Squibs and Discussions

Estimation of Probabilistic Context-Free Grammars[1]

Zhiyi Chi*
Brown University

Stuart Geman*
Brown University

The assignment of probabilities to the productions of a context-free grammar may generate an improper distribution: the probability of all finite parse trees is less than one. The condition for proper assignment is rather subtle. Production probabilities can be estimated from parsed or unparsed sentences, and the question arises as to whether or not an estimated system is automatically proper. We show here that estimated production probabilities always yield proper distributions.

1. Introduction

Context-free grammars (CFG's) are useful because of their relatively broad coverage and because of the availability of efficient parsing algorithms. Furthermore, CFG's are readily fit with a probability distribution (to make **probabilistic** CFG's—or PCFG's), rendering them suitable for ambiguous languages through the maximum a posteriori rule of choosing the most probable parse.

For each nonterminal symbol, a (normalized) probability is placed on the set of all productions from that symbol. Unfortunately, this simple procedure runs into an unexpected complication: the language generated by the grammar may have probability less than one. The reason is that the derivation tree may have probability greater than zero of never terminating—some mass can be lost to infinity. This phenomenon is well known and well understood, and there are tests for "tightness" (by which we mean total probability mass equal to one) involving a matrix derived from the expected growth in numbers of symbols generated by the probabilistic rules (see for example Booth and Thompson [1973], Grenander [1976], and Harris [1963]).

What if the production probabilities are estimated from data? Suppose, for example, that we have a parsed corpus that we treat as a collection of (independent) samples from a grammar. It is reasonable to hope that if the trees in the sample are finite, then an estimate of production probabilities based upon the sample will produce a system that assigns probability zero to the set of infinite trees. For example, there is a simple maximum-likelihood prescription for estimating the production probabilities from a corpus of trees (see Section 2), resulting in a PCFG. Is it tight? If the corpus is unparsed then there is an iterative approach to maximum-likelihood estimation (the EM or Baum-Welsh algorithm—again, see Section 2) and the same question arises: do we get actual probabilities or do the estimated PCFG's assign some mass to infinite trees? We will show that in both cases the estimated probability is tight.[2]

* Division of Applied Mathematics, Brown University, Providence, RI 02912 USA

1 Note added in proof: An alternative proof of one of our main results (see Corollary, Section 3) recently appeared in the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Sánchez and Benedí [1997]).

2 When estimating from an unparsed corpus, we shall assume a model without null or unit productions; see Section 2.

Wetherell (1980) has asked a similar question: a scheme (different from maximum likelihood) is introduced for estimating production probabilities from an unparsed corpus, and it is conjectured that the resulting system is tight. (Wetherell and others use the designation "consistent" instead of "tight," but in statistics, consistency refers to the asymptotic correctness of an estimator.)

A trivial example is the CFG with one nonterminal and one terminal symbol, in Chomsky normal form:

$$\begin{aligned} A &\rightarrow AA \\ A &\rightarrow a \end{aligned}$$

where a is the only terminal symbol. Assign probability p to the first production ($A \rightarrow AA$) and $q = 1 - p$ to the second ($A \rightarrow a$). Let S_h be the total probability of all trees with depth less than or equal to h. For example, $S_2 = q$ corresponding to $A \rightarrow a$, and $S_3 = q + pq^2$ corresponding to $\{A \rightarrow a\} \cup \{A \rightarrow AA, A \rightarrow a, A \rightarrow a\}$. In general, $S_{h+1} = q + pS_h^2$. (Condition on the first production: with probability q the tree terminates and with probability p it produces two nonterminal symbols, each of which must now terminate with depth less than or equal to h.) It is not hard to show that S_h is nondecreasing and converges to $\min(1, \frac{q}{p})$, meaning that a proper probability is obtained if and only if $p \leq \frac{1}{2}$.

