Statistical Properties of Probabilistic Context-Free Grammars

Zhiyi Chi*
University of Chicago

We prove a number of useful results about probabilistic context-free grammars (PCFGs) and their Gibbs representations. We present a method, called the relative weighted frequency method, to assign production probabilities that impose proper PCFG distributions on finite parses. We demonstrate that these distributions have finite entropies. In addition, under the distributions, sizes of parses have finite moment of any order. We show that Gibbs distributions on CFG parses, which generalize PCFG distributions and are more powerful, become PCFG distributions if their features only include frequencies of production rules in parses. Under these circumstances, we prove the equivalence of the maximum-likelihood (ML) estimation procedures for these two types of probability distributions on parses. We introduce the renormalization of improper PCFGs to proper ones. We also study PCFGs from the perspective of stochastic branching processes. We prove that with their production probabilities assigned by the relative weighted frequency method, PCFGs are subcritical, i.e., their branching rates are less than one. We also show that by renormalization, connected supercritical PCFGs become subcritical ones. Finally, some minor issues, including identifiability and approximation of production probabilities of PCFGs, are discussed.

1. Introduction

This article proves a number of useful properties of probabilistic context-free grammars (PCFGs). In this section, we give an introduction to the results and related topics.

1.1 Assignment of Proper PCFG Distributions

Finite parse trees, or parses, generated by a context-free grammar (CFG) can be equipped with a variety of probability distributions. The simplest way to do this is by production probabilities. First, for each nonterminal symbol in the CFG, a probability distribution is placed on the set of all productions from that symbol. Then each finite parse tree is allocated a probability equal to the product of the probabilities of all productions in the tree. More specifically, denote a finite parse tree by τ. For any production rule $A \to \alpha$ of the CFG, let $f(A \to \alpha; \tau)$ be the number of times it occurs in τ. Let R be the set of all production rules. Then

$$p(\tau) = \prod_{(A\to\alpha)\in R} p(A \to \alpha)^{f(A\to\alpha;\tau)}.$$

A CFG with a probability distribution on its parses assigned in this way is called a probabilistic context-free grammar (PCFG) (Booth and Thompson 1973; Grenander

* Department of Statistics, University of Chicago, Chicago, IL 60637, USA. Email: chi@galton.uchicago.edu. This work was supported by the Army Research Office (DAAL03-92-G-0115), the National Science Foundation (DMS-9217655), and the Office of Naval Research (N00014-96-1-0647).

1976)[1] and the probability distribution is called a PCFG distribution. A PCFG may be improper, i.e., the total probability of parses may be less than one. For instance, consider the CFG in Chomsky normal form:

$$\begin{aligned} S &\rightarrow SS \\ S &\rightarrow a \end{aligned} \tag{1}$$

where S is the only nonterminal symbol, and a is the only terminal symbol. If $p(S \rightarrow SS) = p$, then $p(S \rightarrow a) = 1 - p$. Let x_h be the total probability of all parses with height no larger than h. Clearly, x_h is increasing. It is not hard to see that $x_{h+1} = 1 - p + px_h^2$. Therefore, the limit of x_h, which is the total probability of all parses, is a solution for the equation $x = 1 - p + px^2$. The equation has two solutions: 1 and $1/p - 1$. It can be shown that x is the smaller of the two: $x = \min(1, 1/p - 1)$. Therefore, if $p > 1/2$, $x < 1$—an improper probability.

How to assign proper production probabilities is quite a subtle problem. A sufficient condition for proper assignment is established by Chi and Geman (1998), who prove that production probabilities estimated by the maximum-likelihood (ML) estimation procedure (or relative frequency estimation procedure, as it is called in computational linguistics) always impose proper PCFG distributions. Without much difficulty, this result can be generalized to a simple procedure, called "the relative weighted frequency" method, which assigns proper production probabilities of PCFGs. We will give more details of the generalization in Section 3 and summarize the method in Proposition 1.

1.2 Entropy and Moments of Parse Tree Sizes

As a probabilistic model for languages, the PCFG model has several important statistical properties, among which is the entropy of PCFG distribution on parses. Entropy is a measure of the uncertainty of a probability distribution. The larger its entropy, the less one can learn about parses randomly sampled from the distribution. As an example, suppose we have a set S of N parses—or any objects—$\tau_1, \ldots, \tau_N$, where N is very large. We may ask how much one can learn from the sentence "τ is a random sample from S." At one extreme, let the distribution on S be $p(\tau_1) = 1$, and $p(\tau_i) = 0$, for $i \neq 1$. Then, because with probability one, $\tau = \tau_1$, there is no uncertainty about the sample. In other words, we can get full information from the above sentence. At the other extreme, suppose the distribution is $p(\tau_1) = \ldots = p(\tau_N) = 1/N$. In this case, all the elements of S are statistically equivalent. No specific information is given about τ that would make it possible to know it from S. Greater effort is required—for example, enumerating all the elements in S—to find what τ is. Since S is big, the uncertainty about the sample is then much greater. Correspondingly, for the two cases, the entropy is 0 and $N \log N \gg 0$, respectively.

Entropy plays a central role in the theory of information. For an excellent exposition of this theory, we refer the reader to Cover and Thomas (1991). The theory has been applied in probabilistic language modeling (Mark, Miller, and Grenander 1996; Mark et al. 1996; Johnson 1998), natural language processing (Berger, Della Pietra, and Della Pietra 1996; Della Pietra, Della Pietra, and Lafferty 1997), as well as computational vision (Zhu, Wu, and Mumford 1997). In addition, all the models proposed in these articles are based on an important principle called the maximum entropy principle. Chapter 11 of Cover and Thomas (1991) gives an introduction to this principle.

[1] A probabilistic context-free grammar is also called a stochastic context-free grammar (SCFG)

Briefly, the maximum entropy principle says that among all the distributions that satisfy the same given conditions, the one that achieves the largest entropy should be the model of choice. For a distribution p on parses, its entropy is

$$H(p) = \sum_{\tau} p(\tau) \log \frac{1}{p(\tau)}.$$

In order that the maximum entropy principle makes sense, all the candidate distributions should have finite entropies, and this is usually implicitly assumed.

Take Mark, Miller, and Grenander's (1996) model, for example. First, a PCFG distribution p is selected to serve as a "reference" distribution on parses. Then, by invoking the minimum relative entropy principle, which is a variant of the maximum entropy principle, the distribution that minimizes

$$D(q\|p) = \sum_{\tau} q(\tau) \log \frac{q(\tau)}{p(\tau)} = \sum_{\tau} q(\tau) \log \frac{1}{p(\tau)} - H(q)$$

subject to a set of constraints incorporating context-sensitive features is chosen to be the distribution of the model. It is then easy to see that the assumption that $H(q)$ is finite is necessary.

Conceptually, having finite entropy is a basic requirement for a "good" probabilistic model because a probability distribution with infinite entropy has too much uncertainty to be informative.

Problems regarding entropies of PCFGs are relatively easy to tackle because they can be studied analytically. Several authors have reported results on this subject, including Miller and O'Sullivan (1992), who gave analytical results on the rates of entropies of *improper* PCFGs. It is worthwhile to add a few more results on entropies of proper PCFGs. In this paper, we show that the entropies of PCFG distributions imposed by production probabilities assigned by the relative weighted frequency method are finite (Section 4, Corollary 2).

In addition to entropy, we will also study the moment of sizes of parses. The moment is of statistical interest because it gives information on how sizes of parses are distributed. For PCFG distributions, the first moment of sizes of parses, i.e., the mean size of parses, is directly linked with the entropy: the mean size of parses is finite if and only if the entropy is. The second moment of sizes is another familiar quantity. The difference between the second moment and the mean squared is the variance of sizes, which tells us how "scattered" sizes of parses are distributed around the mean. Proposition 2 shows that, under distributions imposed by production probabilities assigned by the relative weighted frequency method, sizes of parses have finite moment of any order.

1.3 Gibbs Distributions on Parses and Renormalization of Improper PCFGs

Besides PCFG distributions, a CFG can be equipped with many other types of probability distributions. Among the most widely studied is the Gibbs distribution (Mark, Miller, and Grenander 1996; Mark et al. 1996; Mark 1997; Abney 1997). Gibbs distributions arise naturally by invoking the maximum entropy principle. They are considered to be more powerful than PCFG distributions because they incorporate more features, especially context-sensitive features, of natural languages, whereas PCFG distributions only consider frequencies of production rules. On the other hand, Gibbs distributions are not always superior to PCFG distributions. A Gibbs distribution, with only frequencies of production rules in parse as its features, turns into a PCFG. More specifically, we will show in Proposition 4 in Section 5, that a CFG equipped with a Gibbs distribution of the form

$$P_\lambda(\tau) = \frac{1}{Z_\lambda} \prod_{(A\to\alpha)\in R} e^{\lambda_{A\to\alpha} f(A\to\alpha;\tau)} \qquad (2)$$

is actually a PCFG, and we can get the production probabilities of the PCFG explicitly from the Gibbs form.

The fact that a Gibbs distribution of the form in (2) is imposed by production probabilities has a useful consequence. Suppose p is an improper PCFG distribution. If we write the sum of p over all parses as Z, and assign to each parse tree a new probability equal to $p(\tau)/Z$, then we renormalize p to a Gibbs distribution $\tilde{p}$ on parses. What (2) implies is that $\tilde{p}$ is also a PCFG distribution (Corollary 3). Moreover, in Section 6 we will show that, under certain conditions, $\tilde{p}$ is subcritical.

There is another issue about the relations between PCFG distributions and Gibbs distributions of the form in (2), from a statistical point of view. Although PCFG distributions are special cases of Gibbs distributions in the sense that the former can be written in the form of the latter, PCFG distributions cannot be put in the framework of Gibbs distributions if they have different parameter estimation procedures. We will compare the maximum-likelihood (ML) estimation procedures for these two distributions. As will be seen in Section 5, numerically these two estimation procedures are different. However, Corollary 4 shows that they are *equivalent* in the sense that estimates by the two procedures impose the same distributions on parses. For this reason, a Gibbs distribution may be considered a generalization of PCFG, not only in form, but also in a certain statistical sense.

1.4 Branching Rates of PCFGs

Because of their context-free nature, PCFG distributions can also be studied from the perspective of stochastic processes. A PCFG can be described by a random branching process (Harris 1963), and its asymptotic behavior can be characterized by its branching rate. A branching process, or its corresponding PCFG, is called subcritical (critical, supercritical), if its branching rate < 1 ($= 1$, > 1). A subcritical PCFG is always proper, whereas a supercritical PCFG is always improper. Many asymptotic properties of supercritical branching processes are established by Miller and O'Sullivan (1992). Chi and Geman (1998) proved the properness of PCFG distributions imposed by estimated production probabilities, and around the same time Sánchez and Benedí (1997) established the subcriticality of the corresponding branching processes, hence their properness. In this paper we will explore properties of branching rate further. First, in Proposition 5, we will show that if a PCFG distribution is imposed by production probabilities assigned by the relative weighted frequency method, then the PCFG is subcritical. The result generalizes that of Sánchez and Benedí (1997), and has a less involved proof. Then in Proposition 7, we will demonstrate that a connected and improper PCFG, after being renormalized, becomes a subcritical PCFG.

