let the numbers $0, \{\alpha\}, \{2\alpha\}, \ldots, \{t\alpha\}, 1$ be arranged in ascending order and let them be labeled $p_0, p_1, p_2, \ldots, p_{t+1}$. Then*

$$\min_{0 \le i \le t} (p_{i+1} - p_i) \ge \frac{1}{(B+2)t}.$$

Proof. Let $(p_k/q_k)_{k\geq 0}$ be the convergents to α. It is a consequence of the three-distance theorem (also called Steinhaus' conjecture) that

$$\min_{0\leq i\leq t}(p_{i+1}-p_i)\geq \|q_{k-1}\alpha\|,$$

where $q_{k-1}\leq t<q_k$. See, for example, [18, Exercise 6.4.8]. (Also see, for example, [29, 30, 28].) Now we know

$$\begin{aligned}
\|q_{k-1}\alpha\| &= |q_{k-1}\alpha - p_{k-1}| \\
&= \frac{1}{a'_k q_{k-1}+q_{k-2}} \qquad \textit{by } [11, p.\ 140] \\
&\geq \frac{1}{(a_k+1)q_{k-1}+q_{k-2}} \\
&= \frac{1}{q_k+q_{k-1}} \\
&= \frac{1}{a_k q_{k-1}+q_{k-2}+q_{k-1}} \\
&\geq \frac{1}{(B+2)q_{k-1}} \\
&\geq \frac{1}{(B+2)t},
\end{aligned}$$

and the result follows. ■

Our next lemma is a version of the inhomogeneous approximation theorem. Unlike the traditional versions of this theorem, the requirement that α has bounded partial quotients allows us to bound the size of the integers that effect the desired approximation.

LEMMA 17. *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 $|p\alpha-\beta-q|\leq\frac{1}{N}$.*

Proof. By Dirichlet's theorem (see, e.g., [4, Theorem I]), there exist integers n,r with $1\leq n\leq N$ and $0\leq r\leq N$, such that $|n\alpha-r|<1/N$. Choose k such that $q_{k-1}\leq N<q_k$, where $(p_k/q_k)_{k\geq 0}$ are the convergents to α. Then, as in the previous theorem,

$$\begin{aligned}
|n\alpha-r| &\geq |q_{k-1}\alpha-p_{k-1}| \qquad (\textit{by } [4, p.\ 2, Eq.\ (4)]) \\
&\geq \frac{1}{(B+2)N}.
\end{aligned}$$

Without loss of generality, assume $n\alpha - r > 0$, and set $p' = \lfloor \beta/(n\alpha - r) \rfloor$. Then $0 \le p' \le (B+2)N$, and $0 \le \beta - (n\alpha - r)p' \le 1/N$. Hence $|p'n\alpha - p'r - \beta| \le 1/N$. Now set $p = p'n$, $q = p'r$; then $p, |q| \le (B+2)N^2$. ■

The next lemma shows that Sturmian sequences corresponding to real numbers with bounded partial quotients have the property that for all pairs of subsequences of the form $(s_{ri+c})_{i\ge 0}$, $(s_{ri+d})_{i\ge 0}$ with $c \ne d$, there is a small witness $i = m$ that shows that these subsequences are different.

LEMMA 18. *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 $i \ge 1$. Let $r \ge 2$ be an integer. Then for all integers c, d with $0 \le c, d < r$, $c \ne d$, there exists an integer m with $0 \le m \le 4(B+2)^3 r^3$ such that $s_{rm+c} \ne s_{rm+d}$.*

Proof. We use the "circular representation" for intervals in $[0,1)$, identifying the point 0 with the point 1, and considering each point modulo 1. Thus, for example, the interval we write as $[2/3, 1/3)$ is really $[2/3, 1) \cup [0, 1/3)$. See, for example, [11, §3.8, §23.2].

It is easy to see that $s_i = 1 \iff \{i\alpha\} \in [1-\alpha, 1)$. Hence if we could find m such that

$$\begin{aligned} \{(rm+c)\alpha\} &\in [1-\alpha, 1); \\ \{(rm+d)\alpha\} &\in [0, 1-\alpha); \end{aligned}$$

it would follow that $s_{rm+c} \ne s_{rm+d}$.