What if p is estimated from data? Given a set of finite parse trees $\omega_1, \omega_2, \ldots, \omega_n$, the maximum-likelihood estimator for p (see Section 2) is, sensibly enough, the "relative frequency" estimator

$$\hat{p} = \frac{\sum_{i=1}^{n} f(A \rightarrow AA; \omega_i)}{\sum_{i=1}^{n} [f(A \rightarrow AA; \omega_i) + f(A \rightarrow a; \omega_i)]}$$

where $f(\cdot; \omega)$ is the number of occurrences of the production "$\cdot$" in the tree ω. The sentence a^m, although ambiguous (there are multiple parses when $m > 2$), always involves $m - 1$ of the $A \rightarrow AA$ productions and m of the $A \rightarrow a$ productions. Hence $f(A \rightarrow AA; \omega_i) < f(A \rightarrow a; \omega_i)$ for each ω_i. Consequently:

$$f(A \rightarrow AA; \omega_i) < \frac{1}{2}[f(A \rightarrow AA; \omega_i) + f(A \rightarrow a; \omega_i)]$$

for each ω_i, and $\hat{p} < \frac{1}{2}$. The maximum-likelihood probability is tight.

If only the **yields** (left-to-right sequence of terminals) $Y(\omega_1), Y(\omega_2), \ldots, Y(\omega_n)$ are available, the EM algorithm can be used to iteratively "climb" the likelihood surface (see Section 2). In the simple example here, the estimator converges in one step and is the same $\hat{p}$ as if we had observed the entire parse tree for each ω_i. Thus, $\hat{p}$ is again less than $\frac{1}{2}$ and the distribution is again tight.

2. Maximum-Likelihood Estimation

More generally, let $G = (V, T, R, S)$ denote a context-free grammar with finite variable set V, start symbol $S \in V$, finite terminal set T, and finite production (or rule) set R. (We use R in place of the more typical P to avoid confusion with probabilities.) Each production in R has the form $A \rightarrow \alpha$, where $A \in V$ and $\alpha \in (V \cup T)^*$. In the usual way, probabilities are introduced through the productions: $P : R \rightarrow [0, 1]$ such that $\forall A \in V$:

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

Given a set of finite parse trees $\omega_1, \omega_2, \ldots, \omega_n$, drawn independently according to the distribution imposed by p, we wish to estimate p.

In terms of the frequency function f, introduced in Section 1, the likelihood of the data is

$$\begin{aligned} L &= L(p; \omega_1, \omega_2, \ldots, \omega_n) \\ &= \prod_{i=1}^{n} \prod_{(A \to \alpha) \in R} p(A \to \alpha)^{f(A \to \alpha; \omega_i)}. \end{aligned}$$

Recall the derivation of the maximum-likelihood estimator of p: The log of the likelihood is:

$$\sum_{A \in V} \sum_{\substack{\alpha \text{ s.t.} \\ (A \to \alpha) \in R}} \sum_{i=1}^{n} f(A \to \alpha; \omega_i) \log p(A \to \alpha). \tag{2}$$

The function $p\colon R \to [0,1]$ subject to (1) that maximizes (2) satisfies:

$$\frac{\delta}{\delta p(B \to \beta)} \sum_{A \in V} \sum_{\substack{\alpha \text{ s.t.} \\ (A \to \alpha) \in R}} \left\{ \lambda_A p(A \to \alpha) + \sum_{i=1}^{n} f(A \to \alpha; \omega_i) \log p(A \to \alpha) \right\} = 0$$

$\forall (B \to \beta) \in R$ where $\{\lambda_A\}_{A \in V}$ are Lagrange multipliers. Denote the maximum-likelihood estimator by $\hat{p}$:

$$\lambda_B + \frac{\sum_{i=1}^{n} f(B \to \beta; \omega_i)}{\hat{p}(B \to \beta)} = 0 \quad \forall (B \to \beta) \in R$$

$$\Rightarrow \left(\text{Since} \sum_{\substack{\beta \text{ s.t.} \\ (B \to \beta) \in R}} \hat{p}(B \to \beta) = 1 \right)$$

$$\hat{p}(B \to \beta) = \frac{\sum_{i=1}^{n} f(B \to \beta; \omega_i)}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in R}} \sum_{i=1}^{n} f(B \to \alpha; \omega_i)}. \tag{3}$$

The maximum-likelihood estimator is the natural, "relative frequency," estimator.