1.5 Identifiability and Approximation of Production Probabilities

Returning to the statistical aspect of PCFGs, we will discuss the identifiability of production probabilities of PCFGs as well as parameters of Gibbs distributions. Briefly speaking, production probabilities of PCFGs are identifiable, which means that different production probabilities always impose different distributions on parses (Proposition 8). In contrast, for the Gibbs distribution given by (2), the λ parameters are not identifiable; in fact, there are infinitely many different λ that impose the same Gibbs distribution.

Finally, in Proposition 9, we propose a method to approximate production probabilities. Perhaps the most interesting part about the result lies in its proof, which is largely information theoretic. We apply the Kullback-Leibler divergence, which is the information distance between two probability distributions, to prove the convergence of the approximation. In information theory literature, the Kullback-Leibler divergence is also called the relative entropy. We also use Lagrange multipliers to solve the constrained minimization problem involved. Both Kullback-Leibler divergence and Lagrange multipliers method are becoming increasingly useful in statistical modeling, e.g., modeling based on the maximum entropy principle.

1.6 Summary

As a simple probabilistic model, the PCFG model is applied to problems in linguistics and pattern recognition that do not involve much context sensitivity. To design sensible PCFG distributions for such problems, it is necessary to understand some of the statistical properties of the distributions. On the other hand, the PCFG model serves as a basis for more expressive linguistic models. For example, many Gibbs distributions are built upon PCFG distributions by defining

$$P(\tau) = \frac{p(\tau)e^{\lambda \cdot U(\tau)}}{Z},$$

where p is a PCFG distribution. Therefore, in order for the Gibbs distribution P to have certain desired statistical properties, it is necessary for p to have those properties first. This paper concerns some of the fundamental properties of PCFGs. However, the methods used in the proofs are also useful for the study of statistical issues on other probabilistic models.

This paper proceeds as follows: In Section 2, we gather the notations for PCFGs that will be used in the remaining part of the paper. Section 3 establishes the relative weighted frequency method. Section 4 proves the finiteness of the entropies of PCFG distributions when production probabilities are assigned using the relative weighted frequency method. In addition, finiteness of the moment of sizes of parses are proved. Section 5 discusses the connections between PCFG distributions and Gibbs distributions on parses. Renormalization of improper PCFGs is also discussed here. In Section 6, PCFGs are studied from the random branching process point of view. Finally, in Section 7, identifiability of production probabilities and their approximation are addressed.

2. Notations and Definitions

In this section, we collect the notations and definitions we will use for the remaining part of the paper.

Definition 1

A context-free grammar (CFG) G is a quadruple (N, T, R, S), where N is the set of variables, T the set of terminals, R the set of production rules, and $S \in N$ is the start symbol.[2] Elements of N are also called nonterminal symbols. N, T, and R are always

[2] Some of our discussion requires that each sentential form have only finitely many parses. For this reason, we shall assume that in G, there are no null or unit productions.

assumed to be finite. Let Ω denote the set of finite parse trees of G, an element of which is always denoted as τ. For each $\tau \in \Omega$ and each production rule $(A \rightarrow \alpha) \in R$, define $f(A \rightarrow \alpha; \tau)$ to be the number of occurrences, or frequency, of the rule in τ, and $f(A; \tau)$ to be the number of occurrences of A in τ. $f(A; \tau)$ and $f(A \rightarrow \alpha; \tau)$ are related by

$$f(A;\tau) = \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \rightarrow \alpha) \in R}} f(A \rightarrow \alpha; \tau).$$

Define $h(\tau)$ as the height of τ, which is the number of nonterminal nodes on the longest route from τ's root to its terminal nodes. Define $|\tau|$ as the size of τ, which is the total number of nonterminal nodes in τ. For any $A \in N$ and any sentential form $\gamma \in (N \cup T)^*$, define $n(A;\ \gamma)$ as the number of instances of A in γ. Define $|\gamma|$ as the length of the sentential form.

Definition 2

Let $A \in \tau$ denote that the symbol A occurs in the parse τ. If $A \in \tau$, let τ_A be the left-most maximum subtree of τ rooted in A, which is the subtree of τ rooted in A satisfying the condition that if $\tau' \neq \tau_A$ is also a subtree of τ rooted in A, then τ' is either a subtree of τ_A, or a right sibling of τ_A, or a subtree of a right sibling of τ_A. Let A_τ be the root of τ_A, which is the left-most "shallowest" instance of A in τ.

Definition 3

For any two symbols $A \in N$ and $B \in N \cup T$, not necessarily different, B is said to be reachable from A in G, if there is a sequence of symbols $A_0, A_1, \ldots, A_n$ with $A_0 = A$ and $A_n = B$, and a sequence of sentential forms $\alpha_0, \ldots, \alpha_{n-1}$, such that each $A_i \rightarrow \alpha_i$ is a production in R and each α_i contains the next symbol A_{i+1}. G is called connected if all elements in $N \cup T$ can be reached from all nonterminal symbols.

We now define the probabilistic version of reachability in CFG. Suppose p is a distribution on Ω. For any two symbols $A \in N$ and $B \in N \cup T$, not necessarily different, B is said to be reachable from A in G under p, if there is a $\tau \in \Omega$ with $p(\tau) > 0$ and there is a subtree τ' of τ, such that τ' is rooted in A and $B \in \tau'$. G is called connected under p if all symbols in $N \cup T$ can be reached from all nonterminal symbols under p.

Definition 4

A system of production probabilities of G is a function $p : R \rightarrow [0,1]$ such that for any $A \in N$,

$$\sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \rightarrow \alpha) \in R}} p(A \rightarrow \alpha) = 1. \tag{3}$$

We will also use p to represent the PCFG probability distribution on parses imposed by p, via the formula

$$p(\tau) = \prod_{(A \rightarrow \alpha) \in R} p(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \tau)}. \tag{4}$$

Similarly, for any estimated system of production probabilities $\hat{p}$, we will also use $\hat{p}$ to represent the probability distribution on parses imposed by $\hat{p}$. We will write $p(\Omega)$ as the total probability of all finite parse trees in Ω.

Definition 5
Now we introduce a notation in statistics. Let p be an arbitrary distribution on Ω and $g(\tau)$ a function of $\tau \in \Omega$. The expected value of g under the distribution p, denoted $E_p g(\tau)$, is defined as

$$E_p g(\tau) = \sum_{\tau \in \Omega} p(\tau) g(\tau).$$

Definition 6
All the parse trees we have so far seen are rooted in S. It is often useful to investigate subtrees of parses, therefore it is necessary to consider trees rooted in symbols other than S. We call a tree rooted in $A \in N$ a parse (tree) rooted in A if it is generated from A by the production rules in R. Let Ω_A be the set of all finite parse trees with root A. Define p_A as the probability distribution on Ω_A imposed by a system of production probabilities p, via (4). Also extend the notions of height and size of parses to trees in Ω_A.

When we write $p_A(\tau)$, we always assume that τ is a parse tree rooted in A. When $p = p_A$, $E_p g(\tau)$ equals $E_p g(\tau) = \sum_{\Omega_A} p_A(\tau) g(\tau)$. We will use $p(\Omega_A)$ instead of $p_A(\Omega_A)$ to denote the total probability of finite parse trees in Ω_A. With no subscripts, Ω and p are assumed to be Ω_S and p_S, respectively.

For convenience, we also extend the notion of trees to terminals. For each terminal $t \in T$, define Ω_t as the set of the single "tree" $\{t\}$. Define $p_t(t) = 1$, $|t| = 0$ and $h(t) = 0$.

For this paper we make the following assumptions:

1. For each symbol $A \neq S$, there is at least one parse τ with root S such that $A \in \tau$. This will guarantee that each $A \neq S$ can be reached from S;

2. When a system of production probabilities p is not explicitly assigned, each production rule $(A \to \alpha) \in R$ is assumed to have positive probability, i.e., $p(A \to \alpha) > 0$. This guarantees that there are no useless productions in the PCFG.

3. Relative Weighted Frequency

The relative weighted frequency method is motivated by the maximum-likelihood (ML) estimation of production probabilities. We shall first give a brief review of ML estimation.

We consider two cases of ML estimation. In the first case, we assume the data are **fully observed**, which means that all the samples are fully observed finite parses trees. Let $\tau_1, \tau_2, \ldots, \tau_n$ be the samples. Then the ML estimate of $p(A \to \alpha)$ is the ratio between the total number of occurrences of the production $A \to \alpha$ in the samples and the total number of occurrences of the symbol A in the samples,

$$\hat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} f(A \to \alpha; \tau_i)}{\sum_{i=1}^{n} f(A; \tau_i)}. \tag{5}$$

Because of the form of the estimator in (5), ML estimation in the full observation case is also called relative frequency estimation in computational linguistics. This simple estimator, as shown by Chi and Geman (1998), assigns proper production probabilities for PCFGs.

In the second case, the parse trees are unobserved. Instead, the yields $Y_1 = Y(\tau_1), \ldots, Y_n = Y(\tau_n)$, which are the left-to-right sequences of terminals of the unknown parses $\tau_1, \ldots, \tau_n$, form the data. It can be proved that the ML estimate $\hat{p}$ is given by

$$\hat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} E_{\hat{p}}[f(A \to \alpha; \tau) | \tau \in \Omega_{Y_i}]}{\sum_{i=1}^{n} E_{\hat{p}}[f(A; \tau) | \tau \in \Omega_{Y_i}]}, \tag{6}$$

where Ω_Y is the set of all parses with yield Y, i.e., $\Omega_Y = \{\tau \in \Omega : Y(\tau) = Y\}$.

Equation (6) cannot be solved in closed form. Usually, the solution is computed by the EM algorithm with the following iteration (Baum 1972; Baker 1979; Dempster, Laird, and Rubin 1977):

$$\hat{p}_{k+1}(A \to \alpha) = \frac{\sum_{i=1}^{n} E_{\hat{p}_k}[f(A \to \alpha; \tau) | \tau \in \Omega_{Y_i}]}{\sum_{i=1}^{n} E_{\hat{p}_k}[f(A; \tau) | \tau \in \Omega_{Y_i}]}. \tag{7}$$

Like $\hat{p}$ in (5), $\hat{p}_k$ for $k > 0$ impose proper probability distributions on Ω (Chi and Geman 1998).

To unify (6) and (7), expand $E_{\hat{p}_k}[f(A \to \alpha; \tau) | \tau \in \Omega_{Y_i}]$, by the definition of expectation, into

$$E_{\hat{p}_k}[f(A \to \alpha; \tau) | \tau \in \Omega_{Y_i}] = \sum_{\tau \in \Omega_{Y_i}} f(A \to \alpha; \tau) \hat{p}_k(\tau | \tau \in \Omega_{Y_i}).$$

Let Λ be the set of parses whose yields belong to the data, i.e., $\Lambda = \{\tau : Y(\tau) \in \{Y_1, \ldots, Y_n\}\}$. For each $\tau \in \Lambda$, let $y = Y(\tau)$ and

$$W(\tau) = \sum_{i: Y_i = y} \hat{p}_k(\tau | \tau \in \Omega_y).$$

Then we observe that, for any production rule $A \to \alpha$,

$$\sum_{i=1}^{n} \sum_{\tau \in \Omega_{Y_i}} f(A \to \alpha; \tau) \hat{p}_k(\tau | \tau \in \Omega_{Y_i}) = \sum_{\tau \in \Lambda} f(A \to \alpha; \tau) W(\tau)$$

$$\Rightarrow \sum_{i=1}^{n} E_{\hat{p}_k}[f(A \to \alpha;\tau)|\tau \in \Omega_{Y_i}] = \sum_{\tau\in\Lambda} f(A \to \alpha;\tau)W(\tau).$$

Therefore, (7) is transformed into

$$\hat{p}_{k+1}(A \to \alpha) = \frac{\sum_{\tau\in\Lambda} f(A \to \alpha;\tau)W(\tau)}{\sum_{\tau\in\Lambda} f(A;\tau)W(\tau)}.$$

The ML estimator in (6) can also be written in the above form, as can be readily checked by letting Λ be the set $\{\tau_1, \ldots, \tau_n\}$ and $W(\tau)$, for each $\tau \in \Lambda$, be the number of occurrences of τ in the data. In addition, in both full observation cases and partial observation cases, we can divide the weights of $W(\tau)$ by a constant so that their sum is 1.