Now

$$\begin{aligned} \{(rm+c)\alpha\} \in [1-\alpha, 1) &\iff \{rm\alpha\} \in I_c := [-(c+1)\alpha, -c\alpha); \\ \{(rm+d)\alpha\} \in [0, 1-\alpha) &\iff \{rm\alpha\} \in I_d := [-d\alpha, -(d+1)\alpha). \end{aligned}$$

We have $\mu(I_c) + \mu(I_d) = 1$, where μ is Lebesgue measure; hence these intervals have nontrivial intersection whenever $c \ne d$. In fact, the endpoints of these intervals are precisely of the form $\{-i\alpha\}$ for some i with $0 \le i \le r$. Let $p_0, p_1, \ldots, p_{m+1}$ denote the points 0, $\{\alpha\}$, $\{2\alpha\}$, ..., $\{r\alpha\}$, 1 arranged in increasing order. It follows that $\mu(I_c \cap I_d) \ge \min_{0\le i\le r}(p_{i+1} - p_i)$, and by Theorem 16, we know this quantity is bounded below by $\frac{1}{(B+2)r}$.

Now let m' be the midpoint of the interval $I_c \cap I_d$. To find m with $s_{rm+c} \ne s_{rm+d}$, it suffices to find integers m, t with

$$\begin{aligned} |rm\alpha - m' - t| &< \frac{\mu(I_c \cap I_d)}{2} \\ &< \frac{1}{2(B+2)r}. \end{aligned}$$

By a folklore result (see, e.g., [23]), since α has partial quotients bounded by B, we know that $r\alpha$ has partial quotients bounded by $r(B+2)$. By Theorem 17, it follows that such an m exists with $m \le r(B+2)(2(B+2)r)^2 = 4(B+2)^3r^3$. ■

THEOREM 19. *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 $i \ge 1$. Then for all $k \ge 2$, the k-automaticity of the sequence $(s_i)_{i\ge 1}$ is $\Omega(n^{1/4}/k)$.*

Proof. Combine Lemmas 12 and 18. ■

It now follows from this result, for example, that the fixed point of the homomorphism $1 \to 10$, $0 \to 1$ is not k-quasiautomatic. This follows because this fixed point can be obtained as a Sturmian sequence by setting $\alpha = (\sqrt{5}-1)/2$. It is known for which α the corresponding Sturmian sequence $(s_i(\alpha))_{i\ge 1}$ is the fixed point of a homomorphism; see [6].

5. Diversity

As we have seen in Section 1, a sequence is k-automatic if and only if its k-kernel (defined in Eq. (1)) is finite. The most spectacular way a sequence can fail to be k-automatic is for *all* the sequences in the k-kernel to be *distinct*; we call such a sequence *strongly k-diverse*. Results of the previous sections suggest that the property of strong diversity and related properties deserve further study.

We make the following definitions:

DEFINITIONS 20. *A sequence $(s(i))_{i\ge 0}$ is weakly k-diverse if the $\varphi(k)$ subsequences $\{(s(ki+a))_{i\ge 0} : \gcd(a,k) = 1, 1 \le a < k\}$ are all distinct. A sequence is weakly diverse if it is weakly k-diverse for all $k \ge 2$.*

A sequence $(s(i))_{i\ge 0}$ is k-diverse if the k subsequences $\{(s(ki+a))_{i\ge 0} : 0 \le a < k\}$ are all distinct. A sequence is diverse if it is k-diverse for all $k \ge 2$.

A sequence $(s(i))_{i\ge 0}$ is strongly k-diverse if the subsequences $\{(s(k^i \cdot j + a))_{i\ge 0} : 0 \le a < k^i,\ i \ge 0\}$ are all distinct. A sequence is strongly diverse if it is strongly k-diverse for all $k \ge 2$.

A sequence is maximally diverse if the subsequences $\{(s(ki+a))_{i\ge 0} : 0 \le a < k,\ k \ge 1\}$ are all distinct.

