LINEAR ALGEBRA AND ITS APPLICATIONS 5, 277–281 (1972)

Context-Free Grammars and Nonnegative Matrices

DAVID SANKOFF
Université de Montréal
Montréal, Québec, Canada

Communicated by A. J. Hoffman

The probabilistic interpretation of certain nonnegative matrix properties can be applied to the study of context-free grammars. The first theorem provides the matrix property central to our discussion.

THEOREM 1. *Let* $P = \begin{pmatrix} I & 0 \\ R & A \end{pmatrix}$ *be an* $(m + n) \times (m + n)$ *nonnegative matrix such that the largest eigenvalue of* A *is* $\lambda < 1$. *Then as* $N \to \infty$, $P^N \to Q$ *where* $Q = \begin{pmatrix} I & 0 \\ L & 0 \end{pmatrix}$ *and* $R = (I - A)L$.

Proof. See [6].

To apply this fact we first define a context-free grammar and describe how to attach probabilities to the rules of such a grammar, in constructing a behavioral model of sentence production.

A context-free grammar consists of two sets of integers, $T = \{1, \ldots, m\}$ and $C = \{m + 1, \ldots, m + n\}$, the *terminal* integers and the *nonterminal* integers respectively, and a set $G = \{r_1, \ldots, r_h\}$ of *rewrite rules*. Each rule is of the form $i \to \alpha_r$, where $i \in C$ and α_r is a finite sequence of elements from $T \cup C$. Each $i \in C$ must be on the left of the arrow for at least one rule in G.

The sentences generated by the grammar (T, C, G) are all those and only those sequences produced as follows. Start with the sequence consisting of the single term $m + 1$. Choose from among the rules in G *any* one which rewrites $m + 1$, i.e., of form $m + 1 \to \alpha$. Rewrite (replace) $m + 1$ with the sequence α. If there are any terms which are elements

of C in this new sequence, these must also be rewritten. When a term is rewritten it is simply deleted from the sequence and replaced by the one or more elements in α_r if r is the rewriting rule selected, without disturbing the neighboring terms of the sequence on the left and right. The process stops when the last sequence contains only elements of T. This sequence is a sentence, of which there are at most, countably many.

Example. $T = \{1, 2\}$, $C = \{3, 4\}$, $G = \{r_1, r_2, r_3, r_4\} = \{3 \to 1;\ 3 \to 4, 3;\ 4 \to 2, 1;\ 4 \to 3, 4\}$. By starting with the sequence consisting of $3 (= m + 1)$ and applying rules in the order $r_2; r_1; r_4; r_3; r_1$, one arrives successively at the sequences 4, 3; 4, 1; 3, 4, 1; 3, 2, 1, 1; and finally at the sentence 1, 2, 1, 1.

In constructing a behavioral or stochastic model of sentence production, we must take into account that it is necessary at each step of a production to make a choice from the subset of rules $G(i) \subset G$ which rewrite a particular $i \in C$. This is most simply accomplished by setting up, in advance, a probability distribution on each $G(i)$. (See [2–5].) The probability distributions may be unambiguously denoted by a single function $\pi(\cdot)$, since the $G(i)$ are disjoint, and

$$\sum_{r \in G(i)} \pi(r) = 1, \quad \text{for all} \quad i \in C.$$

The problem we investigate here is as follows. *If we know the rules of the grammar but not the function* $\pi(\cdot)$, *and we can observe the output of the grammar, e.g., the average number of terms of various types per sentence, can we use this information to find* $\pi(\cdot)$?

$\pi(\cdot)$ determines an $(m + n) \times (m + n)$ nonnegative matrix P, where P_{ij} may be interpreted as the expected number of j terms to be expected when i is rewritten ($i \in C$, $j \in T \cup C$). In other words

$$P_{ij} = \sum_{r \in G(i)} \pi(r)[\text{number of } j \text{ terms in } \alpha_r], \qquad \text{for} \quad i \in C.$$

For $i \in T$, we make the convention $P_{ij} = \delta_{ij}$.

Let e_{m+1} be the unit vector in the $(m + 1)$st coordinate. Then the vectors

$$e_{m+1}P, e_{m+1}P^2, e_{m+1}P^3, \ldots$$

represent the expected numbers of the different integers after $m + 1$ has been rewritten, after *all* the terms belonging to C in the new sequence have been rewritten, after all the terms belonging to C in the sequence

thus derived have been rewritten, and so on. If $\pi(\cdot)$ is such that $A^N \to 0$ as $N \to \infty$ ($\lambda < 1$), then $P^N \to Q$, a matrix with finite entries, and $e_{m+1}Q$ represents the expected numbers of the different integers in the sentences generated by the grammar (zero, for each $i \in C$).

In our problem, we assume $e_{m+1}Q$ can be directly observed. This fact, together with the relation $R = (I - A)L$ of Theorem 1, and further information obtained from the form of the rules in G, can sometimes be combined to solve the problem. Whether or not this is possible depends on the precise nature of the grammar in question as discussed in [6], but a necessary condition is given by the next theorem.

THEOREM 2. *If for any function $\pi(\cdot)$ which specifies probability distributions on the $G(i)$, observation of $e_{m+1}Q$ determines $\pi(\cdot)$, then $m \geqslant h - n$.*

Proof. The only restraint on $\pi(\cdot)$ is that it sum to one over $G(i)$, for each $i \in C$. There being h rules and n elements of C, there are effectively $h - n$ independent quantities to be determined by $e_{m+1}Q$. But

$$e_{m+1}Q = (L_{11}, \ldots, L_{1,m}, 0, \ldots, 0),$$

and no set of more than m independent variables can be functionally determined by the m variables $L_{11}, \ldots, L_{1,m}$.