Suppose $B \in V$ is unobserved among the parse trees $\omega_1, \omega_2, \ldots, \omega_n$. Then we can assign $\hat{p}(B \to \beta)$ arbitrarily, requiring only that (1) be respected. Evidently the likelihood is unaffected by the particular assignment of $\hat{p}(B \to \beta)$. Furthermore, it is not hard to see that any such B has probability zero of arising in any derivation that is based upon the maximum-likelihood probabilities[3]—hence the issue of tightness is independent of this assignment.

We will show that if Ω is the set of all (finite) parse trees generated by G, and if $\hat{p}(\omega)$ is the probability of $\omega \in \Omega$ under the maximum-likelihood production probabilities, then $\hat{p}(\Omega) = 1$.

3 Consider any sequence of productions that leads from S to B. If the parent (antecedent) of B arose in the sample, then the last production has $\hat{p}$ probability zero and hence the sequence has probability zero. Otherwise, move "up" through the ancestors of B until finding the first variable in the S-to-B sequence represented in the sample (certainly S is represented). Apply the same reasoning to the production from that variable, and conclude that the given sequence has $\hat{p}$ probability zero.

2.1 The EM Algorithm

Usually the derivation trees are unobserved—the sample, or corpus, contains only the yields $Y(\omega_1), Y(\omega_2), \ldots, Y(\omega_n)$ ($Y(\omega_i) \in T^*$ for each $1 \leq i \leq n$). The likelihood is substantially more complex, since $p(Y(\omega))$ is now a marginal probability; we need to sum over the set of $\omega \in \Omega$ that yield $Y(\omega)$:

$$p(Y(\omega)) = \sum_{\substack{\omega' \in \Omega \\ Y(\omega')=Y(\omega)}} p(Y(\omega')).$$

In the case where only yields are observed, the treatment is complicated considerably by the possibility of null productions ($A \rightarrow \emptyset$) and unit productions ($A \rightarrow B \in V$). If, however, the language of the grammar does not include the null string, then there is an equivalent grammar (one with the same language) that has no null productions and no unit productions (cf. Hopcroft & Ullman [1979], Theorem 4.4). It is, then, perhaps best to simplify the treatment by assuming that there are no null or unit productions. Therefore, when the corpus consists of yields only, we shall assume a priori a model free of null and unit productions, and study tightness for probabilities estimated under such a model. Based upon the results of Stolcke [1995] it is likely that this restriction can be relaxed, but we have not pursued this.

Letting Ω_Y denote $\{\omega \in \Omega\colon\ Y(\omega) = Y\}$, the likelihood of the corpus becomes

$$\prod_{i=1}^{n} \sum_{\omega \in \Omega_{Y(\omega_i)}} \prod_{(A \rightarrow \alpha) \in R} p(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \omega)}.$$

And the maximum-likelihood equation becomes

$$\lambda_B + \frac{1}{\hat{p}(B \rightarrow \beta)} \sum_{i=1}^{n} \frac{\sum_{\omega \in \Omega_{Y(\omega_i)}} f(B \rightarrow \beta; \omega) \prod_{(A \rightarrow \alpha) \in R} \hat{p}(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \omega)}}{\sum_{\omega \in \Omega_{Y(\omega_i)}} \prod_{(A \rightarrow \alpha) \in R} \hat{p}(A \rightarrow \alpha)^{f(A \rightarrow \alpha; \omega)}} = 0$$