The results of previous sections can now be rephrased in the language of diversity. In Section 2 we showed that the characteristic sequences of the primes and squarefree numbers are weakly diverse. In Section 4 we showed that the sequence $(\nu_2(i+1))_{i\ge 0}$ is k-diverse for all odd k, and we also showed that if α is a real number with bounded partial quotients, then $(s_i(\alpha))_{i\ge 0}$ is diverse.

We now give an example of a sequence that is strongly k-diverse for $k = 2$. Consider the set $X = \{2^j + j \ : \ j \geq 0\}$ introduced in Theorem 11, and let $(c(i))_{i\geq 0}$ be the characteristic sequence of this set. Then we have the following theorem:

THEOREM 21. *The sequence $(c(i))_{i\geq 0}$ is strongly 2-diverse.*

Proof. We must show that, given any four integers j, k, a, b with $j, k \geq 0$, $0 \leq a < 2^j$, $0 \leq b < 2^k$, and $(j, a) \neq (k, b)$, there exists $n \geq 0$ with $c_{2^j n+a} \neq c_{2^k n+b}$. Without loss of generality, assume $j \leq k$ and if $j = k$, then $a < b$. Let $n_0 = 2^{k+1}$ and set $n = 2^{n_0 \cdot 2^j - j + a} + n_0$. Then $2^j n + a = 2^{n_0 \cdot 2^j + a} + n_0 \cdot 2^j + a = 2^i + i$ for $i = n_0 \cdot 2^j + a$. Hence $c_{2^j n+a} = 1$.

It remains to show $c_{2^k n+b} = 0$. To see this, it suffices to show that

$$2^i + i < 2^k n + b < 2^{i+1} + i + 1$$

for $i = n_0 \cdot 2^j + k - j + a$.

To prove the first inequality, it suffices to show that

$$n_0 \cdot 2^j + k - j + a < n_0 \cdot 2^k + b.$$

There are three cases to examine.

(i) If $k = j$, then this inequality follows from the assumption that $a < b$.

(ii) If $k = j + 1$, then we must show $a - b + 1 < n_0(2^k - 2^{k-1}) = n_0 2^{k-1}$. Since $a < 2^j = 2^{k-1}$, it suffices to show $2^{k-1} < n_0(2^k - 2^{k-1}) = n_0 2^{k-1}$, which is true since $n_0 \geq 2$.

(iii) If $k \geq j + 2$, then we must show $k - j + 2^j < n_0(2^k - 2^j)$. Now $k - j < 2^{k-j} < 2^k$, and $2^j < 2^k - 2 \cdot 2^j$ provided $2^k > 3 \cdot 2^j$, which is true since $k \geq j + 2$. Adding these inequalities, we find $k - j + 2^j < 2(2^k - 2^j) \leq n_0(2^k - 2^j)$, as desired.

To prove the second inequality, we must show

$$n_0(2^k - 2^j) + j - k + b - a < 2^{n_0 \cdot 2^j + k - j + a} + 1.$$

There are two cases to consider.

(i) If $k = j$, we must show $b - a < 2^{n_0 \cdot 2^j + a} + 1$. Since $b < 2^k$, it suffices to show $2^k < 2^{n_0} + 1$ and for this it suffices to take $n_0 \geq k$.

(ii) If $k > j$, then $n_0(2^k - 2^j) + j - k + b - a < n_0(2^k - 2^j) + 2^k < (n_0 + 1)2^k$. Thus it suffices to show $(n_0 + 1)2^k < 2^{n_0}$. Choose $n_0 \geq 2^{k+1}$. Then $(n_0 + 1)2^k \leq (n_0 + 1)n_0/2 < 2^{n_0}$ provided $n_0 \geq 2$, which it is, since $n_0 \geq 2^{k+1}$. ■

We now show the following:

THEOREM 22. *Almost all sequences over $\{0, 1\}$ are maximally diverse.*

Proof. Since the set of pairs $\{(k,a) \ : \ 0 \le a < k, \ k \ge 1\}$ is countably infinite, it suffices to show that if $(k,a) \ne (l,b)$, then the set of sequences $(s(i))_{i \ge 0}$ for which $s(ki+a) = s(li+b)$ for all $i \ge 0$ is of measure zero.

