J. Appl. Prob. **8**, 233–240 (1971)
Printed in Israel

BRANCHING PROCESSES WITH TERMINAL TYPES: APPLICATION TO CONTEXT-FREE GRAMMARS

DAVID SANKOFF, *Centre de recherche mathématiques, Université de Montréal*

1. Definitions

In this note we consider multi-type branching processes where certain *terminal types* of particles, once created, are subject neither to death nor to further reproduction.

Let $T = \{t_1, \cdots, t_m\}$ be the set of terminal types. In the terminology of Harris ((1963), p. 46) each $t \in T$ constitutes a *final group*. Let $C = \{c_1, \cdots, c_n\}$ be the set of non-terminal types and $V = T \cup C$. A particle of type c_i has r_1 offspring of type $t_1, \cdots, r_{m+n}$ of type c_n with probability $p^i(r_1, \cdots, r_{m+n})$. These probabilities define P, the $(m+n) \times (m+n)$ *mean matrix* of the process, where

$$P(m+i,j) = \sum_{r_1=0}^{\infty} \cdots \sum_{r_{m+n}=0}^{\infty} r_j p^i(r_1, \cdots, r_{m+n}), \qquad 1 \leqq i \leqq n, \tag{1}$$

$$P(i,j) = \delta_{ij}, \qquad 1 \leqq i \leqq m. \tag{2}$$

We require that each $c_i \in C$ eventually produce terminal types, i.e., for some $t_j \in T$ and some $N > 0$,

$$P^N(m+i,j) > 0. \tag{3}$$

Of special interest is the *subcritical* case where, as $N \to \infty$, the iterates P^N approach a limit matrix Q having finite entries.

The state of the process at the Nth generation is represented by a row vector X^N with $m+n$ non-negative integer entries. X^0 represents the initial population. The expected value of X^N is the product X^0P^N and the expected final population, when the process has reached completion, is

$$E[X^\infty] = X^0Q. \tag{4}$$

From (2), P is of the form

$$P = \left[\begin{array}{c|c} I_m & Z_{mn} \\ \hline R & A \end{array}\right], \tag{5}$$

Received in revised form 8 June 1970.

where I_m is the $m \times m$ identity matrix and Z_{mn} represents an $m \times n$ matrix with all entries zero.

A necessary and sufficient condition for the process to be subcritical is that the restriction of the process to the non-terminal types be subcritical in the usual sense, i.e., $A^N \to Z_{nn}$, as $N \to \infty$, or equivalently, the maximal eigenvalue of A is less than 1 (cf. Karlin (1966), p. 480).

2. Reconstruction of the process

In the applications, we encounter the following situation. We are given some information about P, including which entries are zero and which are non-zero. We are free to set X^0 to take on a number of values and in each case observe the corresponding $E[X^\infty]$; (we will not consider the statistical procedures for estimating $E[X^\infty]$). The problem is: under what conditions can we specify P completely, i.e., by finding the actual values of the non-zero entries?

From (5) and subcriticality we have that Q is of the form

$$Q = \left[\begin{array}{c|c} I_m & Z_{mn} \\ \hline L & Z_{nn} \end{array}\right]. \tag{6}$$

Theorem 1. Let P and Q be as in (5) and (6). Then

$$R = (I - A)L.$$

Proof.

$$P^N = \left[\begin{array}{c|c} I_m & Z_{mn} \\ \hline L_N & A^N \end{array}\right],$$

where it is easily shown that

$$L_N = \sum_{i=0}^{N} A^i R.$$

Then

$$L = \left[\lim_{N \to \infty} \sum_{i=0}^{N} A^i\right] R. \tag{7}$$

But since $A^N \to Z_{nn}$, the limit in (7) is just $(I - A)^{-1}$ (see, e.g., Kemeny and Snell (1960), p. 22), and hence left-multiplying both sides of (7) by $I - A$ proves the theorem.

Equation (4) and Theorem 1 may sometimes be used to solve the reconstruction problem, as follows. By setting X^0 and observing $E[X^\infty]$ we can specify entries of Q with the help of (4). An upper bound on the amount of this sort of information obtainable is given in Theorem 2.

Theorem 2. Let the vectors $X_1^0, \cdots, X_k^0$ be the available initial populations for a subcritical process (as described above), such that for each X_i^0 we can

observe the corresponding $E[X_i^\infty]$. Consider the vector space of dimension $\boldsymbol{d}$ spanned by vectors consisting of the last n entries of each X_i^0, i.e., by

$$(X_1^0(m+1), \cdots, X_1^0(m+n)), \cdots, (X_k^0(m+1), \cdots, X_k^0(m+n)).$$

Then $\boldsymbol{d} \leqq \min(k, n)$ and the same information about Q obtained from the X_i^0 and $E[X_i^\infty]$ could be obtained from any set of basis vectors of this space and the associated final populations.