The above discussion leads us to define a procedure to assign production probabilities as follows. First, pick an arbitrary finite subset Λ of Ω, with every production rule appearing in the trees in Λ. Second, assign to each $\tau \in \Lambda$ a positive weight $W(\tau)$ such that $\sum_{\tau\in\Lambda} W(\tau) = 1$. Finally, define a system of production probabilities p by

$$p(A \to \alpha) = \frac{\sum_{\tau\in\Lambda} f(A \to \alpha;\tau)W(\tau)}{\sum_{\tau\in\Lambda} f(A;\tau)W(\tau)}. \tag{8}$$

Because of the similarity between (5) and (8), we call the procedure to assign production probabilities by (8) the "relative weighted frequency" method.

Proposition 1
Suppose all the symbols of N occur in the parses of Λ, and all the parses have positive weight. Then the production probabilities given by (8) impose proper distributions on parses.

Proof
The proof is almost identical to the one given by Chi and Geman (1998). Let $q_A = p$ (derivation tree rooted in A fails to terminate). We will show that $q_S = 0$ (i.e., derivation trees rooted in S always terminate). For each $A \in V$, let $\tilde{f}(A;\tau)$ be the number of non-root instances of A in τ. Given $\alpha \in (V \cup T)^*$, let α_i be the ith symbol of the sentential form α. For any $A \in V$

$$\begin{aligned}
q_A &= p\left(\bigcup_{(A\to\alpha)\in R} \bigcup_i \{\text{derivation begins } A \to \alpha_i \text{ and } \alpha_i \text{ fails to terminate}\}\right) \\
&= \sum_{(A\to\alpha)\in R} p(A \to \alpha) p\left(\bigcup_i \{\alpha_i \text{ fails to terminate}\}\right) \\
&\le \sum_{(A\to\alpha)\in R} p(A \to \alpha) \sum_i p(\{\alpha_i \text{ fails to terminate}\}) \\
&= \sum_{(A\to\alpha)\in R} p(A \to \alpha) \sum_{B\in V} n(B;\alpha) q_B
\end{aligned}$$

$$= \sum_{B \in V} q_B \left\{ \frac{\sum_{(A \to \alpha) \in R} n(B;\alpha) \sum_{\tau \in \Lambda} f(A \to \alpha;\tau) W(\tau)}{\sum_{\tau \in \Lambda} f(A;\tau) W(\tau)} \right\}$$

$$\Longrightarrow q_A \sum_{\tau \in \Lambda} f(A;\tau) W(\tau) \leq \sum_{B \in V} q_B \sum_{\tau \in \Lambda} \sum_{(A \to \alpha) \in R} n(B;\alpha) f(A \to \alpha;\tau) W(\tau)$$

Sum over $A \in V$:

$$\begin{aligned} \sum_{A \in V} q_A \sum_{\tau \in \Lambda} f(A;\tau) W(\tau) &\leq \sum_{B \in V} q_B \sum_{\tau \in \Lambda} \sum_{A \in V} \sum_{(A \to \alpha) \in R} n(B;\alpha) f(A \to \alpha;\tau) W(\tau) \\ &= \sum_{B \in V} q_B \sum_{\tau \in \Lambda} \tilde{f}(B;\tau) W(\tau) \end{aligned}$$

i.e.,

$$\sum_{A \in V} q_A \sum_{\tau \in \Lambda} (\tilde{f}(A;\tau) - f(A;\tau)) W(\tau) \geq 0$$

Clearly, for every $\tau \in \Lambda$, $\tilde{f}(A;\tau) = f(A;\tau)$ whenever $A \neq S$ and $\tilde{f}(S;\tau) = f(S;\tau) - 1$. Hence $q_S = 0$, completing the proof. □

Corollary 1
Under the same assumption of Proposition 1, for each symbol $A \in N$, $p(\Omega_A) = 1$.

Proof
For any $A \in N$, there is a $\tau \in \Lambda$ such that $A \in \tau$. Since $p(\tau) > 0$, this implies A is reachable from S under p. Using the notation given in Definition 2, we have

$$\begin{aligned} q_S &\geq p(\{A \in \tau \text{ and } \tau_A \text{ fails to terminate}\}) \\ &= p(\{\tau_A \text{ fails to terminate}\} | A \in \tau) p(A \in \tau) \\ \Longrightarrow & \quad p(\{\tau_A \text{ fails to terminate}\} | A \in \tau) = 0, \end{aligned}$$

since $q_S = 0$ and $p(A \in \tau) > 0$. By the nature of PCFGs, the form of τ_A is distributed according to p_A, independent of its location in τ or of the choice of subtrees elsewhere in τ. Therefore the conditional probability of τ_A failing to terminate, given that A occurs in τ, equals q_A, proving that $q_A = 0$. □

4. Entropy and Moments of Parse Tree Sizes

In this section, we will first show that if production probabilities are assigned by the relative weighted frequency method, then they impose PCFG distributions under which parse tree sizes have finite moment of any order. Based on this result, we will then demonstrate that such PCFG distributions have finite entropy and give the explicit form of the entropy.

The mth moment of sizes of parses is given by

$$E_p|\tau|^m = \sum_{\tau \in \Omega} p(\tau)|\tau|^m$$

and the entropy of a PCFG distribution p is given by

$$H(p) = \sum_{\tau \in \Omega} p(\tau) \log \frac{1}{p(\tau)},$$

To make the proofs more readable, we define, for any given $\Lambda = \{\tau_1, \ldots, \tau_n\}$,

$$F(A \to \alpha) = \sum_{\tau \in \Lambda} f(A \to \alpha; \tau) W(\tau),$$

for any $(A \to \alpha) \in R$, and

$$F(A) = \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} F(A \to \alpha) = \sum_{\tau \in \Lambda} f(A; \tau) W(\tau),$$

for any $A \in N$; that is, $F(A \to \alpha)$ is the weighted sum of the number of occurrences of the production rule $A \to \alpha$ in Λ and $F(A)$ is the weighted sum of the number of occurrences of A in Λ.

The relative weighted frequency method given by (8) can be written as

$$p(A \to \alpha) = \frac{F(A \to \alpha)}{F(A)} \tag{9}$$

We have the following simple lemma:

Lemma 1
For any $A \in N$,

$$\sum_{A \in N} F(A) = \sum_{\tau \in \Lambda} |\tau| W(\tau) \tag{10}$$

and

$$\sum_{B \in N} \sum_{\substack{\gamma \text{ s.t.} \\ (B \to \gamma) \in R}} F(B \to \gamma) n(A;\, \gamma) = \begin{cases} F(S) - 1 & \text{if } A = S \\ F(A) & \text{if } A \neq S \end{cases} \tag{11}$$

(If $\sum_{\tau \in \Lambda} W(\tau) \neq 1$, $F(S) - 1$ should be changed to $F(S) - \sum_{\tau \in \Lambda} W(\tau)$.)

Proof
For the first equation,

$$\sum_{A \in N} F(A) = \sum_{A \in N} \sum_{\tau \in \Lambda} f(A; \tau) W(\tau) = \sum_{\tau \in \Lambda} \sum_{A \in N} f(A; \tau) W(\tau) = \sum_{\tau \in \Lambda} |\tau| W(\tau).$$

For the second equation,

$$\begin{aligned} \sum_{B \in N} \sum_{\substack{\gamma \text{ s.t.} \\ (B \to \gamma) \in R}} F(B \to \gamma) n(A;\, \gamma) &= \sum_{B \in N} \sum_{\substack{\gamma \text{ s.t.} \\ (B \to \gamma) \in R}} \sum_{\tau \in \Lambda} f(B \to \gamma; \tau) W(\tau) n(A;\, \gamma) \\ &= \sum_{\tau \in \Lambda} W(\tau) \sum_{B \in N} \sum_{\substack{\gamma \text{ s.t.} \\ (B \to \gamma) \in R}} f(B \to \gamma; \tau) n(A;\, \gamma) \quad (12) \end{aligned}$$

For each A,

$$\sum_{B \in N} \sum_{\substack{\gamma \text{ s.t.} \\ (B \rightarrow \gamma) \in R}} f(B \rightarrow \gamma; \tau) n(A; \gamma)$$

is the number of nonroot instances of A in τ. When $A \neq S$, the number of nonroot instances of A in τ is equal to $f(A; \tau)$. Substitute this into (12) to prove (11) for the case $A \neq S$. The case $A = S$ is similarly proved. □

Proposition 2

Suppose all the symbols in N occur in the parses of Λ, and all parses have positive weights. If the production probabilities p are assigned by the relative weighted frequency method in (8), then for each $m \in \mathbf{N} \cup \{0\}$, $E_p|\tau|^m < \infty$.

Proof

We shall show that for any $A \in N$, if $p = p_A$, then $E_p|\tau|^m < \infty$. When $m = 0$, this is clearly true. Now suppose the claim is true for $0, \ldots, m-1$. For each $A \in N$ and $k \in \mathbf{N}$, define

$$M_{k,A} = \sum_{\substack{\tau \in \Omega_A \\ h(\tau) \leq k}} p_A(\tau)|\tau|^m.$$

It is easy to check

$$M_{k+1,A} = \sum_{\substack{\alpha \in (N \cup T)^* \\ (A \rightarrow \alpha) \in R}} \sum_{\substack{\tau_1, \ldots, \tau_L \\ \tau_i \in \Omega_{\alpha_i} \\ h(\tau_i) \leq k}} (1 + \sum_{i=1}^{L} |\tau_i|)^m p(A \rightarrow \alpha) p_{\alpha_1}(\tau_1) \ldots p_{\alpha_L}(\tau_L), \tag{13}$$

where for ease of typing, we write L for $|\alpha|$. For fixed α, write

$$(1 + \sum_{i=1}^{L} |\tau_i|)^m = P(|\tau_1|, \ldots, |\tau_L|) + \sum_{i=1}^{L} |\tau_i|^m.$$

P is a polynomial in $|\tau_1|, \ldots, |\tau_L|$, each term of which is of the form

$$|\tau_1|^{s_1} |\tau_2|^{s_2} \ldots |\tau_L|^{s_L}, \; 0 \leq s_i < m, \; s_1 + s_2 + \ldots s_L \leq m. \tag{14}$$

By induction hypothesis, there is a constant $C > 1$, such that for all $0 \leq s < m$ and $A \in N \cup T$,

$$\sum_{\tau \in \Omega_A} p_A(\tau)|\tau|^s = E_{p_A}|\tau|^s < C.$$

Then for each term with the form given in (14),

$$\sum_{\substack{\tau_1, \ldots, \tau_L \\ \tau_i \in \Omega_{\alpha_i}}} |\tau_1|^{s_1} \ldots |\tau_L|^{s_L} p_{\alpha_1}(\tau_1) \ldots p_{\alpha_L}(\tau_L) = \prod_{i=1}^{L} \sum_{\tau_i \in \Omega_{\alpha_i}} |\tau_i|^{s_i} p_{\alpha_i}(\tau_i) \leq C^L.$$