In discussing the problem of inferring $\pi(\cdot)$ from average frequencies of words in sentences, we have not taken account of a feature of context-free grammars which distinguishes them from other multitype branching processes [1] having the same expectation matrices. A sentence is not just a collection of integer terms, but a sequence of these terms. By taking advantage of the ordering of the terms in a sequence, we can eliminate the condition in Theorem 2. To do this we consider the procedure of sentence production as it affects *pairs* of adjacent terms in the sequence.

In the example above, the sentence 1, 2, 1, 1 contains one example each of the following pairs (1, 2), (2, 1), (1, 1) as well as (blank, 1) and (1, blank). Inserting blanks at the beginning and end of the sequence is helpful; for example, the initial sequence $m + 1$ contains *no* pairs unless we consider it as blank, $m + 1$, blank, in which case it contains (blank, $m + 1$) and ($m + 1$, blank).

There are $(m + n + 1)^2 - 1$ different pairs possible using elements of T, C and {blank}, (blank, blank) being impossible. In some grammars, of course, some of these pairs will not occur. In any case, we can number

the pairs which can occur and, using the rules of the grammar and the function $\pi(\cdot)$, construct a matrix $\bar{P}$ giving the expected number of the jth type of pair produced by rewriting an ith type pair.

This construction is carried out as follows. If the ith type of pair is (a, b) where $a \in T$, $b \in T$, then $\bar{P}_{ij} = \delta_{ij}$.

If the ith type of pair is (a, b), where $a \in T$ but $b \in C$, then

$$\bar{P}_{ij} = \sum_{r \in G(b)} \pi(r)[\chi_1(j, a, r) + \tfrac{1}{2} \text{ number of } j\text{th type pairs in } \alpha_r]$$

where

$$\chi_1(j, a, r) = 1 \quad \text{if } \alpha_r \text{ is of the form } t, \ldots, u \text{ and } j\text{th pair type is } (a, t),$$
$$= 0 \quad \text{otherwise.}$$

The factor $\frac{1}{2}$ in this formula takes into account that the corresponding pairs are counted again, as being produced by a pair of form $(b, \cdot)$.

Similarly, if the ith type of pair is (a, b), where $a \in C$, $b \in T$, then

$$\bar{P}_{ij} = \sum_{r \in G(a)} \pi(r)[\chi_2(j, b, r) + \tfrac{1}{2} \text{ number of } j\text{th type pairs in } \alpha_r],$$

where

$$\chi_2(j, b, r) = 1 \quad \text{if } \alpha_r \text{ is of the form } t, \ldots, u \text{ and } j\text{th pair type is } (u, b)$$
$$= 0 \quad \text{otherwise.}$$

Finally, if the ith type of pair is (a, b), where $a \in C$, $b \in C$, then

$$\bar{P}_{ij} = \sum_{\substack{r \in G(a) \\ \text{or } r \in G(b)}} \pi(r)[\tfrac{1}{2} \text{ number of } j\text{th type pairs in } \alpha_r] + \sum_{\substack{r \in G(a) \\ s \in G(b)}} \pi(r)\pi(s)\chi_3(j, r, s)$$

where

$$\chi_3(j, r, s) = 1 \quad \text{if } \alpha_r \text{ is of the form } t, \ldots, u \text{ and } \alpha_s \text{ is of the form } v, \ldots, w \text{ and } j\text{th pair type is } (u, v)$$
$$= 0 \quad \text{otherwise.}$$

$\bar{P}_{ij}$ represents the expected number of jth type of pair produced as a consequence of rewriting the terms of an ith type of pair. If $\bar{e}$ is the vector

with 1's in the coordinates representing (blank, $m + 1$) and ($m + 1$, blank) and zeros elsewhere, then $\bar{e}$, $\bar{P}$, and $\bar{Q} = \lim_{N\to\infty} \bar{P}^N$ can play roles analogous to e_{m+1}, P and Q in the solution of our grammatical inference problem. The advantage is that when there are m terminal integers, there are of the order of m^2 terminal pair types which may be observed. Since $h - n$, the number of independent allocations of values of $\pi(\cdot)$, remains the same, the condition in Theorem 2 is clearly no longer necessary.

If $m^2 < h - n$ there is still the possibility of extending these methods to triplets, quadtuples, etc.

To summarize, we have related the output of context-free grammars to their probabilistic structures, in the first instance by using a multitype branching process model; this can work only for grammars satisfying a restriction on the number of rules and the numbers of different types of terms. By taking account of the order of terms within a sentence, the analysis can be extended and the method can be applied to grammars not satisfying this restriction.

REFERENCES

1 T. E. Harris, *The Theory of Branching Processes*, Springer-Verlag, Berlin; Prentice-Hall, Englewood Cliffs, N.J. (1963).

2 J. J. Horning, "A study of grammatical inference," Technical report No. CS 139, Stanford Artificial Intelligence Project, Memo AI-98, Computer Science Department, School of Humanities and Sciences, Stanford University, 1969.

3 M. M. Kherts, "Entropy of languages generated by automated or context-free grammars with a single-value deduction, *Nauchno-Tekhnicheskaia Informatsia Series* **2**(1968).

4 T. J. Li and K. S. Fu, "Automata games, stochastic automata and formal languages," TR-EE 69-1, School of Electrical Engineering, Purdue University, Lafayette, Indiana, 1969.

5 D. B. Peizer and D. L. Olmsted, "Modules of grammar acquisition," *Language* **45** (1969), 60.

6 D. Sankoff, "Branching processes with terminal types: application to context-free grammars," *J. Appl. Prob.* **8**(1971), 233.

Received June, 1970