$\Rightarrow$

$$\hat{p}(B \rightarrow \beta) = \frac{\sum_{i=1}^{n} E_{\hat{p}}[f(B \rightarrow \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \rightarrow \alpha) \in R}} \sum_{i=1}^{n} E_{\hat{p}}[f(B \rightarrow \alpha; \omega) | \omega \in \Omega_{Y(\omega_i)}]} \tag{4}$$

where $E_{\hat{p}}$ is expectation under $\hat{p}$ and where "$|\omega \in \Omega_{Y(\omega_i)}$" means "conditioned on $\omega \in \Omega_{Y(\omega_i)}$."

There is no hope for a closed form solution, but (4) does suggest an iteration scheme, which, as it turns out, "climbs" the likelihood surface (though there are no guarantees about approaching a **global** maximum): Let $\hat{p}_0$ be an arbitrary assignment respecting (1). Define a sequence of probabilities, $\hat{p}_n$, by the iteration

$$\hat{p}_{n+1}(B \rightarrow \beta) = \frac{\sum_{i=1}^{n} E_{\hat{p}_n}[f(B \rightarrow \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \rightarrow \alpha) \in R}} \sum_{i=1}^{n} E_{\hat{p}_n}[f(B \rightarrow \alpha; \omega) | \omega \in \Omega_{Y(\omega_i)}]} \tag{5}$$

The right-hand side is manageable, as long as we can manageably compute all possible parses of a sentence (yield) $Y(\omega)$. (More efficient approaches exist; see Baker [1979].) This iteration procedure is an instance of the EM Algorithm. Baum [1972] first introduced it for hidden Markov models (regular grammars) and Baker [1979] extended it to the problem addressed here (estimation for context-free grammars). Dempster, Laird, and Rubin [1977] put the idea into a much more general setting and coined the

term EM for Expectation-Maximization. The right-hand side of (5) is computed using the *expected* frequencies under $\hat{p}_n$; $\hat{p}_{n+1}$ is then the *maximum*-likelihood estimator, treating the expected frequencies as though they were observed frequencies.

The issue of tightness comes up again. We will show that $\hat{p}_n(\Omega) = 1$ for each $n > 0$.

3. Tightness of the Maximum-Likelihood Estimator

Given a context-free grammar $G = (V, T, R, S)$, let Ω be the set of finite parse trees, let $p: R \to [0, 1]$ be a system of production probabilities satisfying (1), and let $\omega_1, \omega_2, \ldots, \omega_n$ be a set (sample) of finite parse trees $\omega_k \in \Omega$. For now, null and unit productions are permitted. Finally, let $\hat{p}$ be the maximum-likelihood estimator of p, as defined by (3). (See also the remarks following [3] concerning variables unobserved in $\omega_1, \omega_2, \ldots, \omega_n$.) More generally, $\hat{p}$ will refer to the probability distribution on (possibly infinite) parse trees induced by the maximum-likelihood estimator.

Theorem
$\hat{p}(\Omega) = 1$

Proof
Let $q_A = \hat{p}$ (derivation tree rooted with A fails to terminate). We will show that $q_S = 0$ (i.e., derivation trees rooted with S always terminate).

For each $A \in V$, let $F(A;\omega)$ be the number of instances of A in ω and let $\tilde{F}(A;\omega)$ be the number of nonroot instances of A in ω. Given $\alpha \in (V \cup T)^*$, let $n_A(\alpha)$ be the number of instances of A in the string α, and, finally, let α_i be the ith component of the string α. For any $A \in V$:

