ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

439

ON LANGUAGES REPRESENTABLE IN RATIONAL PROBABILISTIC AUTOMATA

BY

PAAVO TURAKAINEN

HELSINKI 1969
SUOMALAINEN TIEDEAKATEMIA

doi:10.5186/aasfm.1969.439

Communicated 13 December 1968 by P. J. Myrberg and K. Inkeri

KESKUSKIRJAPAINO
HELSINKI 1969

On languages representable in rational probabilistic automata

In a probabilistic automaton the initial vector as well as the transition matrices are stochastic and the final vector consists of 0's and 1's only. If the elements of these vectors and matrices are replaced by arbitrary real numbers, then we get the so-called generalized automaton. However, this generalization is not essential as far as the family of representable languages is concerned; we have proved in [2] that a language can be represented in a generalized automaton if and only if it can be represented in a probabilistic automaton. This result is a useful tool in the investigation of languages representable in probabilistic automata.

In this paper, we first (sections 1 and 2) present two different notions of the representability of languages in probabilistic automata and prove, by using generalized automata, that they lead to the same family of languages, which we call the family of stochastic languages.

In section 3, we investigate probabilistic automata where the elements of the initial vector and of the transition matrices are rational numbers. The family $\mathcal{L}_{rat}$ of languages representable in these automata with rational cut-points contains all regular languages as a proper subfamily. It turns out that $\mathcal{L}_{rat}$ coincides with the family $\mathcal{L}_{int}$ by which we mean the family of languages representable with integer-valued cut-points in generalized automata where the elements of the initial vector, of the final vector and of the matrices are integers. By using generalized automata, we prove that the family $\mathcal{L}_{rat}$ is closed under complementation. As it is well-known, the corresponding problem for the whole family of stochastic languages is open. In section 3, we also introduce another subfamily of stochastic languages containing all regular languages as a proper subfamily.

Finally, in section 4 generalized automata are used in establishing that the language $\{x^n y^n | n \geqq 1\}$ is a stochastic language.

1. By an *alphabet* I we mean a finite non-empty set. The set of words, including the empty word Λ, over the alphabet I is denoted by $W(I)$. Subsets of $W(I)$ are called *languages* over I. The union, the intersection and the product of two languages L_1 and L_2 are denoted, respectively, by $L_1 + L_2$, $L_1 \cap L_2$ and $L_1 L_2$. The complement of a language L with respect to $W(I)$ is denoted by $\bar{L}$. We also use the notations

$$L_1 - L_2 = L_1 \cap \bar{L}_2\,, \quad L^* = \sum_{i=0}^{\infty} L^i$$

where $L^0 = \{\Lambda\}$.

Definition. *A generalized automaton* over the alphabet I is an ordered quadruple $\mathfrak{GA} = (S, M, \pi_0, f_0)$ where $S = \{s_1, \ldots, s_n\}$ is a finite non-empty set (the set of states), M is a mapping of I into the set of $n \times n$ matrices with real elements, π_0 is an n-dimensional row vector with real components (the *initial vector*) and f_0 is an n-dimensional column vector with real components (the *final vector*).

The domain of M is extended from I to $W(I)$ by defining

$$M(\Lambda) = E_n \quad (n \times n \text{ identity matrix}),$$
$$M(x_1 x_2 \cdots x_k) = M(x_1)\, M(x_2) \cdots M(x_k)$$

where $k > 1$ and $x_i \in I$.

If the initial vector π_0 as well as the matrices $M(x)$ $(x \in I)$ are stochastic and the final vector f_0 consists of 0's and 1's only, then $\mathfrak{GA}$ is called a *probabilistic automaton*. In this case we also use the notation $\mathfrak{PA}$.

By a *rational probabilistic automaton* $\mathfrak{RPA}$ we mean a probabilistic automaton $\mathfrak{PA} = (S, M, \pi_0, f_0)$ where the elements of π_0 and of the matrices $M(x)$ $(x \in I)$ are rational numbers.

An *integer-valued generalized automaton* $\mathfrak{IGA} = (S, M, \pi_0, f_0)$ is a generalized automaton where the elements of π_0, f_0 and of the matrices $M(x)$ $(x \in I)$ are integers.