There are less than $L^m = |\alpha|^m$ terms in $P(|\tau_1|,\dots,|\tau_L|)$. Hence

$$\sum_{\substack{\tau_1,\dots,\tau_L\\ \tau_i\in\Omega_{\alpha_i}\\ h(\tau_i)\le k}} P(|\tau_1|,\dots,|\tau_L|)p(A\to\alpha)p_{\alpha_1}(\tau_1)\dots p_{\alpha_L}(\tau_L) \le |\alpha|^m C^{|\alpha|}.$$

So we get

$$\begin{aligned}
M_{k+1,A} \;\le\;& \sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} |\alpha|^m C^{|\alpha|} p(A\to\alpha)\\
&+ \sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} \sum_{\substack{\tau_1,\dots,\tau_{|\alpha|}\\ \tau_i\in\Omega_{\alpha_i}\\ h(\tau_i)\le k}} \sum_{i=1}^{|\alpha|} |\tau_i|^m p(A\to\alpha)p_{\alpha_1}(\tau_1)\dots p_{\alpha_{|\alpha|}}(\tau_{|\alpha|})\\
\le\;& \sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} |\alpha|^m C^{|\alpha|} p(A\to\alpha) + \sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i} p(A\to\alpha).
\end{aligned}$$

Because the set of production rules is finite, the length of a sentential form that occurs on the right-hand side of a production rule is upper bounded, i.e.,

$$\sup\{|\alpha| : \text{ for some } A\in N,\ (A\to\alpha)\in R\} < \infty.$$

Therefore we can bound $(|\alpha|+1)^m C^{|\alpha|}$ by a constant, say, K. Then we get

$$M_{k+1,A} \le K + \sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i} p(A\to\alpha). \tag{15}$$

Replace $p(A\to\alpha)$ by $F(A\to\alpha)/F(A)$, then multiply both sides of (15) by $F(A)$ and sum over all $A\in N$ with $F(A)>0$. By (10) and (11), we then get

$$\begin{aligned}
\sum_{A\in N} M_{k+1,A}F(A) \;\le\;& K\sum_{\tau\in\Lambda}|\tau|W(\tau) + \sum_{A\in N}\sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i}F(A\to\alpha) \quad \text{(by (10))}\\
=\;& K\sum_{\tau\in\Lambda}|\tau|W(\tau) + \sum_{A\in N}\sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} \sum_{B\in N} n(B;\alpha)M_{k,B}F(A\to\alpha)\\
=\;& K\sum_{\tau\in\Lambda}|\tau|W(\tau) + \sum_{B\in N} M_{k,B}\sum_{A\in N}\sum_{\substack{\alpha\in(N\cup T)^*\\ \text{s.t. } (A\to\alpha)\in R}} n(B;\alpha)F(A\to\alpha)\\
=\;& K\sum_{\tau\in\Lambda}|\tau|W(\tau) + \sum_{B\ne S} M_{k,B}F(B) + M_{k,S}(F(S)-1). \quad \text{(by (11))}
\end{aligned}$$

Because for each $A\in N$, $M_{k+1,A}\ge M_{k,A}$, we get

$$\begin{aligned}
&\sum_{A\in N} M_{k,A}F(A) \le K\sum_{\tau\in\Lambda}|\tau|W(\tau) + \sum_{A\ne S} M_{k,A}F(A) + M_{k,S}(F(S)-1)\\
&\Longrightarrow\quad M_{k,S} \le K\sum_{\tau\in\Lambda}|\tau|W(\tau) < \infty.
\end{aligned}$$

Letting $k \to \infty$, by $M_{k,S} \uparrow E_{p_S}|\tau|^m$, we get $E_{p_S}|\tau|^m \leq K\sum_{\tau\in\Lambda}|\tau|W(\tau) < \infty$. To complete the induction, we shall show for every $A \in N \cup T$ other than S, $E_{p_A}|\tau|^m < \infty$.

By conditional expectation, there is (see Definition 2 for the notations $A \in \tau$ and τ_A)

$$E_{p_S}(|\tau|^m) = E_{p_S}(|\tau|^m \; |A \in \tau)p_S(A \in \tau) + E_{p_S}(|\tau|^m \; |A \notin \tau)p_S(A \notin \tau)$$
$$\Longrightarrow \quad E_{p_S}(|\tau|^m \; |A \in \tau) \leq \frac{E_{p_S}|\tau|^m}{p_S(A \in \tau)} < \infty, \tag{16}$$

since $p_S(A \in \tau) > 0$. Because $|\tau_A| < |\tau|$, $E_{p_S}(|\tau_A|^m \; |A \in \tau) < \infty$.

As in the proof of Corollary 1, τ_A is independent of its location and other part of τ, and is distributed by p_A. Therefore

$$p_S(\tau_A \; |A \in \tau) = p_A(\tau_A),$$

which leads to $E_{p_A}|\tau|^m = E_{p_S}(|\tau_A|^m \; |A \in \tau) < \infty$. □

From Proposition 2 it follows that the mean size of parses is finite under p. Since $f(A \to \alpha; \tau) \leq |\tau|$ for each production $A \to \alpha$, it follows that the mean frequency of $f(A \to \alpha; \tau)$ is finite. The next proposition gives the explicit form of the mean frequency in terms of the production probabilities assigned by the relative weighted frequency method.

Proposition 3
Under the same conditions of Proposition 2, the mean frequency of the production rule $(A \to \alpha) \in R$ is the weighted sum of the numbers of its occurrences in parses of Λ, with weights $W(\tau)$, i.e.,

$$E_p f(A \to \alpha; \tau) = \sum_{\tau\in\Lambda} f(A \to \alpha; \tau)W(\tau) \tag{17}$$

Proof
Fix $(A \to \alpha) \in R$. For each $C \in N$, write $E(C)$ for $E_{p_C} f(A \to \alpha; \tau)$. We shall find the linear relations between $E(C)$. To this end, for each $\tau \in \Omega_C$, let $C \to \gamma$ be the production rule applied to τ's root. Suppose γ is composed of m symbols, $\gamma_1, \ldots, \gamma_m$, and $\tau_1, \ldots, \tau_m$ are the daughter subtrees of τ rooted in $\gamma_1, \ldots, \gamma_m$, respectively. Then

$$f(A \to \alpha; \tau) = \chi(C \to \gamma) + \sum_{k=1}^{m} f(A \to \alpha; \tau_k),$$

where

$$\chi(C \to \gamma) = \begin{cases} 0 & \text{if } C \to \gamma \neq A \to \alpha, \\ 1 & \text{otherwise.} \end{cases}$$

Multiply both sides by $p(\tau)$ and sum over all $\tau \in \Omega_C$ which have $C \to \gamma$ as the production rule applied at the root. By the definition of PCFG, $p(\tau) = p(C \to$

$\tau)p(\tau_1)\cdots p(\tau_m)$, and τ_k can be any parse in Ω_{γ_k}. Therefore, by factorization, we get

$$\begin{aligned}\sum_{\tau\in\Omega_{C\to\gamma}} \chi(C\to\gamma)p(\tau) &= \sum_{\tau\in\Omega_{C\to\gamma}} \chi(C\to\gamma)p(C\to\gamma)p(\tau_1)\cdots p(\tau_m)\\ &= p(C\to\gamma)\chi(C\to\gamma)\prod_{k=1}^{m} p(\Omega_{\gamma_k})\\ &= p(C\to\gamma)\chi(C\to\gamma),\quad (\text{All } p(\Omega_{\gamma_k})=1 \text{ by Proposition 1}),\end{aligned}$$

where $\Omega_{C\to\gamma}$ stands for the set of trees in which $C\to\gamma$ is the rule applied at the root.

Similarly, for each k,

$$\begin{aligned}\sum_{\tau\in\Omega_{C\to\gamma}} f(C\to\gamma;\tau_k)p(\tau) &= p(C\to\gamma)\sum_{\tau\in\Omega_{C\to\gamma}} f(C\to\gamma;\tau_k)p(\tau_k)\prod_{\substack{i=1\\ i\neq k}}^{m} p(\tau_i)\\ &= p(C\to\gamma)E(\gamma_k).\end{aligned}$$

Therefore we get

$$\begin{aligned}\sum_{\tau\in\Omega_{C\to\gamma}} p(\tau)f(A\to\alpha;\tau) &= p(C\to\gamma)(\chi(C\to\gamma)+\sum_{k=1}^{m}E(\gamma_k))\\ &= p(C\to\gamma)(\chi(C\to\gamma)+\sum_{\substack{B\in N\\ \text{s.t. } B\in\gamma}} n(B;\gamma)E(\gamma_k)).\end{aligned}$$

Sum over all production rules for C. The left-hand side totals $E(C)$ and

$$E(C) = \sum_{\substack{\gamma\in(N\cup T)^*\\ \text{s.t. } (C\to\gamma)\in R}} p(C\to\gamma)(\chi(C\to\gamma)+\sum_{\substack{B\in N\\ \text{s.t. } B\in\gamma}} n(B;\,\gamma)E(B)).$$

Replace $p(C\to\gamma)$ by $F(C\to\gamma)/F(C)$, according to (9). Then multiply both sides by $F(C)$ and sum both sides over all $C\in N$. We get

$$\begin{aligned}\sum_{C\in N} F(C)E(C) &= \sum_{C\in N}\sum_{\substack{\gamma\in(N\cup T)^*\\ \text{s.t. } (C\to\gamma)\in R}} F(C\to\gamma)\chi(C\to\gamma)\\ &\quad+\sum_{C\in N}\sum_{\substack{\gamma\in(N\cup T)^*\\ \text{s.t. } (C\to\gamma)\in R}} F(C\to\gamma)\sum_{\substack{B\in N\\ \text{s.t. } B\in\gamma}} n(B;\,\gamma)E(B)\\ &= F(A\to\alpha)+\sum_{B\in N}E(B)\sum_{C\in N}\sum_{\substack{\gamma \text{ s.t.}\\ (C\to\gamma)\in R}} F(C\to\gamma)n(B;\,\gamma)\\ &= F(A\to\alpha)+\sum_{B\in N}E(B)F(B)-E(S)\quad(\text{By (11)})\\ \Longrightarrow&\quad E(S)=F(A\to\alpha),\end{aligned}$$

completing the proof of (17). □

Now we can calculate the entropy of p in terms of production probabilities.

Corollary 2
Under the conditions in Proposition 2,

$$H(p) = \sum_{(A\to\alpha)\in R} F(A\to\alpha)\log\frac{1}{F(A\to\alpha)} - \sum_{A\in N} F(A)\log\frac{1}{F(A)},$$

which is clearly finite.