Proof. Writing

$$X^0 = (X^0(1), \cdots, X^0(m), Z_{1,n}) + (Z_{1,m}, X^0(m+1), \cdots, X^0(m+n)),$$

by (4),

$$\begin{aligned} E[X^\infty] &= (X^0(1), \cdots, X^0(m), Z_{1,n})Q + (Z_{1,m}, X^0(m+1), \cdots, X^0(m+n))Q \\ &= (X^0(1), \cdots, X^0(m), Z_{1,n}) + (Z_{1,m}, X^0(m+1), \cdots, X^0(m+n))Q. \end{aligned}$$

Then the contribution of the first m entries of X^0 to $E[X^\infty]$ is not affected by Q, and hence gives us no information about Q. It therefore suffices to consider vectors consisting of only the last n entries. Consider the $k \times n$ matrix M constructed by stacking these vectors. By simultaneously carrying out the same elementary row operations on M and the $E[X_i^\infty]$, $i = 1, \cdots, k$, we can construct a matrix B whose non-zero rows are linearly independent and at the same time, by (4), observe the expected final populations resulting from the initial populations corresponding to the rows of B. Elementary row operations being invertible, M and the $E[X_i^\infty]$ could just as well be derived from B and its corresponding final populations, which proves the theorem.

In reconstruction problems, we are usually given certain information about P, even *before* applying (4) and Theorem 1. For example, we usually know which entries of P are zero and perhaps the specific values of some of the other entries. As well, we may be given that some of the unknown entries of P may be functionally dependent. Whether or not in any particular problem P can be completely specified depends strongly on the nature of this additional information.

Generally speaking, if the maximal set of functionally *independent* entries of P is not too large compared to $\boldsymbol{d}$ and m, there is the possibility of reconstruction.

Theorem 3. Given the process and initial populations of Theorem 2. Suppose all entries of P are either specified constants or continuous functions of some maximal set of functionally independent entries $a_1, \cdots, a_s$ (considered as functions of the X_i^0 and $E[X_i^\infty]$). That is, $(a_1, \cdots, a_s)$ is free to take on all values in some open set of $\boldsymbol{R}^s$. Then a necessary condition for the X_i^0 and $E[X_i^\infty]$, $i = 1, \cdots, k$, to implicitly determine the values of $a_1, \cdots, a_s$, is that $s \leqq \boldsymbol{d}m$.

Proof. By Theorem 2, it suffices to consider $\boldsymbol{d}$ different initial populations where the vectors consisting of the last n entries in each are linearly independent.

Each of these is associated with one observed expected final population, which is an m-vector. Then Equation (4) can determine no more than dm independent variable entries in Q (i.e., in L).

Now if the equation

$$F = R - (I - A)L$$

implicitly defines, near $F = 0$, $a_1, \cdots, a_s$ as functions of σ out of these dm entries of L in some open set of R^σ, then $s \leqq \sigma$. For if not, the inverse function theorem (Rudin (1953), p. 177) implies that any σ of the a serve to specify the entries of L and hence the remaining $s - \sigma$ variables a, contrary to the hypothesis in the theorem that $(a_1, \cdots, a_s)$ could take on values in an open set of R^s.

The condition in Theorem 3 is not sufficient, however, since some of the dm variables in L may be functionally related as a consequence of the functional relations imposed on the entries of P.

3. Context-free grammars and probability

For our purposes, a context-free grammar (cf. Ginsburg (1966)) consists of a set $V = T \cup C$ as before, and a finite set G of grammatical rules or *rewrite rules* of the form

$$c_i \to v_1 \cdots v_h,$$

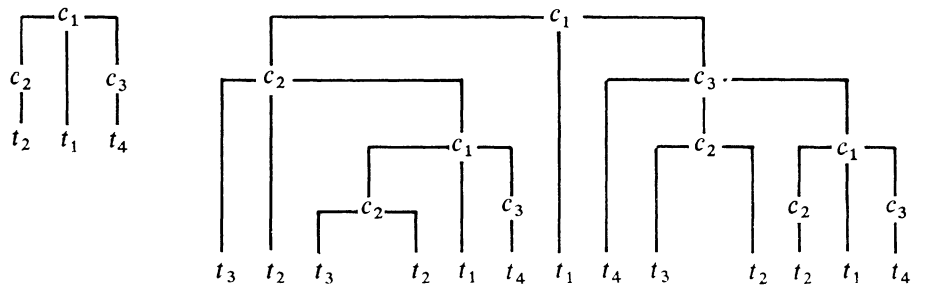
where $h \geqq 1$, $c_i \in C$, $v_j \in V$ for $j = 1, \cdots, h$.