For any real number η, the language represented in $\mathfrak{GA}$ with the cut-point η is defined to be the set

$$L(\mathfrak{GA}, \eta) = \{P \in W(I) \mid \pi_0 M(P) f_0 > \eta\}\,.$$

If $\mathfrak{GA}$ is a probabilistic automaton, then $L(\mathfrak{GA}, \eta)$ is called a *stochastic language*.

For any probabilistic automaton $\mathfrak{PA} = (S, M, \pi_0, f_0)$ and for any real numbers η and ε $(\varepsilon \geqq 0)$, we define

$$L(\mathfrak{PA}, \eta, \varepsilon) = \{P \in W(I) \mid |\pi_0 M(P) f_0 - \eta| < \varepsilon\}\,. \tag{1}$$

Let $\mathcal{L}$ be the family consisting of the languages L over I such that, for some $\mathfrak{PA}$, η and ε, $L = L(\mathfrak{PA}, \eta, \varepsilon)$. In what follows, we also use the notation $\mathcal{L}_{rat}$ $(\mathcal{L}_{int})$ to mean the family of languages L over I such that, for some $\mathfrak{RPA}$ $(\mathfrak{IGA})$ and some rational number η (integer η), $L = L(\mathfrak{RPA}, \eta)$ $(L = L(\mathfrak{IGA}, \eta))$.

It should be noted that the family of regular languages over I is a

proper subfamily of $\mathcal{L}_{rat}$ (cf., for instance, the example of [2], p. 19). On the other hand, $\mathcal{L}_{rat}$ is denumerable and, consequently, a proper subfamily of stochastic languages.

Lemma 1. *A language L can be represented in a generalized automaton if and only if it can be represented in a probabilistic automaton, i.e., if and only if it is a stochastic language.*

This lemma has been proved in [2] by using a constructive method. The following lemma has been proved in [2] for the so-called generalized probabilistic automata, but the same proof is valid for generalized automata.

Lemma 2. *For two generalized automata $\mathfrak{GA}_1 = (S_1, M_1, \pi_1, f_1)$ and $\mathfrak{GA}_2 = (S_2, M_2, \pi_2, f_2)$ over I, there exist generalized automata $\mathfrak{GA} = (S, M, \pi_0, f_0)$ and $\mathfrak{GA}' = (S', M', \pi_0', f_0')$ such that, for any word $P \in W(I)$,*

$$\pi_0 M(P) f_0 = \pi_1 M_1(P) f_1 + \pi_2 M_2(P) f_2$$

and

$$\pi_0' M'(P) f_0' = (\pi_1 M_1(P) f_1)\,(\pi_2 M_2(P) f_2)\,.$$

2. In this section, we show that the family $\mathcal{L}$ obtained from the definition (1) coincides with the family of stochastic languages.

Theorem 1. *A language L belongs to the family $\mathcal{L}$ if and only if it is a stochastic language.*

Proof. For the »if»-part, assume that L is a stochastic language over I, i.e., $L = L(\mathfrak{PA}, \eta)$ $(\eta \leqq 1)$ for a probabilistic automaton $\mathfrak{PA} = (S, M, \pi_0, f_0)$. Thus,

$$L = \{P \mid \pi_0 M(P) f_0 > \eta\}\,,$$

which can be expressed in the form

$$L = \{P \mid |\pi_0 M(P) f_0 - 1| < 1 - \eta\}\,.$$

This implies that $L \in \mathcal{L}$.

To establish the »only if»-part, let $L \in \mathcal{L}$ be arbitrary. By the definition of $\mathcal{L}$, there exist a probabilistic automaton $\mathfrak{PA} = (S_1, M_1, \pi_1, f_1)$ and real numbers η, ε $(\varepsilon \geqq 0)$ such that

$$L = \{P \mid |\pi_1 M_1(P) f_1 - \eta| < \varepsilon\}\,.$$

Since $\varepsilon \geqq 0$, this can be written in the form

$$L = \{P \mid (\pi_1 M_1(P) f_1 - \eta)^2 < \varepsilon^2\} . \tag{2}$$

There exists a generalized automaton $\mathfrak{GA}' = (S_2, M_2, \pi_2, f_2)$ such that, for any word $P \in W(I)$, $\pi_2 M_2(P) f_2 = -\eta$. This implies, by Lemma 2, that for some generalized automaton $\mathfrak{GA}'' = (S, M, \pi_0, f_0)$ the equation

$$\pi_0 M(P) f_0 = (\pi_1 M_1(P) f_1 - \eta)^2 \tag{3}$$

holds for any word $P \in W(I)$. From formulas (2) and (3) it now follows that

$$L = \{P \mid \pi_0 M(P) (-f_0) > -\varepsilon^2\} .$$

In other words, $L = L(\mathfrak{GA}, -\varepsilon^2)$ for $\mathfrak{GA} = (S, M, \pi_0, -f_0)$. By Lemma 1, this implies that L is a stochastic language. Theorem 1 is thus proved.

