A new family of nonstochastic languages

Rūsiņš Freivalds a, 1, Abuzer Yakaryılmaz b, 2, A.C. Cem Say b, *, 2

a Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulvāris 29, Rīga, LV-1459, Latvia
b Boğaziçi University, Department of Computer Engineering, Bebek 34342 İstanbul, Turkey

Article info
Article history:
Received 12 May 2009
Accepted 9 March 2010
Available online 16 March 2010
Communicated by P.M.B. Vitanyi

Keywords:
Theory of computation
Nonstochastic languages

1. Introduction

The class of stochastic languages, i.e. those recognized by probabilistic finite automata (PFA’s) with nonisolated cutpoint, has the cardinality of the real numbers, so the question of whether there exists any nonstochastic language at all is a nontrivial matter. This problem was studied by Bukharaev and many others [1–4], who identified several nonstochastic languages. The compilation of a rich list of nonstochastic languages is important for understanding the limitations of not just the PFA, but also other significant models of computation, like constant space probabilistic Turing machines, and one-way quantum finite automata (QFA’s), which are known [5–7] to be able to recognize only stochastic languages. In this paper, we prove a new subclass of context-free languages, which are useful in the study of the computational power of QFA’s, to be nonstochastic.

2. Preliminaries

The probability that a probabilistic finite automaton (PFA) [1] \( P \), with state set \( A = \{a_1, \ldots, a_n\} \), will be in state \( a_i \) after reading the input prefix \( u \) is denoted by \( p_i(u) \), and is determined by the program of \( P \). The acceptance probability of \( u \) is

\[
P_{\text{acc}}(u) = \sum_{i \in F} p_i(u),
\]

where \( F \subseteq A \) is the set of accepting states. The language \( L \) recognized by \( P \) with cutpoint \( \lambda \in [0, 1) \) is defined as \( L = \{u \mid P_{\text{acc}}(u) > \lambda\} \). Any language recognized by a PFA with cutpoint is a stochastic language. If a language is not stochastic, then it is nonstochastic.

The following notation and definitions will be used in this paper:

- \( w[i] \): the \( i \)th symbol (from the left) of string \( w \),
- \( |w|_\sigma \): the number of occurrences of symbol \( \sigma \) in string \( w \),
- \( S_k \): the set of all \( k \)-bit binary strings,
- \( J_k(s) = \{i \mid s[i] = 1\} \), where \( k \in \mathbb{Z}^+ \), \( s \in S_k \), \( 1 \leq i \leq k \).

3. New nonstochastic languages

Consider the following two languages on the alphabet \( \Sigma = \{a, b\} \):

\[
L_1 = \{w \mid \exists s, t, u, v \in \Sigma^*, \ w = sat = ubv, \ |s| = |v|\},
\]

\[
L_2 = \{w \mid \exists s, t, u, v \in \Sigma^*, \ w = sbt = ubv, \ |s| = |v|\}.
\]
Let $d_1$, which is the set of nonpalindromes on this alphabet, is stochastic [7,8], whereas $L_2$ is not, as we will show, for the first time, in this paper.

Note that the membership conditions of both $L_1$ and $L_2$ are defined in terms of the existence of a pair of symbols, whose positions, expressed as offsets from the left and right ends of the string, say, $x$ and $y$, respectively, satisfy the simple equation $x = y$. In the following, we will examine a family of languages where the relationship between $x$ and $y$ is generalized to an arbitrary linear Diophantine equation, and we will demonstrate exactly which members of this family are nonstochastic.

**Definition 1.** Let $d_1, d_2, d \in \mathbb{Z}$, and $\sigma_1, \sigma_2 \in \Sigma$, where $\Sigma$ is a finite alphabet. Language $L_{d_1, \sigma_1, d_2, \sigma_2, d, \Sigma}$ is defined as follows:

$$L_{d_1, \sigma_1, d_2, \sigma_2, d, \Sigma} = \{ w \in \Sigma^* \mid \exists x, y \{ w[x] = \sigma_1, w[y] = \sigma_2, d_1(x+y) = d \} \}.$$

It is easy to show that all languages of this form are context-free. Well-known properties of linear Diophantine equations can be used to demonstrate that $L_{d_1, \sigma_1, d_2, \sigma_2, d, \Sigma}$ is regular (and therefore stochastic) whenever (i) the alphabet $\Sigma$ is unary, (ii) $d_1$ and $d_2$ have different signs, and (iii) $\gcd(d_1, d_2)$ does not divide $d$. Note that, if $\gcd(d_1, d_2) = g > 1$, and $g$ divides $d$, then $L_{d_1, \sigma_1, d_2, \sigma_2, d, \Sigma} = L_{\frac{d}{g}, \sigma_1, \frac{d}{g}, \sigma_2, \frac{d}{g}, \Sigma}$.