The *sentences* generated by the grammar are all those (and only those) finite concatenations (*strings*) of terminal types (word types) constructed as follows. Start with the string consisting of the single symbol c_1. Rewrite (replace) the symbol c_1 with a new string consisting of the right hand side of any one of the rules in G rewriting c_1. Do the same with any (and all) $c \in C$ (grammatical category types) occurring in the new string. Continue in this manner until a string is obtained consisting entirely of terminal types (all t's and no c's).

For example, let $T = \{t_1, t_2, t_3, t_4\}$, $C = \{c_1, c_2, c_3\}$ and

$$\begin{aligned} G &= \{g_1, \cdots, g_{11}\} \\ &= \{c_1 \to c_2 t_1 c_3, c_2 \to c_2 c_1, c_2 \to t_2 c_1, c_2 \to t_2, \\ &\quad c_2 \to t_3 t_2 c_1, c_2 \to t_3 t_2, c_3 \to t_4 c_2 c_1, c_3 \to t_4 c_1, \\ &\quad c_3 \to t_4, c_3 \to t_4 c_2 c_2, c_3 \to t_4 c_2\}. \end{aligned} \tag{8}$$

Two elements of the infinite set of sentences generated by this grammar are $t_2 t_1 t_4$ and $t_3 t_2 t_3 t_2 t_1 t_4 t_1 t_4 t_3 t_2 t_2 t_1 t_4$. The *structure* of a sentence is its "family tree" or "family history" (Harris (1963), pp. 33, 122), e.g.,



This representation contains the sequence of rules used in producing the sentence, and where they are applied, and incorporates the notion of left-right order inherent in strings.

In general, context-free grammars are restricted so that, in analogy to (4) and the condition of subcriticality, it must be possible for each $c \in C$ to rewrite it, its offspring, etc., until a string containing only t's is attained.

There are a number of probabilistic extensions of the theory of context-free grammars. The approach we shall take is similar to those of Kherts (1968) [as cited by Li and Fu (1969)], Horning (1969), and Peizer and Olmsted (1969). This approach entails the assignment of a suitable probability $\pi(g) > 0$ to each $g \in G$ subject to the condition

$$\sum_{\substack{g \\ \text{rewrites} \\ c}} \pi(g) = 1, \quad \text{for all } c \in C. \tag{9}$$

In this way a non-zero probability is also imposed on each sentence structure which can be generated by the grammar, namely the product of all the $\pi(g)$ for all g used in its production.

The motivation behind these considerations derives from the wish to use formal grammars (of which context-free grammars are a special case) as models within the context of the behavioral study of linguistic performance. The behavioral linguist wishes to know not only the logical possibilities and restraints on sentences, but is also interested in which sentences or sets of sentences are likely and which improbable. He is thus concerned with the theoretical consequences of imposing probability measures on formal grammars. This is true for example in the investigation of language acquisition, in the comparison of related dialects, styles and levels of speech, and in the study of grammatical inference. Context-free grammars are among the simplest formal grammars bearing any strong resemblance to portions of the grammars of natural language (and programming languages) and hence provide a natural starting point for this type of inquiry.

The function $\pi(\cdot)$ on G determines a branching process with terminal types where P is defined by

$$(10)\quad P(m+i,j) = \sum_{\substack{g \\ \text{rewrites} \\ c_i}} \pi(g)[\text{no. of } t_j \text{ (or } c_{j-m}) \text{ on RHS of } g], \quad 1 \leqq i \leqq n$$

in analogy to (1), and

$$P(i,j) = \delta_{ij} \quad \text{for } 1 \leqq i \leqq m;$$

and the initial population is

$$X^0 = (Z_{1,m}, 1, Z_{1,n-1}).$$

By observing a large number of sentences, i.e., strings of terminal types, we can estimate the average number of each type (of word) per sentence. This and (4) determine the first row of L, ($\boldsymbol{d} = 1$), and if the conditions of Theorem 3 are met we may be able to reconstruct P. Note that only certain $\pi(\cdot)$ produce a subcritical process.