Remark. For any generalized automaton $\mathfrak{GA}$, the language $L(\mathfrak{GA}, \eta, \varepsilon)$ can be defined in the same way as $L(\mathfrak{PA}, \eta, \varepsilon)$ in (1). Also in this case the corresponding family of languages coincides with the family of stochastic languages. The proof is the same as that of Theorem 1.

3. This section deals with rational probabilistic automata. We need the following two lemmas.

Lemma 3. *The family* $\mathscr{L}_{int}$ *is a subfamily of* $\mathscr{L}_{rat}$.

The validity of this lemma is verified by considering the constructive proof of Lemma 1 in [2].

Lemma 4. *If* $L \in \mathscr{L}_{rat}$, *then there exists an integer-valued generalized automaton* $\mathfrak{TGA}$ *such that* $L = L(\mathfrak{TGA}, 0)$.

Proof. Let $L \in \mathscr{L}_{rat}$ be arbitrary. By definition, $L = L(\mathfrak{RPA}, \eta)$ for some $\mathfrak{RPA} = (S_1, M_1, \pi_1, f_1)$ and some rational number η. Denote

$$f_1' = f_1 - \begin{bmatrix} \eta \\ \cdot \\ \cdot \\ \cdot \\ \eta \end{bmatrix} .$$

Note that the components of f_1' are rational numbers. Since $\pi_0 M(P)$ is a stochastic vector for every word $P \in W(I)$, we have

$$\pi_1 M_1(P) f_1' = \pi_1 M_1(P) f_1 - \eta \text{ for all } P \in W(I) .$$

This implies that, for $\mathfrak{GA} = (S_1, M_1, \pi_1, f_1')$, $L = L(\mathfrak{GA}, 0)$. Since the elements of π_1, f_1' and of the matrices $M_1(x)$ $(x \in I)$ are rational, there exists a natural number K such that the elements of

$$\pi_0 = K\pi_1\,, \quad f_0 = Kf_1'\,, \quad M(x) = KM_1(x) \quad (x \in I)$$

are integers. Clearly, for any word $P \in W(I)$, $\pi_1 M_1(P) f_1' > 0$ if and only if $\pi_0 M(P) f_0 > 0$. This implies that, for $\mathfrak{TGA} = (S_1, M, \pi_0, f_0)$, $L = L(\mathfrak{TGA}, 0)$, whence the lemma follows.

As an immediate consequence of Lemmas 3 and 4, we obtain the following

Theorem 2. *The families* $\mathcal{L}_{rat}$ *and* $\mathcal{L}_{int}$ *are equal.*

It is not known whether or not the family of stochastic languages is closed under complementation. In the following theorem we solve this problem for the subfamily $\mathcal{L}_{rat}$ of stochastic languages.

Theorem 3. *The family* $\mathcal{L}_{rat}$ *is closed under complementation. Thus, if* $L \in \mathcal{L}_{rat}$, *then* $\bar{L}$ *is a stochastic language.*

Proof. Let $L \in \mathcal{L}_{rat}$ be arbitrary. By Lemma 4, $L = L(\mathfrak{TGA}, 0)$ for an integer-valued generalized automaton $\mathfrak{TGA} = (S, M, \pi_0, f_0)$. It follows that

$$\pi_0 M(P) f_0 \text{ is an integer for all } P \in W(I)\,. \tag{4}$$

Since

$$\bar{L} = \{P \mid \pi_0 M(P) f_0 \leqq 0\}\,,$$

we infer from (4) that

$$\begin{aligned} \bar{L} &= \{P \mid \pi_0 M(P) f_0 < 1\} \\ &= \{P \mid \pi_0 M(P)\,(-f_0) > -1\}\,. \end{aligned}$$

Hence $\bar{L} = L(\mathfrak{TGA}', -1)$ for the integer-valued generalized automaton $\mathfrak{TGA}' = (S, M, \pi_0, -f_0)$. Theorem 3 now follows from Theorem 2.