Proof
The calculation goes as follows,

$$\begin{aligned}
H(p) &= \sum_{\tau\in\Omega} p(\tau)\log\frac{1}{p(\tau)} \\
&= \sum_{\tau\in\Omega} p(\tau)\log\frac{1}{\prod\limits_{(A\to\alpha)\in R} p(A\to\alpha)^{f(A\to;\tau)}} \\
&= \sum_{\tau\in\Omega} p(\tau)\sum_{(A\to\alpha)\in R} f(A\to\alpha;\tau)\log\frac{1}{p(A\to\alpha)} \\
&= \sum_{(A\to\alpha)\in R}\sum_{\tau\in\Omega} p(\tau)f(A\to\alpha;\tau)\log\frac{1}{p(A\to\alpha)} \quad \text{(Exchange the order of summation)} \\
&= \sum_{(A\to\alpha)\in R} E_p f(A\to\alpha;\tau)\log\frac{1}{p(A\to\alpha)} \\
&= \sum_{(A\to\alpha)\in R}\sum_{\tau\in\Lambda} f(A\to\alpha;\tau)W(\tau)\log\frac{F(A)}{F(A\to\alpha)} \\
&= \sum_{(A\to\alpha)\in R}\sum_{\tau\in\Lambda} f(A\to\alpha;\tau)W(\tau)\log F(A) \\
&\quad - \sum_{(A\to\alpha)\in R}\sum_{\tau\in\Lambda} f(A\to\alpha;\tau)W(\tau)\log F(A\to\alpha) \\
&= \sum_{A\in N} F(A)\log F(A) - \sum_{(A\to\alpha)\in R} F(A\to\alpha)\log F(A\to\alpha). \qquad \square
\end{aligned}$$

5. Gibbs Distributions on Parses and Renormalization of Improper PCFGs

A Gibbs distribution on parses has the form

$$P_\lambda(\tau) = \frac{e^{\lambda\cdot U(\tau)}}{Z_\lambda},$$

where $Z_\lambda = \sum e^{\lambda\cdot U(\tau)}$, and $\lambda = \{\lambda_i\}$ and $U(\tau) = \{U_i(\tau)\}$ are constants and functions on Ω, respectively, both indexed by elements in a finite set I. The inner product $\lambda\cdot U = \sum\lambda_i U_i$ is called the potential function of the Gibbs distribution and Z_λ is called the partition number for the exponential $e^{\lambda\cdot U}$.

The functions U_i are usually considered features of parses and the constants λ_i are weights of these features. The index set I and the functions $U_i(\tau)$ can take various forms. Among the simplest choices for I is R, the set of production rules, and correspondingly,

$$U(\tau) = f(\tau) = \{f(A \to \alpha; \tau)\}_{(A\to\alpha)\in R}. \tag{18}$$

Given constants λ, if $Z_\lambda < \infty$, then we get a Gibbs distribution on parses given by

$$P_\lambda(\tau) = \frac{e^{\lambda \cdot f(\tau)}}{Z_\lambda}. \tag{19}$$

A proper PCFG distribution is a Gibbs distribution of the form in (19). To see this, let $\lambda_{A\to\alpha} = \log p(A \to \alpha)$ for each $(A \to \alpha) \in R$. Then

$$Z_\lambda = \sum_{\tau\in\Omega} e^{\lambda\cdot f(\tau)} = \sum_{\tau\in\Omega} p(\tau) = 1$$

$$\Longrightarrow \quad p(\tau) = \prod_{(A\to\alpha)\in R} p(A\to\alpha)^{f(A\to\alpha;\,\tau)} = e^{\lambda\cdot f(\tau)} = \frac{1}{Z_\lambda} e^{\lambda\cdot U(\tau)},$$

which is a Gibbs form.

A Gibbs distribution usually is not a PCFG distribution, because its potential function in general includes features other than frequencies of production rules. What if its potential function only has frequencies as features? More specifically, is the Gibbs distribution in (19) a PCFG distribution? The next proposition gives a positive answer to this question.

Proposition 4
The Gibbs distribution P_λ given by (19) is a PCFG distribution. That is, there are production probabilities p, such that for every $\tau \in \Omega$,

$$P_\lambda(\tau) = \prod_{(A\to\alpha)\in R} p(A\to\alpha)^{f(A\to\alpha;\tau)}.$$

Proof
The Gibbs distributions we have seen so far are only defined for parses rooted in S. By obvious generalization, we can define for each nonterminal symbols A the partition number

$$Z_\lambda(A) = \sum_{\tau\in\Omega_A} e^{\lambda\cdot f(\tau)}$$

and the Gibbs distribution $P(\tau)$ on parses rooted in A. For simplicity, also define $Z_\lambda(t) = 1$ and $P_t(t) = 1$ for each $t \in T$.

We first show $Z_\lambda(A) < \infty$ for all A. Suppose $(S \to \alpha) \in R$ with $|\alpha| = n$. The sum of $e^{\lambda\cdot f(\tau)}$ over all $\tau \in \Omega_S$ with $S \to \alpha$ being applied at the root is equal to $e^{\lambda_{S\to\alpha}} Z_\lambda(\alpha_1) \dots Z_\lambda(\alpha_n)$, while less than the sum of $e^{\lambda\cdot f(\tau)}$ over all $\tau \in \Omega_S$, which is $Z_\lambda(S)$. Therefore,

$$Z_\lambda(S) \geq e^{\lambda_{S\to\alpha}} Z_\lambda(\alpha_1) \dots Z_\lambda(\alpha_n).$$

Since $Z_\lambda < \infty$ and $Z_\lambda(A) > 0$, for all A, it follows that $Z_\lambda(\alpha_i)$ is finite. For any variable A, there are variables $A_0 = S, A_1, \dots, A_n = A \in N$ and sentential forms $\alpha^{(0)}, \dots, \alpha^{(n-1)} \in$

$(N \cup T)^*$, such that $(A_i \to \alpha^{(i)}) \in R$ and $A_{i+1} \in \alpha^{(i)}$. By the same argument as above, we get

$$Z_\lambda(A_i) \geq e^{\lambda_{A_i \to \alpha^{(i)}}} \prod_{k=1}^{|\alpha^{(i)}|} Z_\lambda(\alpha_k^{(i)}),$$

where $\alpha_k^{(i)}$ is the kth element in $\alpha^{(i)}$. By induction, $Z_\lambda(A) < \infty$.

Now for $(A \to \alpha) \in R$, with $|\alpha| = n$, define

$$p(A \to \alpha) = \frac{1}{Z_\lambda(A)} e^{\lambda_{A \to \alpha}} Z_\lambda(\alpha_1) \dots Z_\lambda(\alpha_n), \tag{20}$$

Since $Z_\lambda(A)$ and $Z_\lambda(\alpha_i)$ are finite, $p(A \to \alpha)$ is well defined.

The p's form a system of production probabilities, because for each $A \in N$,

$$\sum_{(A \to \alpha) \in R} p(A \to \alpha) = \frac{1}{Z_\lambda(A)} \sum_{(A \to \alpha) \in R} e^{\lambda_{A \to \alpha}} \prod_{k=1}^{|\alpha|} Z_\lambda(\alpha_k) = \frac{1}{Z_\lambda(A)} \sum_{\tau \in \Omega_A} e^{\lambda \cdot f(\tau)} = 1$$

We shall prove, by induction on $h(\tau)$, that

$$P_\lambda(\tau) = \prod_{(A \to \alpha) \in R} p(A \to \alpha)^{f(A \to \alpha; \tau)}.$$

When $h(\tau) = 0$, τ is just a terminal, and the equation is obviously true. Suppose the equation is true for all $\tau \in \Omega_A$ with $h(\tau) < h$, and all $A \in N$. For any $\tau \in \Omega_A$ with $h(\tau) = h$, let $A \to \beta = \beta_1 \dots \beta_m$ be the production rule applied at the root. Then

$$P_A(\tau) = \frac{1}{Z_\lambda(A)} e^{\lambda \cdot f(\tau)} = \frac{1}{Z_\lambda(A)} e^{\lambda_{A \to \beta}} \prod_{k=1}^{m} e^{\lambda \cdot f(\tau_k)},$$

where τ_k is the daughter subtree of τ rooted in β_k. Each τ_k has height $< h$. Hence, by induction assumption,

$$\begin{aligned}
& \frac{1}{Z_\lambda(\beta_k)} e^{\lambda \cdot f(\tau_k)} = P_\lambda(\tau_k) = \prod_{(B \to \alpha) \in R} p(B \to \alpha)^{f(B \to \alpha; \tau_k)} \\
\Longrightarrow \quad & e^{\lambda \cdot f(\tau_k)} = Z_\lambda(\beta_k) \prod_{(B \to \alpha) \in R} p(B \to \alpha)^{f(B \to \alpha; \tau_k)} \\
\Longrightarrow \quad & P_A(\tau) \\
& = \frac{1}{Z_\lambda(A)} e^{\lambda_{A \to \beta}} \prod_{k=1}^{m} e^{\lambda \cdot f(\tau_k)} \\
& = \frac{1}{Z_\lambda(A)} e^{\lambda_{A \to \beta}} \prod_{k=1}^{m} Z_\lambda(\beta_k) \prod_{(B \to \alpha) \in R} p(B \to \alpha)^{f(B \to \alpha; \tau_k)} \\
& = p(A \to \beta) \prod_{k=1}^{m} \prod_{(B \to \alpha) \in R} p(B \to \alpha)^{f(B \to \alpha; \tau_k)} \\
& = \prod_{(B \to \alpha) \in R} p(B \to \alpha)^{f(B \to \alpha; \tau)},
\end{aligned}$$

proving P_λ is imposed by p. □

Proposition 4 has a useful application to the renormalization of improper PCFGs. Suppose a PCFG distribution p on $\Omega = \Omega_S$ is improper. We define a new, proper distribution $\tilde{p}$ on Ω, by

$$\tilde{p}(\tau) = \frac{p(\tau)}{p(\Omega)}, \quad \tau \in \Omega.$$

We call $\tilde{p}$ the renormalized distribution of p on Ω. We can also define the renormalized distribution of p_A on Ω_A, for each $A \in N$, by

$$\tilde{p}_A(\tau) = \frac{p_A(\tau)}{p(\Omega_A)}, \quad \tau \in \Omega_A. \tag{21}$$

Comparing $\tilde{p}$ with (19), we see that $\tilde{p}$ is a Gibbs distribution with frequencies of production rules as features. Therefore, by Proposition 4, $\tilde{p}$ is a PCFG distribution, and from the proof of Proposition 4, we get Corollary 3.

Corollary 3
Suppose the production probabilities of the improper distribution p are positive for all the production rules. Then the renormalized distributions $\tilde{p}$ are induced by the production probabilities

$$\tilde{p}(A \to \alpha) = \frac{1}{p(\Omega_A)} p(A \to \alpha) \prod_{B \in N} p(\Omega_B)^{n(B;\,\alpha)}. \tag{22}$$

Therefore, $\tilde{p}$ on Ω is a PCFG distribution.

Proof
The only thing we have not mentioned is that $\lambda_{A\to\alpha} = \log p(A \to \alpha)$ are all bounded, since p are all positive. □

We have seen that PCFG distributions can be expressed in the form of Gibbs distributions. However, from the statistical point of view, this is not enough for regarding PCFG distributions as special cases of Gibbs distributions. An important statistical issue about a distribution is the estimation of its parameters. To equate PCFG distributions with special cases of Gibbs distributions, we need to show that estimators for production probabilities of PCFGs and parameters of Gibbs distributions produce the same results.

Among many estimation procedures, the maximum-likelihood (ML) estimation procedure is commonly used. In the full observation case, if the data is composed of $\tau_1, \ldots, \tau_n$, then the estimator for the system of production probabilities is

$$\hat{p} = \{\hat{p}(A \to \alpha)\} = \arg\max \prod_{i=1}^{n} \prod_{(A\to\alpha)\in R} p(A \to \alpha)^{f(A\to\alpha;\tau_i)}, \tag{23}$$

subject to

$$\sum_{(A\to\alpha)\in R} p(A \to \alpha) = 1,$$

for any $A \in N$ and the estimator for parameters of Gibbs distributions with λ of the form in (19) is

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{\lambda \cdot U(\tau_i)}}{Z_\lambda}, \tag{24}$$

In addition, the ML estimate $\hat{p}$ in (23) can be analytically solved and the solution is given by Equation (5).