We will therefore focus on the case where $\Sigma$ contains at least two symbols, $d_1, d_2 \in \mathbb{Z}^+$, and $\gcd(d_1, d_2) = 1$ in the rest of this paper.

**Theorem 1.** For $d \in \mathbb{Z}$, language $L_{1,a,1,b,d,[a,b]}$ is stochastic.

**Proof.** If $d \geq 0$, then $L_{1,a,1,b,d,[a,b]} = \{a b^d \} L_1$, where $L_1$ is the language of nonpalindromes, defined in Eq. (1). Since $L_1$ is stochastic, and the PFA for $[a,b]^d$ has a simple structure that makes it easy to "append" the PFA for $L_1$ to it, $L_{1,a,1,b,d,[a,b]}$ is also recognized by a PFA. If $d < 0$, then $L_{1,a,1,b,d,[a,b]} = L_1 \{a b^d \}$. Since stochastic languages are closed under reversal [1], $L_{1,a,1,b,d,[a,b]}$ is stochastic in this case as well.

We will show that all remaining languages of the form $L_{d_1, \sigma_1, d_2, \sigma_2, d, \Sigma}$, including $L_2$ (Eq. (2)), are nonstochastic. Our proofs in this regard will use the following fact:

**Theorem 2.** (See [3,4].) Suppose that the language $L$ has the following property: For any $k \in \mathbb{Z}^+$, there exist two sets $U = \{u_s \mid s \in S_k\}$ and $V = \{v_1, v_2, \ldots, v_k\}$, such that $u_s v_i \in L$ if and only if $s[i] = 1$,

where $1 \leq i \leq k$, $u_s \in U$, $v_i \in V$. Then $L$ is nonstochastic.

In what follows, $(x_1, y_1) \in (\mathbb{Z}^+, \mathbb{Z}^+)$ is the pair that has the smallest value of $(x_1 + y_1)$ satisfying $d_1 x_1 = d_2 y_1 + d$.

Additionally, we define three lengths, parametric on the natural number $k$:

$$l_u = x_1 + (k-1)d_2, \quad l_{v'} = y_1 + (k-1)d_1,$$

and

$$l_{v'} = \begin{cases} \frac{d_1 l_u + l_{v'} + 1 - d_2 l_u}{d_2 - d_1}, & d_1 < d_2, \\ 0, & d_1 = d_2, \\ \frac{d_2 l_u + l_{v'} + 1 - d_1 l_u + d}{d_1 - d_2}, & d_1 > d_2. \end{cases}$$

Here, $l_{v'}$ is assigned the minimum values that satisfy the following conditions:

1. If $y = l_{v'} + l_{v'}$ then $x = \frac{d_1 y + d}{d_1} \geq l_u + l_{v'} + l_{v'} + 1$ (when $d_1 < d_2$).
2. If $x = l_u + l_{v'}$ then $y = \frac{d_1 y - d}{d_2} \geq l_u + l_{v'} + l_{v'} + 1$ (when $d_1 > d_2$).

Let us demonstrate our basic approach by showing that $L_2$, which is simply $L_{1,b,1,b,0,[a,b]}$, is nonstochastic. (Note that $x_1 = y_1 = 1$, $l_u = l_{v'} = k$, and $l_{v'} = 0$ in this case.) Let $U = \{a, b\}^k$. Each member $u_s$ of $U$ is associated with the string $s \in S_k$ such that $u_s[1] = b$ if and only if $s[1] = 1$, where $1 \leq i \leq k$. Let $V$ be the set of all strings $v_i$ of length $k$ which contain the symbol $b$ at the ith position from the right, and the symbol $a$ in all other positions, where $1 \leq i \leq k$. It is easy to see that $u_s v_i \in L_2$ if and only if $s[i] = 1$, for all $u_s$ and $v_i$, and we therefore conclude by Theorem 2 that $L_2$ is nonstochastic. We now proceed to prove the generalized version of this result.

**Theorem 3.** For $d \in \mathbb{Z}$ and $d_1, d_2 \in \mathbb{Z}^+$, language $L_{d_1,b,d_2,b,d,[a,b]}$, where $[a, b] \subseteq \Sigma$, is nonstochastic.

**Proof.** We will show that the language under consideration has the property mentioned in Theorem 2. For any $k \in \mathbb{Z}^+$, let $U$ and $V$ be defined as follows:

$$U = \{u_s \in \{a, b\}^* \mid |u_s| = l_u, |u_s|[a] = |s|_1, \forall i \in J_k(s) (u_s[x_1 + (i-1)d_2] = b) \},$$

where $s \in S_k$;

$$V = \{v_i \in \{v' a b^* \mid |v_i| = l_{v'}, v'_i[1] - (y_1 + (i-1)d_1) = b \},$$

where $1 \leq i \leq k$ and $v'_i = a^{l_{v'}}$.