For any probabilistic automaton $\mathfrak{PA} = (S, M, \pi_0, f_0)$ over I, we denote

$$L(\mathfrak{PA}\,, \eta\,, =) = \{P \in W(I) \mid \pi_0 M(P) f_0 = \eta\}\,.$$

Let $\mathcal{L}(=)$ be the family consisting of the languages L over I such that, for some $\mathfrak{PA}$ and some real number η, $L = L(\mathfrak{PA}, \eta, =)$. Let $\mathcal{L}_{rat}(=)$ be the family of languages L over I such that, for some $\mathfrak{RPA}$ and

some rational number η, $L = L(\mathfrak{RPA}, \eta, =)$. (Thus, $\mathcal{L}_{rat}(=)$ is a subfamily of $\mathcal{L}(=)$.)

Theorem 4. *The family of regular languages over I is a proper subfamily of $\mathcal{L}_{rat}(=)$. The family $\mathcal{L}_{rat}(=)$ is a proper subfamily of stochastic languages.*

Proof. Every deterministic automaton can be rewritten as a probabilistic automaton where the initial vector and the rows of the transition matrices are co-ordinate vectors. Thus, for any regular language L over I, there exists a deterministic automaton $\mathfrak{PA} = (S, M, \pi_0, f_0)$ such that $L = \{P \mid \pi_0 M(P) f_0 = 1\}$. This implies that $L \in \mathcal{L}_{rat}(=)$. The first part of Theorem 3 now follows from the fact that, for example, the non-regular language $\{x^n yx^n y \mid n \geqq 1\} + xyy$ belongs to $\mathcal{L}_{rat}(=)$ (cf. [1]).

To prove the last sentence of the theorem, let $L \in \mathcal{L}_{rat}(=)$ be arbitrary. From the proof of Lemma 4 we conclude that

$$L = \{P \mid \pi_0 M(P) f_0 = 0\}$$

for an integer-valued generalized automaton $\mathfrak{GPA} = (S, M, \pi_0, f_0)$. It follows that $\pi_0 M(P) f_0$ is an integer for all $P \in W(I)$. Consequently,

$$L = \{P \mid (\pi_0 M(P) f_0)^2 < 1\} .$$

As in the proof of the »only if»-part of Theorem 1, it is verified that L is a stochastic language. The proof is now complete, because $\mathcal{L}_{rat}(=)$ is denumerable.

The problem whether or not $\mathcal{L}(=)$ is a subfamily of stochastic languages is open. We have established in [2] that if there exists a language $L \in \mathcal{L}(=)$ which is not stochastic, then the family of stochastic languages is not closed under complementation.

4. Finally, we show that the language $\{x^n y^n \mid n \geqq 1\}$ is a stochastic language. The following lemma is needed.

Lemma 5. *There exists a generalized automaton $\mathfrak{GA} = (S, M, \pi_0, f_0)$ over $\{x, y\}$ such that the elements of π_0, f_0 and of the matrices $M(x)$, $M(y)$ are rational and*

$$\{P \mid \pi_0 M(P) f_0 = 0\} = \{x^n y^n \mid n \geqq 1\} + ((x + y)^* - xx^* yy^*) .$$

Proof. Denote

$$L_1 = \{x^n y^n \mid n \geqq 1\} , \quad L_2 = (x + y)^* - xx^* yy^* , \quad L = L_1 + L_2 .$$