In the partial observation case, if $Y_1, \ldots, Y_n$ are the observed yields, then the estimators for the two distributions are

$$\hat{p} = \{\hat{p}(A \to \alpha)\} = \arg\max \prod_{i=1}^{n} \sum_{Y(\tau)=Y_i} \prod_{(A\to\alpha)\in R} p(A \to \alpha)^{f(A\to\alpha;\tau)}, \tag{25}$$

subject to

$$\sum_{(A\to\alpha)\in R} p(A \to \alpha) = 1,$$

for any $A \in N$, and

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \sum_{\tau\in\Omega_{Y_i}} \frac{e^{\lambda \cdot U(\tau)}}{Z_\lambda}, \tag{26}$$

respectively.

We want to compare the ML estimators for the two distributions and see if they produce the same results in some sense. Since the parameters p serve as base numbers in PCFG distributions, whereas λ are exponents in Gibbs distributions, to make the comparison sensible, we take the logarithms of $\hat{p}$ and ask whether or not $\log p$ and $\hat{\lambda}$ are the same. Since the ML estimation procedure for PCFGs involves constrained optimization, whereas the estimation procedure for Gibbs distributions only involves unconstrained optimization, it is reasonable to suspect $\log \hat{p} \neq \hat{\lambda}$. Indeed, numerically $\log \hat{p}$ and $\hat{\lambda}$ are different. For example, the estimator (23) only gives one estimate of the system of production probabilities, whereas the estimator (24) may yield infinitely many solutions. Such uniqueness and nonuniqueness of estimates is related to the identifiability of parameters. We will discuss this in more detail in Section 7.

Despite their numerical differences, the ML estimators for PCFG distributions and Gibbs distributions with the form (19) are *equivalent*, in the sense that the estimates produced by the estimators impose the same distributions on parses. Because of this, in the context of ML estimation of parameters, we can regard PCFG distributions as special cases of Gibbs distributions.

Corollary 4
If $\hat{p}$ is the solution of (23), then $\log \hat{p}$ is a solution of ML estimation (24). Similarly, if $\hat{p}$ is a solution of (25), then $\log \hat{p}$ is a solution of ML estimation (26). Hence, the estimates of production probabilities of PCFG distributions and parameters of Gibbs distributions with the form (19) impose the same distributions on parses.

Proof
Suppose $\hat{\lambda}$ is a solution for (24). By Proposition 4, the Gibbs distribution $P_{\hat{\lambda}}$ is imposed by a system of production probabilities $\hat{p}$. Then $\hat{p}$ is the solution of (23). Let $\tilde{\lambda} = \log \hat{p}$, i.e., $\tilde{\lambda}(A \to \alpha) = \log \hat{p}(A \to \alpha)$. Then $\tilde{\lambda}$ impose the same distribution on parses as $\hat{\lambda}$. Therefore $\tilde{\lambda}$ are also a solution to (24). This proves the first half of the result. The second half is similarly proved. □

6. Branching Rates of PCFGs

In this section, we study PCFGs from the perspective of stochastic branching processes. Adopting the set-up given by Miller and O'Sullivan (1992), we define the mean matrix **M** of p as a $|N| \times |N|$ square matrix, with its (A, B)th entry being the expected number of variables B resulting from rewriting A:

$$\mathbf{M}(A, B) = \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} p(A \to \alpha) n(B; \alpha). \tag{27}$$

Clearly, **M** is a nonnegative matrix.

We say $B \in N$ can be reached from $A \in N$, if for some $n > 0$, $\mathbf{M}^{(n)}(A, B) > 0$, where $\mathbf{M}^{(n)}(A, B)$ is the (A, B)th element of $\mathbf{M}^n$. **M** is irreducible if for any pair $A, B \in N$, B can be reached from A. The corresponding branching process is called connected if **M** is irreducible (Walters 1982). It is easy to check that these definitions are equivalent to Definition 3.

We need the result below for the study of branching processes.

Theorem 1: (Perron-Frobenius)
Let $\mathbf{M} = [m_{ij}]$ be a nonnegative $k \times k$ matrix.

1. There is a nonnegative eigenvalue ρ such that no eigenvalue of A has absolute value greater than ρ.
2. Corresponding to the eigenvalue ρ there is a nonnegative left (row) eigenvector $\nu = (\nu_1, \ldots, \nu_k)$ and a nonnegative right (column) eigenvector

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}.$$

3. If **M** is irreducible then ρ is a simple eigenvalue (i.e., the multiplicity of ρ is 1), and the corresponding eigenvectors are strictly positive (i.e. $u_i > 0$, $v_i > 0$ all i).

The eigenvalue ρ is called the branching rate of the process. A branching process is called subcritical (critical, supercritical), if $\rho < 1$ ($\rho = 1$, $\rho > 1$). We also say a PCFG is subcritical (critical, supercritical), if its corresponding branching process is. When a PCFG is subcritical, it is proper. When a PCFG is supercritical, it is improper.

The next result demonstrates that production probabilities assigned by the relative weighted frequency method impose subcritical PCFG distributions.

Proposition 5
For p assigned by the relative weighted frequency method (8) and **M** by (27),

$$\rho < 1. \tag{28}$$

Proof
Let μ be the right eigenvector of ρ, as described in item (2) of Theorem 1. We have $\mathbf{M}\mu = \rho\mu$. For each variable A,

$$\sum_{B \in N} \mathbf{M}(A,B)\mu(B) = \rho\mu(A).$$

Therefore

$$\sum_{B \in N} \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} p(A \to \alpha) n(B;\alpha)\mu(B) = \rho\mu(A).$$

Replacing $p(A \to \alpha)$ by $F(A \to \alpha)/F(A)$, according to (9),

$$\sum_{B \in N} \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} \frac{F(A \to \alpha)}{F(A)} n(B;\alpha)\mu(B) = \rho\mu(A).$$

Multiply both sides by $F(A)$ and sum over $A \in N$. By (11),

$$\sum_{A \in N} F(A)\mu(A) - \mu(S) = \rho \sum_{A \in N} F(A)\mu(A). \tag{29}$$

We need to show that $\mu(S) > 0$. Assume $\mu(S) = 0$. Then for any $n > 0$, since $\mathbf{M}^n\mu = \rho^n\mu$, we have

$$\sum_{A \in N} \mathbf{M}^{(n)}(S,A)\mu(A) = \rho^n\mu(S) = 0.$$

For each $A \in N$, $\mathbf{M}^{(n)}(S,A)\mu(A) = 0$. Because each $A \in N$ is reachable from S under p, there is $n > 0$ such that $\mathbf{M}^{(n)}(S,A) > 0$. So we get $\mu(A) = 0$. Hence $\mu = 0$. This contradicts the fact that μ is a nonnegative eigenvector of **M**. Therefore $\mu(S) > 0$. By (29),

$$\sum_{A \in N} F(A)\mu(A) > \rho \sum_{A \in N} F(A)\mu(A) \geq 0$$

This proves $\rho < 1$. □

We will apply the above result to give another proof of Proposition 2. Before doing this, we need to introduce a spectral theorem, which is well-known in matrix analysis.

Theorem 2
Suppose **M** is an $n \times n$ real matrix. Let $\sigma(\mathbf{M})$ be the largest absolute value of **M**'s eigenvalues. Then

$$\sigma(\mathbf{M}) = \lim_{n \to \infty} \|\mathbf{M}^n\|^{1/n},$$

where $||\mathbf{M}||$ is the norm of $\mathbf{M}$ defined by

$$||\mathbf{M}|| = \sup_{|\vec{v}|=1} |\mathbf{M}\vec{v}|.$$

Now we can prove the following result.

Proposition 6
If $\mathbf{M}$ given by (27) has branching rate $\rho < 1$, then for each $m \in \mathbf{N} \cup \{0\}$,

$$E_p|\tau|^m < \infty. \tag{30}$$

Proof
We repeat the proof of Proposition 2 from Section 4 up to (15). Then, instead of summing over A, we observe that (15) can be written as

$$M_{k+1,\,A} \leq K + \sum_{B \in N} \mathbf{M}(A,B)M_{k,\,B}.$$

Write $\{M_{k,\,A}\}_{A\in N}$ as $\vec{M}_k$, which is a vector indexed by $A \in N$. We then have

$$\vec{M}_{k+1} \leq K\mathbf{1} + \mathbf{M}\vec{M}_k,$$

where $\mathbf{1}$ is defined as $\{1,\ldots,1\}$, and for two-column vectors $\vec{\mu}$ and $\vec{\nu}$, $\vec{\mu} \leq \vec{\nu}$ means each component of μ is $\leq$ the corresponding component of ν. Since the components in $K\mathbf{1}$, $\mathbf{M}$ and $\vec{M}_k$ are positive, the above relation implies

$$\mathbf{M}\vec{M}_{k+1} \leq K\mathbf{M}\mathbf{1} + \mathbf{M}^2\vec{M}_k.$$

Hence, we get

$$\vec{M}_{k+2} \leq K\mathbf{1} + \mathbf{M}\vec{M}_{k+1} \leq K\mathbf{1} + K\mathbf{M}\mathbf{1} + \mathbf{M}^2\vec{M}_k.$$

By induction, we get

$$\vec{M}_k \leq K\sum_{j=0}^{k-2} \mathbf{M}^j\mathbf{1} + \mathbf{M}^{k-1}\vec{M}_1,$$

$$\Longrightarrow \quad |\vec{M}_k| \leq K\sum_{j=0}^{k-2} ||\mathbf{M}^j|||\mathbf{1}| + ||\mathbf{M}^{k-1}|||\vec{M}_1|. \tag{31}$$

By Theorem 2, for any $\rho < \rho' < 1$, $||\mathbf{M}^n|| = o(\rho'^n)$. Then (31) implies that $|\vec{M}_k|$ is bounded. Since $\vec{M}_k$ are positive and increasing, it follows that $\vec{M}_k$ converge. □

Next we investigate how branching rates of improper PCFGs change after renormalization. First, let us look at a simple example. Consider the CFG given by (1). Assign probability p to the first production ($S \to SS$), and $1-p$ to the second one ($S \to a$). It was proved that the total probability of parses is $\min(1, 1/p - 1)$. If $p > 1/2$, then $\min(1, 1/p - 1) = 1/p - 1 < 1$, implying the PCFG is improper. To get the renormalized distribution, take a parse τ with yield a^m. Since $f(S \to SS; \tau) = m - 1$ and $f(S \to a; \tau) = m$, $p(\tau) = p^{m-1}(1-p)^m$. Then the renormalized probability of τ equals

$$\tilde{p}(\tau) = \frac{p(\tau)}{1/p - 1} = \frac{p^{m-1}(1-p)^m}{1/p-1} = p^m(1-p)^{m-1}.$$

Therefore, $\tilde{p}$ is assigned by a system of production probabilities $\tilde{p}$ with $\tilde{p}(S \to SS) = 1 - p < 1/2$, and $\tilde{p}(S \to a) = p$. So the renormalized PCFG is subcritical.

More generally, we have the following result, which says a connected, improper PCFG, after being renormalized, becomes a subcritical PCFG.

Proposition 7
If p is a connected, improper PCFG distribution on parses, then its renormalized version $\tilde{p}$ is subcritical.