$$
\begin{aligned}
q_A &= \hat{p}\Big(\cup_{B\in V} \cup_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \cup_{i \text{ s.t. } \alpha_i = B} \{\alpha_i \text{ fails to terminate}\}\Big) \\
&\le \sum_{B\in V} \hat{p}\Big(\cup_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \cup_{i \text{ s.t. } \alpha_i = B} \{\alpha_i \text{ fails to terminate}\}\Big) \\
&= \sum_{B\in V} \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \hat{p}(A \to \alpha)\hat{p}\Big(\cup_{i \text{ s.t. } \alpha_i = B} \{\alpha_i \text{ fails to terminate}\} \mid A \to \alpha\Big) \\
&\le \sum_{B\in V} \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \hat{p}(A \to \alpha) n_B(\alpha) q_B \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) \sum_{i=1}^n f(A \to \alpha; \omega_i)}{\sum_{\substack{\alpha \text{ s.t.} \\ (A\to\alpha)\in R}} \sum_{i=1}^n f(A \to \alpha; \omega_i)} \right\} \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)}{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t.} \\ (A\to\alpha)\in R}} f(A \to \alpha; \omega_i)} \right\} \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)}{\sum_{i=1}^n F(A; \omega_i)} \right\} \\
\Rightarrow q_A \sum_{i=1}^n F(A;\omega_i) &\le \sum_{B\in V} q_B \sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)
\end{aligned}
$$

Sum over $A \in V$:

$$\sum_{A \in V} q_A \sum_{i=1}^{n} F(A;\omega_i) \leq \sum_{B \in V} q_B \sum_{i=1}^{n} \sum_{A \in V} \sum_{\substack{\alpha \text{ s.t. } B \in \alpha \\ (A \to \alpha) \in R}} n_B(\alpha) f(A \to \alpha;\omega_i)$$

$$= \sum_{B \in V} q_B \sum_{i=1}^{n} \tilde{F}(B;\omega_i)$$

i.e.,

$$\sum_{A \in V} q_A \sum_{i=1}^{n} (\tilde{F}(A;\omega_i) - F(A;\omega_i)) \geq 0$$

Clearly, for every $i = 1, 2, \ldots, n$ $\tilde{F}(A;\omega_i) = F(A;\omega_i)$ whenever $A \neq S$ and $\tilde{F}(S;\omega_i) < F(S;\omega_i)$. Hence $q_S = 0$, completing the proof of the theorem. □

Now let $\hat{p}_n$ be the system of probabilities produced by the nth iteration of the EM Algorithm (5):

Corollary
If R contains no null productions and no unit productions, then $\hat{p}_n(\Omega) = 1 \; \forall n \geq 1$.

Proof
Almost identical, except that we use (5) in place of (3) and end up with:

$$\sum_{A \in V} q_A \sum_{i=1}^{n} E_{\hat{p}_{n-1}}[\tilde{F}(A;\omega_i) - F(A;\omega_i) | \omega \in \Omega_{Y(\omega_i)}] \geq 0. \quad (6)$$

In the absence of unit productions and null productions, $F(A;\omega) < 2|\omega|$ (twice the length of the string ω). Hence the expectations in (6) are finite. Furthermore, $\tilde{F}(A;\omega)$ and $F(A;\omega)$ satisfy the same conditions as before: $\tilde{F}(A;\omega) = F(A;\omega)$ except when $A = S$, in which case $\tilde{F}(A;\omega) < F(A;\omega)$. Again, we conclude that $q_S = 0$. □

Acknowledgments
We are indebted to Mark Johnson for encouraging us to look at this problem in the first place, and for much good advice along the way. 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).

References

Baker, J. 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, L. E. 1972. An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

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

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

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

Harris, T. E. 1963. *The Theory of Branching Processes.* Springer-Verlag, Berlin.

Hopcroft, J. E. and J. D. Ullman. 1979. *Introduction to Automata Theory, Languages,*

and Computation. Addison Wesley.
Sánchez, J. A. and J. M. Benedí. 1997. Consistency of stochastic context-free grammars from probabilistic estimation based on growth transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1052–1055.
Stolcke, A. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21:165–201.
Wetherell, C. S. 1980. Probabilistic languages: A review and some open questions. *Computing Surveys*, 12:361–379.