Let us show that $u_s v_i \in L_{d_1,b,d_2,b,d,[a,b]}$ if and only if $s[i] = 1$.

If $s[i] = 1$, there are $b$’s at positions $x = x_1 + (i-1)d_2$ and $y = y_1 + (i-1)d_1$. It is easily followed that

$$d_1 x = d_1 x_1 + (i-1)d_2 = d_2 y_1 + (i-1)d_1 + d = d_2 y + d,$$

and so $u_s v_i \in L_{d_1,b,d_2,b,d,[a,b]}$.

If $s[i] \neq 1$, and further that $u_s v_i \in L_{d_1,b,d_2,b,d,[a,b]}$. We look for an $(x, y)$ pair that satisfies
For any $k \in \mathbb{Z}^+$, let $U$ and $V$ be defined as follows:

$$U = \{ u_5 \in [a, b]^* \mid |u_5| = 8, |u_5|_a = |s|_1, \forall i \in J_k(s)(u_5[x_1 + (i - 1)d_1] = a) \},$$

where $s \in S_k$;

$$V = \{ v_i \in v'ab^*a^* \mid |v_i| = l_w + l_w', v_i[|v_i| - (y_i + (i - 1)d_1) + 1] = b \},$$

where $1 \leq i \leq k$, and $v' = a^{6k-2b}$ if $d_1 < d_2$, and $v' = b^{l_{2b}}$ if $d_1 > d_2$.

The proof of Theorem 3 can be easily modified to show that $u_5v_i \in L_{d_1, a, d_2, b, d, x}$ if and only if $s[i] = 1$. The adaptation of the $i$ part, and the case where $d_1 < d_2$ in the only if part, is trivial. The only possibility not covered in the previous proof is that of $x$ indexing a symbol (except the last one) in $v'$ when $d_1 > d_2$, and this is ruled out by setting all symbols of $v'$ to $b$.

Theorem 5.

For $d \in \mathbb{Z}$ and $d_1, d_2 \in \mathbb{Z}^+$ with $d_1 = d_2 = 1$, language $L_{d_1, a, d_2, b, d, x}$, where $[a, b, c] \subseteq \Sigma$, is nonstochastic.

The method presented above can also be used for languages defined in terms of relationships among the positions of more than two symbols. We will just give a specific example in the next result.

Theorem 6. Language $L_3 = \{ w \mid \exists x, y, z(w|x) = w[y] = w[z] = b, 3x + y = z \}$, on the alphabet $\Sigma = \{a, b, c\}$, is nonstochastic.

Proof. For any $k \in \mathbb{Z}^+$, let $U$ and $V$ be defined as follows:

$$U = \{ u_5 \in [a, b]^* \mid |u_5| = 2k - 1, |u_5|_b = |s|_1, \forall i \in J_k(s)(u_5[2i - 1] = b) \},$$

where $s \in S_k$;

$$V = \{ v_i \in v'ab^*a^* \mid |v_i| = |v_i| = 2k - 4, v_i[6k + 6i - 4] = b \},$$

where $1 \leq i \leq k$ and $v' = a^{6k-2b}$.

We will now show that, for any $w = u_5v_i \in L_3$, the values of $x$, $y$ and $z$ must be as follows, in order to satisfy $3x + y = z$.

- $1 \leq x \leq 2k - 1$.
- $y = 8k - 2$.
- $z \leq 14k - 5$.

$x$ cannot index a $b$ in $v_i$, because $x$ would be at least $8k - 2$ in that case, causing $3x$ to be greater than $|u_5v_i|$, precluding the existence of any $y, z$ to satisfy $3x = z - y$. Since $x$ indexes a $b$ in $u_5$, it must be odd, and thus $y \neq z$ mod (2), with $y < z$. If $y$ is odd, and so $z$ is even, then $y$ must be an index of $u_5$, and $z$ must equal $8k - 2$, the only even index. In this case, the maximum possible value of $3x + y$ becomes $8k - 4$, less than $8k - 2$. Hence, $y$ must be even, surely $8k - 2$, and so $z$ is odd, and $8k + 1 \leq z \leq 14k - 5$. With these constraints in place, it is straightforward to use the approach of the preceding proofs in this case.

4. Concluding remarks

Languages of the form $L_{d_1, a_1, d_2, a_2, d, x}$, defined in Section 3, are useful in the study of QFA’s, since they can be shown to be unrecognizable by one-way, but recognizable with “1.5-way” QFA’s [7].

Our results in Section 3 can evidently be extended further to cover more languages, whose membership conditions involve more general Diophantine equations, and it would be interesting to examine the recognizability of those languages by various kinds of QFA.

Acknowledgement

We thank Ahmet Feyzioğlu for his helpful pointers to useful literature.

References