Proof
We have $0 < p(\Omega_S) < 1$, and we shall first show, based on the fact that the PCFG is connected, that all $0 < p(\Omega_A) < 1$. Recall the proof of Corollary 1. There we got the relation $q_S \geq q_A p_S(A \in \tau)$, where q_A is the probability that trees rooted in A fail to terminate. Because the PCFG is connected, S is reachable from A, too. By the same argument, we also have $q_A \geq q_S p_A(S \in \tau)$. Since both q_S and $p_A(S \in \tau) > 0$, $q_A > 0$, then $p(\Omega_A) = 1 - q_A < 1$. Similarly, we can prove $p(\Omega_A) \geq p(\Omega_S) p_A(S \in \tau) > 0$.

For each A, define generating functions $\{g_A\}$ as in Harris (1963, Section 2.2),

$$g_A(s) = \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} p(A \to \alpha) \prod_{B \in N} s_B^{n(B;\alpha)}, \tag{32}$$

where $s = \{s_A\}_{A \in N}$. Write $g = \{g_A\}_{A \in N}$ and define $g^{(n)} = \{g_A^{(n)}\}$ recursively as

$$\left.\begin{aligned} g_A^{(1)}(s) &= g_A(s) \\ g_A^{(n)}(s) &= g_A(g^{(n-1)}(s)) \end{aligned}\right\} \tag{33}$$

It is easy to see that $g_A(\mathbf{0})$ is the total probability of parses with root A and height 1. By induction, $g_A^{(n)}(\mathbf{0})$ is the total probability of parses with root A and height $\leq n$. Therefore, $g_A^{(n)}(\mathbf{0}) \uparrow p(\Omega_A) < 1$.

Write $r = \{p(\Omega_A)\}_{A \in N}$. Then

$$g(r) = g(\lim_{n\to\infty} g^{(n)}(\mathbf{0})) = \lim_{n\to\infty} g(g^{(n)}(\mathbf{0})) = \lim_{n\to\infty} g^{(n+1)}(\mathbf{0}) = r.$$

Therefore, r is a nonnegative solution of $g(s) = s$. It is also the smallest among such solutions. That is, if there is another nonnegative solution $r' \neq r$, then $r \leq r'$. This is because $\mathbf{0} \leq r'$ implies $g^{(n)}(\mathbf{0}) \leq g^{(n)}(r') = r'$ for all $n > 0$, and by letting $n \to \infty$, $r \leq r'$. Clearly, $\mathbf{1}$ is also a solution of $g(s) = s$.

We now renormalize p to get $\tilde{p}$ by (22). Define generating functions $f = \{f_A\}$ of $\tilde{p}$ and $f^{(n)}$ in the same way as (32) and (33). Then

$$f_A(s) \quad = \sum_{\substack{\alpha \in (N \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} \tilde{p}(A \to \alpha) \prod_{B \in N} s_B^{n(B;\alpha)}$$

$$= \sum_{\substack{\alpha\in(N\cup T)^* \\ \text{s.t. } (A\to\alpha)\in R}} \frac{1}{r_A} p(A\to\alpha) \prod_{B\in N} r_B^{n(B;\alpha)} \prod_{B\in N} s_B^{n(B;\alpha)}, \tag{34}$$

For two vectors $r = \{r_A\}$ and $\{s_A\}$, write rs for $\{r_A s_A\}$, and r/s for $\{r_A/s_A\}$. Then (34) can be written

$$f(s) = \frac{g(rs)}{r}.$$

Since all $r_A \doteq p(\Omega_A)$ are positive, $f(s)$ are well defined by the fractions.

Because r is the smallest nonnegative solution of $g(s) = s$, by the above equation, $\mathbf{1}$ is the only solution of $f(s) = s$ in the unit cube. Since $g(s) = s$ also has a solution $\mathbf{1}$, $f(s) = s$ has a solution $\mathbf{1}/r$, which is strictly larger than $\mathbf{1}$.

We want to know how f changes on the line segment connecting $\mathbf{1}$ and $\mathbf{1}/r$. Let $u = \mathbf{1}/r - \mathbf{1}$. Then u is strictly positive. Elements on the line segment between $\mathbf{1}$ and $1/r$ can be represented by $\mathbf{1} + tu$, with $t \in [0,1]$. Define $h(t) = f(\mathbf{1} + tu) - \mathbf{1} - u$. Then

$$h_A(t) = \sum_{\substack{\alpha\in(N\cup T)^* \\ \text{s.t. } (A\to\alpha)\in R}} \tilde{p}(A\to\alpha) \prod_{B\in N} (1 + tu_B)^{n(B;\alpha)} - 1 - tu_A. \tag{35}$$

Differentiate h at $t = 0$. Then $h'(0) = \mathbf{M}u - u$, where $\mathbf{M}$ is the mean matrix corresponding to $\tilde{p}$. Every $h_A(t)$ is a convex function. Then, because $h_A(0) = h_A(1) = 0$, $h'_A(0) \le 0$, which leads to $\mathbf{M}u \le u$.

We now show that for at least one A, $(\mathbf{M}u)_A < u_A$. First of all, note that $h'_A(0) = 0$ only if $h_A(t)$ is linear. Assume $\mathbf{M}u = u$, which leads to $h'(0) = \mathbf{0}$ and the linearity of $h(t)$. Together with $h(0) = \mathbf{0}$, this implies $h(t) \equiv \mathbf{0}$. Choose $t < 0$ such that $1 + tu_A > \mathbf{0}$ for all A. Then $f(\mathbf{1} + tu) - \mathbf{1} - tu = h(t) = \mathbf{0}$. Therefore $\mathbf{1} + tu$ is a nonnegative solution of $f(s) = s$ and is strictly less than $\mathbf{1}$. This contradicts the fact that $\mathbf{1}$ is the smallest nonnegative solution of $f(s) = s$.

Now we have $\mathbf{M}u \le u$, and $\exists A$, s.t. $(\mathbf{M}u)_A < u_A$. Because p is connected, $\mathbf{M}$ is irreducible. By item (3) of Theorem 1, u is strictly positive, and there is a strictly positive left eigenvector ν such that $\nu\mathbf{M} = \rho\nu$. Therefore $\nu\mathbf{M}u < \nu u$, or $\rho\nu u < \nu u$. Hence $\rho < 1$. This completes the proof. □

7. Identifiability and Approximation of Production Probabilities of PCFGs

Identifiability of parameters is related to the consistency of estimates, both being important statistical issues. Proving the consistency of the ML estimate of a system of production probabilities given in (5) is relatively straightforward. Consistency in this case means that, if p imposes a proper distribution, then as the size of the data composed of independent and identically distributed (i.i.d.) samples goes to infinity, with probability one, the estimate $\hat{p}$ converges to p. To see this, think of the sample parses as taken independently from a branching process governed by p. By the context-free nature of the branching process, for $A \in N$, each instance of A selects a production $A \to \alpha$ by probability $p(A \to \alpha)$ independently of the other instances of A. As the size of the data set goes to infinity, the number of occurrences of A goes to infinity. Therefore, by the law of large numbers, the ratio between the number of occurrences of $A \to \alpha$ and the number of occurrences of A, which is $\hat{p}(A \to \alpha)$, converges to $p(A \to \alpha)$, with probability one.

By the consistency of the ML estimate of a system of production probabilities, we can prove that production probabilities are identifiable parameters of PCFGs. In other words, different systems of production probabilities impose different PCFG distributions.

Proposition 8
If p_1, p_2 impose distributions P_1, P_2, respectively, and $p_1 \neq p_2$, then $P_1 \neq P_2$.

Proof
Assume $P_1 = P_2$. Then draw n i.i.d. samples from P_1. Because the ML estimator $\hat{p}$ is consistent, as $n \to \infty$, $\hat{p} \to p_1$, with probability 1. Because the n i.i.d. samples can also be regarded as drawn from P_2, with the same argument, $\hat{p} \to p_2$, with probability 1. Hence $p_1 = p_2$, a contradiction. □

We mentioned in Section 5 that the ML estimators (24) and (26) may produce infinitely many estimates if the Gibbs distributions on parses have the form (19). This phenomenon of multiple solutions results from the nonidentifiability of parameters of the Gibbs distributions (19), which means that different parameters may yield the same distributions.

To see why parameters of Gibbs distribution (19) are nonidentifiable, we note that the frequencies of production rules are linearly dependent,

$$\sum_{(A \to \alpha) \in R} f(A \to \alpha; \tau) \quad = \quad \sum_{(B \to \alpha) \in R} n(A; \alpha) f(B \to \alpha; \tau),$$

if $A \neq S$,

$$\sum_{(S \to \alpha) \in R} f(S \to \alpha; \tau) \quad = \quad \sum_{(B \to \alpha) \in R} n(S; \alpha) f(B \to \alpha; \tau) + 1.$$

Therefore, there exists $\lambda_0 \neq 0$, such that for any τ, $\lambda_0 \cdot f(\tau) = 0$. If $\hat{\lambda}$ is a solution for (24), then for any number t,

$$\begin{aligned} & (\hat{\lambda} + t\lambda_0) \cdot f(\tau) = \hat{\lambda} \cdot f(\tau) \\ \Longrightarrow \quad & e^{(\hat{\lambda}+t\lambda_0)\cdot f(\tau)} = e^{\hat{\lambda}\cdot f(\tau)}, Z_{\hat{\lambda}+t\lambda_0} = Z_{\hat{\lambda}} \\ \Longrightarrow \quad & P_{\hat{\lambda}+t\lambda_0}(\tau) = P_{\hat{\lambda}}(\tau). \end{aligned}$$

Thus for any t, $\hat{\lambda} + t\lambda_0$ is also a solution for (24). This shows that the parameters of Gibbs distribution (19) are nonidentifiable.

Finally, we consider how to approximate production probabilities by mean frequencies of productions. Given i.i.d. samples of parses $\tau_1, \ldots, \tau_n$ from the distribution imposed by p, by the consistency of the ML estimate of p given by (5),

$$\hat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} f(A \to \alpha; \tau_i)/n}{\sum_{i=1}^{n} f(A; \tau_i)/n} \to p(A \to \alpha),$$

with probability 1, as $n \to \infty$. If the entropy of the distribution p is finite, then for every production rule $(A \to \beta) \in R$,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} f(A \to \beta;\tau) \to E_p(f(A \to \beta;\tau)) \quad \text{with probability 1,} \\ \Longrightarrow \quad & \hat{p}(A \to \alpha) \to \frac{E_p(f(A \to \alpha;\tau))}{E_p(f(A;\tau))} \quad \text{with probability 1,} \\ \Longrightarrow \quad & p(A \to \alpha) = \frac{E_p(f(A \to \alpha;\tau))}{E_p(f(A;\tau))}. \end{aligned}$$

If the entropy is infinite, the above argument does not work, because both the numerator and the denominator of the fraction are infinity. Can we change the fraction a little bit so that it still makes sense, and at the same time yields good approximation to $p(A \to \alpha)$?

One way to do this is to pick a large finite subset Ω' of Ω and replace the fraction by

$$\frac{E_p(f(A \to \alpha;\tau)|\tau \in \Omega')}{E_p(f(A;\tau)|\tau \in \Omega')}.$$

where $E_p(f(A \to \alpha;\tau)|\tau \in \Omega')$ is the conditional expectation of $f(A \to \alpha;\tau)$ given Ω', which is defined as

$$E_p(f(A \to \alpha;\tau)|\tau \in \Omega') = \frac{\sum_{\tau\in\Omega'} f(A \to \alpha;\tau)p(\tau)}{\sum_{\tau\in\Omega'} p(\tau)}$$

Because Ω' is finite, the top of the fraction on the right-hand side is finite. Also the bottom of the fraction is positive. Therefore the conditional expectation of f is finite. The conditional expectation $E_p(f(A;\tau)|\tau \in \Omega')$ is similarly defined.