4. Example

According to the analysis of Jacobs and Rosenbaum ((1968), p. 57) the rewrite rules in (8) generate the "deep structure" strings of English. This may be seen by making the following correspondence between their notation and ours:

$$\begin{aligned} c_1 &= \text{S (sentence)} & t_1 &= \text{AUX (auxiliary)} \\ c_2 &= \text{NP (noun phrase)} & t_2 &= \text{N (noun)} \\ c_3 &= \text{VP (verb phrase)} & t_3 &= \text{ART (article)} \\ & & t_4 &= \text{VB (verbal).} \end{aligned}$$

Inspection of the rules in (8) gives us the *a priori* information that P is of the form

$$\left[\begin{array}{c|c} I_4 & Z_{43} \\ \hline R & A \end{array}\right] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & a & b & 0 & c & d & 0 \\ 0 & 0 & 0 & 1 & e & f & 0 \end{bmatrix},$$

where a, b, c, d, e, and f depend on the probabilities $\pi(\cdot)$ as follows. By (10),

$$\begin{aligned} a &= \sum_{\substack{g \\ \text{rewrites} \\ c_2}} \pi(g)[\text{number of } t_2 \text{ on RHS of } g]. \\ &= \pi(g_3) + \pi(g_4) + \pi(g_5) + \pi(g_6). \end{aligned}$$

Note that this may be put in the form of (1) without ambiguity in this example.

$$a = p^2(0,1,0,0,1,0,0) + p^2(0,1,0,0,0,0,0) + p^2(0,1,1,0,1,0,0) + p^2(0,1,1,0,0,0,0).$$

Similarly

$$b = \pi(g_5) + \pi(g_6),$$

$$c = \pi(g_2) + \pi(g_3) + \pi(g_5),$$

$$d = \pi(g_2),$$

$$e = \pi(g_7) + \pi(g_8),$$

$$f = \pi(g_7) + 2\pi(g_{10}) + \pi(g_{11}).$$

Condition (9) implies that $d = 1-a$ and it is also clear that $a \geqq b$, $c \geqq d$, but we still have five functionally independent variables to determine in R and A, and Theorem 3 permits at most $\boldsymbol{d}m = 4$. Hence this process cannot be reconstructed without further constraints on the $\pi(\cdot)$.

It is easy, however, to invent processes which may be reconstructed in this sense, simply by ensuring a large number of terminal types and a small number of rules.

Once the matrix P has been reconstructed there still remains the problem of calculating the individual $\pi(g)$, which may or may not be possible, depending on the particular example.

It should be noted that this branching process approach is weak in the sense that it makes no reference to the order of the terminal types in the sentence, and according to certain theories of language this order contains much information about the nature of the grammar. Indeed, for a special class of context-free grammars, the unambiguous grammars (of which (8) is not one), where each initial sentence is associated with a *unique* structure, the order of the words in a sentence gives enough information so that the associated structure may be explicitly determined by one or other algorithm (Griffiths and Petrick (1965)). Thus $\pi(\cdot)$ is easily estimated through observing a large sample of sentences. This is not true, however, in general. In natural languages, moreover, direct observation of the "deep structure" sentences is seldom possible. What is observable, the so-called "surface structure" sentences, are, in some grammars at least, fairly closely related to the deep structure sentences in content but not in order.

Acknowledgements

I wish to thank Martin Goldstein, Norman Pullman and Donald Dawson for their help, and the referee for his suggestions.

References

[1] GINSBURG, S. (1966) *The Mathematical Theory of Context-free Languages.* McGraw-Hill, New York.

[2] GRIFFITHS, T. V. AND PETRICK, S. R. (1965) On the relative efficiencies of context-free grammar recognizers. *Comm. ACM* **8**, 289–300.

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

[4] HORNING, J. J. (1969) *A study of grammatical inference. Technical Report* No. CS 139, Stanford Artificial Intelligence Project, Memo A1–98. Computer Science Department, School of Humanities and Sciences, Stanford University.

[5] JACOBS. R. A. AND ROSENBAUM, P. S. (1968) *English Transformational Grammar.* Blaisdell, Waltham, Mass.

[6] KARLIN, S. (1966) *A First Course in Stochastic Processes.* Academic Press, New York.

[7] KEMENY, J. G. AND SNELL, J. L. (1960) *Finite Markov Chains.* Van Nostrand, Princeton, N.J.

[8] KHERTS, M. M. (1968) Entropy of languages generated by automated or context-free grammars with a single-value detection. *Nauchno-Tekhnicheskaia Informatsia Series* **2**, 29–34.

[9] LI, T. J. AND FU, K. S. (1969) Automata games, stochastic automata and formal languages. TR-EE 69–1. School of Electrical Engineering, Purdue University, Lafayette, Indiana.

[10] PEIZER, D. B. AND OLMSTED. D. L. (1969) Modules of grammar acquisition. *Language* **45**, 60–96.

[11] RUDIN, W. (1953) *Principles of Mathematical Analysis.* McGraw-Hill, New York.