Let $g = \gcd(k,l)$. If $g \not| b-a$, or if $k = l$ and and $a \ne b$, then the linear progressions $(ki+a)_{i \ge 0}$ and $(lj+b)_{j \ge 0}$ contain no terms at all in common. Therefore the subsequences $(s(ki+a))_{i \ge 0}$ and $(s(lj+b))_{j \ge 0}$ are independent and hence the probability that they are identical is 0.

Otherwise, assume $k \ne l$ and $g \,|\, b-a$. In order that $(k/g)i - (l/g)j = (b-a)/g$, we must have $i = (l/g)i' + i_0$, and $j = (k/g)i' + j_0$, for some constants i_0, j_0. Since $k \ne l$, at least one of $l/g, k/g$ must be different from 1. Without loss of generality, assume it is l/g. Choose a constant $i_1 \ne i_0$; then the set

$$\{k((l/g)i' + i_0) + a \ : \ i' \ge 0\}$$

contains no terms in common with

$$\{lj + b \ : \ j \ge 0\}.$$

Hence the subsequences $(s(k((l/g)i' + i_0) + a))_{i' \ge 0}$ and $(s(l((l/g)i' + i_0) + b))_{i' \ge 0}$ are independent, and so the probability that they are identical is zero. ■

Although almost all 0, 1-sequences are maximally diverse, it is not so easy to prove that any individual sequence has the maximally diverse property. We now give some examples of maximally diverse sequences.

Let α be a real irrational number with $0 < \alpha < 1$. Recall the definition of the Sturmian infinite word $(s_i)_{i \ge 0}$ from Section 4:

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

THEOREM 23. *All Sturmian sequences are maximally diverse.*

Proof. We must show that given $j,k \ge 1$, $0 \le a < j$, $0 \le b < k$, and $(j,a) \ne (k,b)$, there exists an $n \ge 0$ such that $s_{jn+a} \ne s_{kn+b}$.

It is easy to see that $s_n = 1 \iff \{n\alpha\} \in [1-\alpha, 1)$. Hence it suffices to exhibit n such that

$$\begin{aligned} \{(jn+a)\alpha\} &\in [1-\alpha, 1) \\ \{(kn+b)\alpha\} &\in [0, 1-\alpha) \end{aligned}$$

Now

$$\begin{aligned} \{(jn+a)\alpha\} \in [1-\alpha, 1) &\iff \{jn\alpha\} \in I_a := [-(a+1)\alpha, -a\alpha); \\ \{(kn+b)\alpha\} \in [0, 1-\alpha) &\iff \{kn\alpha\} \in I_b := [-b\alpha, -(b+1)\alpha). \end{aligned}$$

First, let us consider the case $j = k$. We have $a \ne b$. In this case, we have $\mu(I_a) + \mu(I_b) = 1$, and since $a \ne b$, these intervals must have nontrivial

intersection. Define $I = I_a \cap I_b$; then $\mu(I) > 0$. It now suffices to choose n such that $\{jn\alpha\} \in I$. Such an n exists by Kronecker's theorem (e.g., [11, Theorem 438]).

Second, let us consider the case $j \neq k$. Without loss of generality, let us assume $j < k$. Define $p(I)$, the projection of an ordinary interval I, to be $p(I) = \{\{x\} \ : \ x \in I\}$. Thus, for example, $p([e, \pi)) = [e-2, \pi-3)$.

Consider I_a, and let its left and right endpoints be t and u respectively. If I_a "wraps around" 0, then choose $u \in [1,2)$ so that $p([t,u)) = I_a$. Define $I_0 = p([t/j, u/j))$, and $I_1 = p([kt/j, ku/j))$.

I claim that $\mu(I_1) > \mu(I_a)$. For if I_a contained a subinterval of measure $\geq j/k$, then $I_1 = [0,1)$, and so $\mu(I_1) = 1 > \mu(I_a)$. Otherwise, I_a contains no subinterval of measure $\geq j/k$, so $\mu(I_a) < j/k$. In this case, $\mu(I_1) = k/j\mu(I_a) > \mu(I_a)$.

