

RESEARCH NOTES ON A PROPERTY OF PROBABILISTIC CONTEXT-FREE GRAMMARS

R. CHAUDHURI

Computer Science Group
Department of Mathematics
East Carolina University
Greenville, North Carolina 27834 U.S.A.

and

A.N.V. RAO

Centre of Applied Mathematics
Department of Mathematics
University of South Florida
Tampa, Florida 33620 U.S.A.

(Received February 14, 1983)

ABSTRACT. It is proved that for a probabilistic context-free language $L(G)$, the population density of a character (terminal symbol) is equal to its relative density in the words of a sample S from $L(G)$ whenever the production probabilities of the grammar G are estimated by the relative frequencies of the corresponding productions in the sample.

KEY WORDS AND PHRASES. Probabilistic grammars, stochastic expectation matrix, consistency and character density.

1980 MATHEMATICS SUBJECT CLASSIFICATION CODES. 68F05, 68F15.

1. INTRODUCTION.

The production probabilities of a context-free grammar may be estimated by the relative frequencies of the corresponding productions in the words of a sample from the language generated by the grammar. It was conjectured by Wetherell [3] that in such a case the expected derivation length and the expected word length of the words in a language would be exactly equal to the mean derivation length and the mean word length of the words in the corresponding sample. The conjecture of Wetherell was proved to be true by Chaudhuri, Pham and Garcia [2]. In this note, we prove that an analogous result holds for character densities also.

2. PRELIMINARIES.

It is assumed that the reader is familiar with elements of formal language theory. Let $G = (V_N, V_T, \sigma, P)$ be a context-free grammar. Assume that, $V_N = \{\sigma = N_1, N_2, \dots, N_r\}$

and $V_T = \{t_1, t_2, \dots, t_s\}$ are the finite sets of non-terminal and terminal symbols respectively. For each $N_i \in V_N$, let $N_i \rightarrow u_{ia}$ ($a=1, 2, \dots, n_i$) be the productions that rewrite the non-terminal N_i . The set P consists of all the productions and σ is the special start symbol. The language corresponding to G will be denoted by $L(G)$.

Definition 1. A probabilistic context-free grammar is a pair (G, π) where G is a context-free grammar and $\pi = (\pi_1, \pi_2, \dots, \pi_r)$ is a vector satisfying the following properties:

- i) for each i , $\pi_i = (p_{i1}, p_{i2}, \dots, p_{in_i})$ where $0 \leq p_{ia} \leq 1$,
 $a = 1, 2, \dots, n_i$ and $p_{ia} = \text{Prob}(N_i \rightarrow u_{ia})$
- ii) for each i , $\sum_{a=1}^{n_i} p_{ia} = 1$.

In the following, we always assume G to be unambiguous, i.e. each word $w \in L(G)$ has a unique derivation starting from σ . The distribution vector π of a probabilistic context-free grammar (G, π) induces a measure μ on $L(G)$. For each word $w \in L(G)$, $\mu(w)$ is precisely the product of the probabilities of the productions used to derive w from the start symbol σ . For details, see [1]. A probabilistic context-free grammar (G, π) is called consistent iff $\sum_{w \in L(G)} \mu(w) = 1$.

Definition 2. Let (G, π) be a probabilistic context-free grammar. Define the matrix $M = (m_{ij})$, $1 \leq i, j \leq r$, as follows:

$$m_{ij} = \sum_{a=1}^{n_i} p_{ia} f_{aj}^{(i)}$$

where n_i is the number of productions of the type $N_i \rightarrow u_{ia}$ and $f_{aj}^{(i)}$ is the number of occurrences of the non-terminal N_j in u_{ia} .

The $(r \times r)$ matrix M is called the stochastic expectation matrix corresponding to (G, π) . If the spectral radius $R(M)$ of the matrix M is less than unity, then (G, π) is called strongly consistent. Also, it was noted in [1,3] that if $R(M) < 1$ then the matrix series $I + M + M^2 + \dots$ converges to a matrix $M^\infty = (I - M)^{-1}$. The matrix M^∞ must be known in order to find the expected derivation length, $EDL(\sigma)$, and the expected word length, $EWL(\sigma)$, and many other statistical characteristics of the language $L(G)$. For details, see [1,2,3].