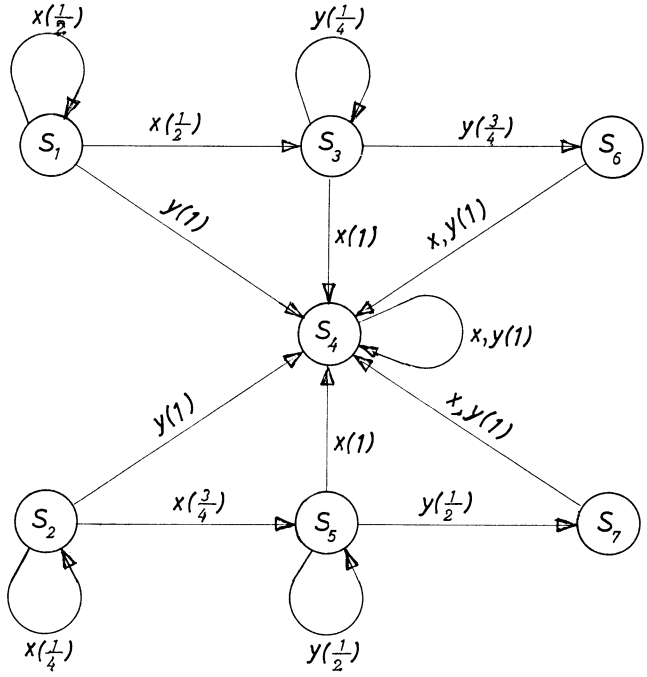

Fig. 1.

Consider the generalized automaton $\mathfrak{GA} = (S, M, \pi_0, f_0)$ over $\{x, y\}$ where $S = \{s_1, \ldots, s_7\}$,

$$\pi_0 = (\tfrac{1}{2}, \tfrac{1}{2}, 0, \ldots, 0), \quad f_0^T = (0, \ldots, 0, 2, -2)$$

(f_0^T means the transpose of f_0) and the mapping M is defined by the graph of Figure 1.

It is verified that, for any nonnegative integers n and k,

$$\pi_0 M(P) f_0 = \begin{cases} 3(\tfrac{1}{2})^{n+k+3}((\tfrac{1}{2})^k - (\tfrac{1}{2})^n) & \text{is } P \text{ if of the form } x^{n+1}y^{k+1}, \\ 0 & \text{otherwise.} \end{cases}$$

This implies our lemma.

Theorem 5. *The language* $\{x^n y^n \mid n \geqq 1\}$ *is a stochastic language.*

Proof. Let L_1, L_2 and L be as in the proof of Lemma 5. From this lemma and the proof of Lemma 4 it follows that

$$L = \{P \mid \pi_0 M(P) f_0 = 0\}$$

for an integer-valued generalized automaton $\mathfrak{IGA} = (S, M, \pi_0, f_0)$. We consider the language L_1 in the form

$$L_1 = L \cap \bar{L}_2 . \tag{5}$$

Since $\bar{L}_2$ is a regular language, there exists a deterministic automaton $\mathfrak{PA} = (S_1, M_1, \pi_1, f_1)$ such that

$$\bar{L}_2 = \{P \mid \pi_1 M_1(P) f_1 = 1\} . \tag{6}$$

Denote $f_1' = f_1 - (1, \ldots, 1)^T$. Formula (6) now implies that, for the integer-valued generalized automaton $\mathfrak{IGA}_1 = (S_1, M_1, \pi_1, f_1')$,

$$\bar{L}_2 = \{P \mid \pi_1 M_1(P) f_1' = 0\} .$$

By Lemma 2, there exists a generalized automaton $\mathfrak{GA}_2 = (S_2, M_2, \pi_2, f_2)$ over $\{x, y\}$ such that, for all $P \in (x + y)^*$,

$$\pi_2 M_2(P) f_2 = (\pi_0 M(P) f_0)^2 + (\pi_1 M_1(P) f_1')^2 .$$

Consequently, by formula (5),

$$L_1 = \{P \mid \pi_2 M_2(P) f_2 = 0\} .$$

For all $P \in (x+y)^*$, the numbers $\pi_0 M(P) f_0$ and $\pi_1 M_1(P) f_1'$ are integers. Thus also $\pi_2 M_2(P) f_2$ is an integer. This implies that

$$L_1 = \{P \mid |\pi_2 M_2(P) f_2| < 1\} .$$

As in the proof of the »only if»-part of Theorem 1, it is now verified that L_1 is a stochastic language.

University of Turku
Turku, Finland

References

[1] Starke, P. H.: Stochastische Ereignisse und Wortmengen. - Z. Math. Logik Grundlagen Math. 12 (1966), 61–68.

[2] Turakainen, P.: On probabilistic automata and their generalizations. - Ann. Acad. Sci. Fenn. A I 429 (1968).

Printed April 1969