The following result shows that as Ω' expands, the approximation gets better.

Proposition 9

Suppose a system of production probabilities p imposes a proper distribution. Then for any increasing sequence of finite subsets Ω_n of Ω with $\Omega_n \uparrow \Omega$, i.e., $\Omega_1 \subset \Omega_2 \ldots \subset \Omega$, Ω_n finite and $\cup\Omega_n = \Omega$,

$$p(A \to \alpha) = \lim_{n\to\infty} \frac{E_p(f(A \to \alpha;\tau)|\tau \in \Omega_n)}{E_p(f(A;\tau)|\tau \in \Omega_n)}.$$

To prove the proposition, we introduce the Kullback-Leibler divergence. For any two probability distributions p and q on Ω, the Kullback-Leibler divergence between p and q is defined as

$$D(p\|q) = \sum_{\tau\in\Omega} p(\tau)\log\frac{p(\tau)}{q(\tau)},$$

where $0\log\frac{0}{q(\tau)}$ is defined as 0 for any $q(\tau) \geq 0$. $D(p\|q)$ is nonnegative and equal to 0 if and only if $p = q$. One thing to note is that q need not be proper in order to make $D(p\|q)$ nonnegative. Even when $\sum q(\tau) < 1$, it is still true that $D(p\|q) \geq 0$. For more about the Kullback-Leibler divergence, we refer the readers to Cover and Thomas (1991).

The Kullback-Leibler divergence has the simple property described below, which will be used in the proof of Proposition 9.

Lemma 2

If Ω' is an arbitrary subset of Ω, then

$$D(p\|q) \geq p(\Omega') \log \frac{p(\Omega')}{q(\Omega')} + (1 - p(\Omega')) \log \frac{1 - p(\Omega')}{1 - q(\Omega')}.$$

Proof

Consider the Kullback-Leibler divergence between the conditional distributions $p(\tau|\Omega')$ and $q(\tau|\Omega')$, which equals

$$\begin{aligned} &\sum_{\tau \in \Omega'} p(\tau|\Omega') \log \frac{p(\tau|\Omega')}{q(\tau|\Omega')} = \frac{1}{p(\Omega')} \sum_{\tau \in \Omega'} p(\tau) \log \frac{p(\tau)}{q(\tau)} - \log \frac{p(\Omega')}{q(\Omega')} \geq 0 \\ \Rightarrow \quad &\sum_{\tau \in \Omega'} p(\tau) \log \frac{p(\tau)}{q(\tau)} \geq p(\Omega') \log \frac{p(\Omega')}{q(\Omega')}. \end{aligned}$$

Similarly,

$$\sum_{\tau \in \Omega \setminus \Omega'} p(\tau) \log \frac{p(\tau)}{q(\tau)} \geq p(\Omega \setminus \Omega') \log \frac{p(\Omega \setminus \Omega')}{q(\Omega \setminus \Omega')} \geq (1 - p(\Omega')) \log \frac{1 - p(\Omega')}{1 - q(\Omega')}.$$

The second "$\geq$" is because $q(\Omega) \leq 1$. These two inequalities together prove the lemma. □

Proof of Proposition 9

Given n, for production probabilities q, let $K_n(q)$ be the Kullback-Leibler divergence between the conditional distribution $p(\tau|\Omega_n)$ and the distribution imposed by q,

$$K_n(q) = \sum_{\tau \in \Omega_n} p(\tau|\Omega_n) \log \frac{p(\tau|\Omega_n)}{\prod_{(A \to \alpha) \in R} q(A \to \alpha)^{f(A \to \alpha; \tau)}}. \tag{36}$$

We want to find q that minimizes $K_n(q)$. This can be achieved by applying the Lagrange multiplier method. The condition that q is subject to is,

$$\sum_{(A \to \alpha) \in R} q(A \to \alpha) = 1, \tag{37}$$

for every $A \in N$. There are $|N|$ such constraints. To incorporate them into the minimization, we consider the function

$$L(q) = K_n(q) + \sum_{A \in N} \lambda_A \left(\sum_{(A \to \alpha) \in R} q(A \to \alpha) - 1 \right),$$

where the unknown coefficients $\{\lambda_A\}_{A\in N}$ are called Lagrange multipliers.

The q that minimizes $K_n(q)$ subject to (37) satisfies

$$\frac{\partial L(q)}{\partial q(A\to\alpha)}=0,$$

for all $(A\to\alpha)\in R$. By simple computation, this is equivalent to

$$\sum_{\tau\in\Omega_n} f(A\to\alpha;\tau)P(\tau|\Omega_n)=\lambda_A q(A\to\alpha).$$

Sum both sides over all $\alpha\in(N\cup T)^*$ such that $(A\to\alpha)\in R$. By the constraints (37),

$$\lambda_A=\sum_{\tau\in\Omega_n} f(A;\tau)P(\tau|\Omega_n).$$

Therefore we prove that *if* there is a minimizer, it has to be $\hat{p}_n$, where

$$\hat{p}_n(A\to\alpha)=\frac{\sum\limits_{\tau\in\Omega_n} f(A\to\alpha;\tau)p(\tau)}{\sum\limits_{\tau\in\Omega_n} f(A;\tau)p(\tau)}.$$

To see that there *is* a minimizer of $K_n(q)$ subject to (37), consider the boundary points of the region

$$\{q=\{q(A\to\alpha)\}:q(A\to\alpha)>0,\ \sum_{\substack{\alpha\text{ s.t.}\\(A\to\alpha)\in R}} q(A\to\alpha)=1\}$$

Any boundary point of the region has a component equal to zero, hence for some $\tau\in\Omega_n$, $q(\tau)=0$, implying $K_n(q)=\infty$. Because $K_n(q)$ is a continuous function, K_n must attain its minimum inside the above region, and this minimizer, as has been shown, is $\hat{p}_n$.

We need to show $\hat{p}_n\to p$. Let $\Omega'=\Omega_n$ and apply Lemma 2 to $p(\tau|\Omega_n)$ and $\hat{p}_n(\tau)$. Since $p(\Omega_n|\Omega_n)=1$, we get $0\le-\log\hat{p}_n(\Omega_n)\le K_n(\hat{p}_n)$. On the other hand, because $\hat{p}_n$ is the minimizer of K_n, $K_n(\hat{p}_n)\le K_n(p)=-\log p(\Omega_n)$.

Because $\Omega_n\to\Omega$ and p is proper, $p(\Omega_n)\to1$. Therefore $0\le-\log\hat{p}_n(\Omega_n)\le-\log p(\Omega_n)\to0$. Hence $\hat{p}_n(\Omega_n)\to1$.

Choose an arbitrary $\tau\in\Omega$. For all n large enough, $\tau\in\Omega_n$. Apply Lemma 2 to $\{\tau\}$ and get

$$\begin{aligned}
&0\le p(\tau|\Omega_n)\log\frac{p(\tau|\Omega_n)}{\hat{p}_n(\tau|\Omega_n)}+(1-p(\tau|\Omega_n))\log\frac{1-p(\tau|\Omega_n)}{1-\hat{p}_n(\tau|\Omega_n)}\le K_n(\hat{p}_n)\to0\\
\Longrightarrow\quad & p(\tau|\Omega_n)\log\frac{p(\tau|\Omega_n)}{\hat{p}_n(\tau|\Omega_n)}+(1-p(\tau|\Omega_n))\log\frac{1-p(\tau|\Omega_n)}{1-\hat{p}_n(\tau|\Omega_n)}\to0\\
\Longrightarrow\quad & \lim_{n\to\infty}\frac{p(\tau|\Omega_n)}{\hat{p}_n(\tau|\Omega_n)}=1.
\end{aligned}$$

Together with $p(\tau|\Omega_n)\to p(\tau)>0$ and $\hat{p}_n(\Omega_n)\to1$, this implies

$$\lim_{n\to\infty}\hat{p}_n(\tau)=\lim_{n\to\infty}\hat{p}_n(\tau|\Omega_n)=\lim_{n\to\infty}p_n(\tau|\Omega_n)=p(\tau).$$

This nearly completes the proof. By the identifiability of production probabilities, $\hat{p}_n$ should converge to p. To make the argument more rigorous, by compactness of $\hat{p}_n$, every subsequence of $\hat{p}_n$ has a limit point. Let p' be a limit point of a subsequence $\hat{p}_{n_i}$. For any τ, since $\hat{p}_{n_i}(\tau) \to p(\tau)$, $p'(\tau) = p(\tau)$. By the identifiability of production probabilities, $p' = p$. Therefore p is the only limit point of $\hat{p}_n$. This proves $\hat{p}_n \to p$.

Acknowledgments

I am indebted to Stuart Geman and Mark Johnson for encouraging me to look at problems in this paper in the first place, and for various discussions and suggestions along the way. I am also grateful to the referees' careful reading of the original manuscript.

References

Abney, Steven P. 1997. Stochastic attribute-value grammars. *Computational Linguistics*, 23(4):597–618.

Baker, James K. 1979. Trainable grammars for speech recognition. In *Speech Communications Papers of the 97th Meeting of the Acoustical Society of America*, pages 547–550, Cambridge, MA.

Baum, Leonard E. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

Berger, Adam L., Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Booth, Taylor L. and Richard A. Thompson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22:442–450.

Chi, Zhiyi and Stuart Geman. 1998. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24(2):299–305.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, Inc.

Della Pietra, Stephen, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):1–13, April.

Dempster, Arthur Pentland, N. M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B*, 39:1–38.

Grenander, Ulf. 1976. *Lectures in Pattern Theory Volume 1, Pattern Synthesis*. Springer-Verlag, New York.

Harris, Theodore Edward. 1963. *The Theory of Branching Processes*. Springer-Verlag, Berlin.

Johnson, Mark. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*. To appear.

Mark, Kevin E. 1997. *Markov Random Field Models for Natural Language*. Ph.D. thesis, Department of Electrical Engineering, Washington University, May.

Mark, Kevin E., Michael I. Miller, and Ulf Grenander. 1996. Constrained stochastic language models. In S. E. Levinson and L. Shepp, editors, *Image Models (and Their Speech Model Cousins)*. Springer-Verlag, pages 131–140.

Mark, Kevin E., Michael I. Miller, Ulf Grenander, and Steven P. Abney. 1996. Parameter estimation for constrained context-free language models. In *Proceedings of the DARPA Speech and Natural Language Workshop, Image Models (and Their Speech Model Cousins)*, pages 146–149, Harriman, NY, February. Morgan Kaufmann.

Miller, Michael I. and Joseph A. O'Sullivan. 1992. Entropies and combinatorics of random branching processes and context-free languages. *IEEE Transactions on Information Theory*, 38(4), July.

Sánchez, Joan-Andreu and José-Miguel Benedí. 1997. Consistency of stochastic context-free grammars from probabilistic estimation based on growth transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1052–1055, September.

Walters, Peter. 1982. *An Introduction to Ergodic Theory*. Springer-Verlag, NY.

Zhu, Song Chun, Ying Nian Wu, and David B. Mumford. 1997. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660.