The production probabilities of a grammar (G, π) can be estimated by the relative frequencies of the corresponding productions in the derivation processes of all the words in a sample S from $L(G)$. The estimated probability of the production $N_i \rightarrow u_{ia}$ is:

$$\bar{p}_{ia} = b_{ia} / \left(\sum_{y=1}^{n_i} b_{iy} \right),$$

where b_{ia} is the frequency of the production $N_i \rightarrow u_{ia}$. Thus, each sample S from $L(G)$ will induce a distribution vector $\bar{\pi}$ and will give rise to a probabilistic context-free grammar $(G, \bar{\pi})$.

Definition 3. Let S be a sample of size λ from $L(G)$. The mean derivation length and the mean word length of the words in S are defined as follows:

$$\begin{aligned} \text{MDL}(S) &= (\text{sum of the derivation lengths of the words in } S) / \lambda \\ \text{MWL}(S) &= (\text{sum of the word lengths of the words in } S) / \lambda. \end{aligned}$$

The following conjecture of Wetherell [3] was proved in [2].

Theorem 1. Let G be any unambiguous context-free grammar and let S be any sample from $L(G)$. Then the probabilistic context-free grammar $(G, \bar{\pi})$ is strongly consistent and

$$\begin{aligned} \text{EDL}(\sigma) &= \text{MDL}(S) \\ \text{and } \text{EWL}(\sigma) &= \text{MWL}(S) \end{aligned}$$

where σ is the start symbol of G and $\bar{\pi}$ is the distribution vector induced by S on G .

3. MAIN RESULT.

The concept of character density was introduced in [1].

Definition 4. The character density $d(t_j)$ of a character (terminal symbol) $t_j \in V_T$ is the relative number of times the character t_j appears in the words of $L(G)$. Let $g_{aj}^{(i)}$ be the number of t_j 's in the consequence u_{ia} of the production $N_i \rightarrow u_{ia}$. Let $T^{(j)}$ be the column vector whose i -th element is

$$T_i^{(j)} = \sum_{a=1}^{n_i} p_{ia} g_{aj}^{(i)} \quad (i=1,2,\dots,r).$$

This term may be interpreted as the average number of t_j 's produced directly from the non-terminal N_i . If (G, π) is a strongly consistent grammar, then it was proved in [1] that for any character $t_j \in V_T$

$$d(t_j) = (1 \ 0 \ 0 \ \dots \ 0) M^\infty \cdot T^{(j)} / \text{EWL}(\sigma).$$

Definition 5. The relative density of a terminal symbol (character) $t_j \in V_T$ in a sample S from $L(G)$ is defined as

$$\text{RD}(t_j, S) = \frac{\text{Total number of times } t_j \text{ appears in the words of } S}{\text{Sum of the word lengths of all words in } S}$$

Note that,

$$\text{RD}(t_j, S) = \left(\sum_{i=1}^r h_{ij} \right) / \left(\sum_{i=1}^r \sum_{j=1}^s h_{ij} \right)$$

where, $h_{ij} = \sum_{a=1}^{n_i} b_{ia} g_{aj}^{(i)}$ is the number of t_j 's produced from the non-terminal N_i in the derivation processes of all the words in the sample S .

The following theorem relates the population density and relative density of a terminal symbol (character) t_j .