Now $\mu(I_a) + \mu(I_b) = 1$. Hence $\mu(I_1) + \mu(I_b) > 1$ and so I_1 and I_b have nontrivial intersection. Let $I_2 = I_1 \cap I_b$; then $\mu(I_2) > 0$. By our definition of I_0 and I_1, there is an interval $I_3 \subseteq I_0$ such that if $I_3 = [v,w]$, then $[kv, kw] \subseteq I_2$. Also, since $I_3 \subseteq I_0$, it is clear that $[jv, jw] \subseteq I_a$. Again, by Kronecker's theorem, we can find n such that $\{n\alpha\} \in I_3$. For this n we have $\{jn\alpha\} \in I_a$ and $\{kn\alpha\} \in I_2 \subseteq I_b$, as desired. ■

If a sequence $(d(i))_{i \geq 0}$ is diverse, then we know that for all r, a, b with $0 \leq a, b < r$ and $a \neq b$, there exists an m such that $d(rm+a) \neq d(rm+b)$. If there is a function f such that $m = O(f(n))$, then $f(n)$ is said to be a *diversity measure* for d. In a previous section we showed, for example, that the diversity measure for Sturmian sequences corresponding to real numbers with bounded partial quotients is $O(r^3)$.

We now show that the diversity measure for almost all sequences is low:

THEOREM 24. *Almost all binary sequences have the property that for all $r \geq 1$, and for all a, b with $0 \leq a, b < r$ and $a \neq b$, there exists $m = O(\log r)$ such that $s_{rm+a} \neq s_{rm+b}$.*

Proof. By Theorem 22, we may restrict our attention to sequences that are diverse.

We have

$$\Pr[s_{rm+a} = s_{rm+b} \ for \ 0 \leq m < f(r)] = 2^{-f(r)}.$$

It follows that

(3)

$$\Pr[\exists \ at \ least \ one \ pair \ (a,b) \ such \ that \ s_{rm+a} = s_{rm+b} \ for \ 0 \leq n < f(r)] =$$

$$= \frac{r(r-1)}{2} 2^{-f(r)}.$$

Choose $f(r) = \lceil 4 \log_2 r \rceil$; then $\frac{r(r-1)}{2} 2^{-f(r)} = \Theta(r^{-2})$. (Here $f = \Theta(g)$ means $f = O(g)$ and $g = O(f)$.) Then $\sum_{r \geq 1} \frac{r(r-1)}{2} 2^{-f(r)}$ converges. Hence by the Borel-Cantelli lemma, with probability 1 at most finitely many of the events of the form (3) occur. That is, with probability 1, the event

$$\forall \; pairs \; (a,b) \; \exists m < \lceil 4 \log_2 r \rceil \; such \; that \; s_{rm+a} \neq s_{rm+b}$$

occurs all but finitely many times. Hence with probability 1 we have $m = O(\log r)$. ■

Interestingly enough, I do not know a single *explicit* example of a diverse sequence with diversity measure $O(\log n)$.

6. Acknowledgments.

I thank Mark Turner for suggesting the term "diversity". Drew Vandeth, Eric Bach, and Jean-Paul Allouche read a draft of this paper and made many useful suggestions. I thank the referee for his comments.

References

[1] D. Allen, Jr. On a characterization of the nonregular set of primes. *J. Comput. System Sci.* **2** (1968), 464–467.

[2] J.-P. Allouche, A. Arnold, J. Berstel, S. Brlek, W. Jockush, S. Plouffe, and B. E. Sagan. A relative of the Thue-Morse sequence. *Discrete Math.* **139** (1995), 455–461.

[3] J. Berstel. Recent results on Sturmian words. To appear, Proc. of DLT '95. Also available at
`http://www-litp.ibp.fr/berstel/Liaisons/magdeburg.ps.gz`

[4] J. W. S. Cassels. *An Introduction to Diophantine Approximation.* Cambridge University Press, 1957.

[5] A. Cobham. Uniform tag sequences. *Math. Systems Theory* **6** (1972), 164–192.

[6] D. Crisp, W. Moran, A. Pollington, and P. Shiue. Substitution invariant cutting sequences. *J. Théorie Nombres Bordeaux* **5** (1993), 123–137.