Theorem 2. Let G be any unambiguous context-free grammar and let S be any

sample from $L(G)$. Then for any terminal symbol (character) t_j we have:

$$d(t_j) = RD(t_j, S)$$

Proof: As noted in [2], the estimation of the production probabilities of a context-free grammar G by choosing a sample from $L(G)$ gives rise to a probabilistic context-free grammar $(G, \bar{\pi})$ which is strongly consistent. Let M be the stochastic expectation matrix corresponding to $(G, \bar{\pi})$. It was proved in [2] that in such a case the matrix series $I + M + M^2 + \dots$ converges to a matrix M^∞ and moreover for each i ,

$$M_{1i}^\infty = \frac{e_i}{\lambda} \quad (\lambda = \text{sample size})$$

where $e_i = \sum_{a=1}^{n_i} b_{ia}$, b_{ia} being the frequency of the production $N_i \rightarrow u_{ia}$ in the sample S . Note that,

$$\begin{aligned} (1 \ 0 \ 0 \ \dots \ 0) M^\infty \cdot T^{(j)} &= \sum_{i=1}^r M_{1i}^\infty \cdot T_i^{(j)} \\ &= \sum_{i=1}^r \frac{e_i}{\lambda} \cdot \frac{h_{ij}}{e_i} = \frac{1}{\lambda} \sum_{i=1}^r h_{ij} \end{aligned}$$

By virtue of Theorem 1, we have

$$EWL(\sigma) = MWL(S) = \frac{1}{\lambda} \sum_{j=1}^s \sum_{i=1}^r h_{ij}$$

Hence, $d(t_j) = (M^\infty \cdot T^{(j)})_{11} / EWL(\sigma)$

$$\begin{aligned} &= \frac{\sum_{i=1}^r h_{ij}}{\sum_{j=1}^s \sum_{i=1}^r h_{ij}} = RD(t_j, S) . \end{aligned}$$

REFERENCES

1. BOOTH, T.L. and THOMPSON, R.A. Applying Probability Measures to Abstract Languages, IEEE Trans. Comp. C-22 (1973) 442-450.
2. CHAUDHURI, R., PHAM, S., and GARCIA, O.N. Solution of an Open Problem on Probabilistic Context-Free Grammars, IEEE Trans. Comp. (1982) to appear.
3. WETHERELL, C.S. Probabilistic Languages: A Review and Some Open Questions, Computing Surveys 12, No. 4 (1980) 361-379.

Special Issue on Modeling Experimental Nonlinear Dynamics and Chaotic Scenarios

Call for Papers

Thinking about nonlinearity in engineering areas, up to the 70s, was focused on intentionally built nonlinear parts in order to improve the operational characteristics of a device or system. Keying, saturation, hysteretic phenomena, and dead zones were added to existing devices increasing their behavior diversity and precision. In this context, an intrinsic nonlinearity was treated just as a linear approximation, around equilibrium points.

Inspired on the rediscovering of the richness of nonlinear and chaotic phenomena, engineers started using analytical tools from "Qualitative Theory of Differential Equations," allowing more precise analysis and synthesis, in order to produce new vital products and services. Bifurcation theory, dynamical systems and chaos started to be part of the mandatory set of tools for design engineers.

This proposed special edition of the *Mathematical Problems in Engineering* aims to provide a picture of the importance of the bifurcation theory, relating it with nonlinear and chaotic dynamics for natural and engineered systems. Ideas of how this dynamics can be captured through precisely tailored real and numerical experiments and understanding by the combination of specific tools that associate dynamical system theory and geometric tools in a very clever, sophisticated, and at the same time simple and unique analytical environment are the subject of this issue, allowing new methods to design high-precision devices and equipment.

Authors should follow the Mathematical Problems in Engineering manuscript format described at <http://www.hindawi.com/journals/mpe/>. Prospective authors should submit an electronic copy of their complete manuscript through the journal Manuscript Tracking System at <http://mts.hindawi.com/> according to the following timetable:

Manuscript Due	February 1, 2009
First Round of Reviews	May 1, 2009
Publication Date	August 1, 2009

Guest Editors

José Roberto Castilho Piqueira, Telecommunication and Control Engineering Department, Polytechnic School, The University of São Paulo, 05508-970 São Paulo, Brazil; piqueira@lac.usp.br

Elbert E. Neher Macau, Laboratório Associado de Matemática Aplicada e Computação (LAC), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 12227-010 São Paulo, Brazil ; elbert@lac.inpe.br

Celso Grebogi, Department of Physics, King's College, University of Aberdeen, Aberdeen AB24 3UE, UK; grebogi@abdn.ac.uk