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

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

[9] S. Eilenberg. *Automata, Languages, and Machines*, Vol. A. Academic Press, 1974.

[10] I. Glaister and J. Shallit. Automaticity III: Polynomial automaticity and context-free languages. Submitted, 1996.

[11] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, 1989.

[12] J. Hartmanis and H. Shanks. On the recognition of primes by automata. *J. Assoc. Mach.* **15** (1968), 382–389.

[13] D. R. Heath-Brown. The least square-free number in an arithmetic progression. *J. Reine Angew. Math.* **332** (1982), 204–220.

[14] D. R. Heath-Brown. Zero-free regions for Dirichlet L-functions and the least prime in an arithmetic progression. *Proc. Lond. Math. Soc.* **64** (1992), 265–338.

[15] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 1979.

[16] J. Kaneps and R. Freivalds. Minimal nontrivial space complexity of probabilistic one-way Turing machines. In B. Rovan, editor, *MCFS '90 (Mathematical Foundations of Computer Science*, Vol. 452 of *Lecture Notes in Computer Sciences,* pages 355–361. Springer-Verlag, 1990.

[17] J. Kaneps and R. Freivalds. Running time to recognize nonregular languages by 2-way probabilistic automata. In J. Leach Albert, B. Monien, and M. Rodriguez Artalejo, editors, *ICALP '91 (18th international Colloquium on Automata, Languages, and Programming*, Vol 510 of *lecture notes in Computer sciences*, pages 174–185. Springer-Verlag, 1991,

[18] D. E. Knuth. *The art of Computer Programming*, Vol. III: Sorting and Searching. Addison-Wesley, 1973.

[19] M. Minsky and S. Papert. Unrecognizable sets of numbers. *J. Assoc. Comput. Mach.* **13** (1966), 281–286.

[20] C. Pomerance, J. M. Robson, and J. O. Shallit. Automaticity II: Descriptional complexity in the unary case. To appear, *Theoret. Comput. Sci.*

[21] J. B. Rosser and L. Schoenfeld. approximate formulas for some functions of prime numbers. *Ill. J. Math.* **6** (1962), 64–94.

[22] L. Schoenfeld. Sharper bounds for the Chebyshev functions $\theta(x)$ and $\psi(x)$. II. *Math. Comp.* **30** (1976), 337–360. Corrigenda in *Math. Comp.* **30** (1976), 900.

[23] J. Shallit. Some facts about continued fractions that should be better known. Technical report CS-91-30, University of Waterloo, Department of Computer science, July 1991.

[24] J. Shallit. The fixed pointof $1 \rightarrow 121$, $2 \rightarrow 12221$ is not 2-automatic. Unpublished manuscript, dated September 10, 1992.

[25] J. Shallit. Real numbers with bounded partial quotients: a survey. *Enseign. Math.* **38** (1992), 151–187.

[26] J. Shallit and Y. Breitbart. Automaticity: Properties of a measure of descriptional complexity. In P. Enjalbert, E. W. Mayr, and K. W. Wagner, editors, *STACS 94: 11th Annual Symposium on Theoretical aspects of Computer Science,* Vol. 775 of Lecture Notes in Computer Science, pages 619–630. Springer-Verlag, 1994.

[27] J. Shallit and Y. Breitbart. AutomaticityI: Properties of a measure of descriptional complexity. To appear, *J. Comput. System Sci.*

[28] N. B. Slater. Gaps and Steps for the sequences $n\theta$ mod 1. *Proc. Cambridge Phil. Soc.* **63** (1967), 115–1123.

[29] V. T. Sós. On the distribution mod 1 of the sequence $n\alpha$. *Ann. Univ. Sci. Budapest Eötvös Sect. Math.* **1** (1958), 127–134.

[30] S. Świerczkowski. On successive settings of an arc on the circumference of a circle. *Fundamenta Math.* **46**, 187–189.
[31] S. S. Wagstaff, Jr. Greatest of the least primes in arithmetic progressions having a given modulus. *math. Comp.* **33** (1979), 1073–1080.

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