Chapter 6

Probabilistic Feature Based Grammars

6.1 Introduction

Statistical language models are becoming increasingly important in linguistics. The development of such models aims to solve problems that traditional categorical grammars face. Sometimes called non-probabilistic grammars, categorical grammars provide extremely detailed syntactic and semantic analyses of a range of sentences. They also have the merit of being sensitive to a wide variety of linguistic interactions. However, categorical grammars have several drawbacks which hinder their utility. First of all, because the grammars fail to address the ranking of grammatical analyses, they suffer serious inefficiency problem when dealing with sentences which have tremendous amount of different analyses. For the same reason, they also lack robustness when coming across unexpected or ill-formed input. Furthermore, with no practical automatic learning mechanism to categorical grammars, such grammars have to be hand-crafted and usually become so complex that they are difficult or impossible to understand and maintain.

Statistical language models are probabilistic versions of categorical grammars, with all analyses allowed by the grammars being equipped with probabilities. The assignment of probability measures automatically enables statistical language models to systematically treat grammatical analyses differently. When good statistical models are established for languages, analyses empirically more likely to be chosen are allocated higher probabilities, hence more likely to be selected by parsing algorithms. Good statistical models also make it possible that analyses of ill-formed input have very low probabilities, making them easily detected by the parsing algorithms. Because of the discriminating power of probabilities, the rules by statistical models need not be as detailed and complex as categorical grammars when modeling the same languages. In addition, statistical models can be adjusted by tuning their parameters and can be learned from the training corpus because the parameters can be estimated.

The simplest statistical language models are probabilistic regular grammars (PRGs) and probabilistic context-free grammars (PCFGs). They are actually the same things as Markov chains and stochastic branching processes, respectively. Both models have had remarkable applications to simple tasks in speech recognition and computer vision (Chou [4]). However, these grammars' non-probabilistic prototypes, i.e., regular grammars (RGs) and context-free grammars (CFGs), are widely deemed linguistically inadequate, because they lack the context sensitivity that is ubiquitous in natural languages. In order to apply statistical methods more effectively to linguistics, it is necessary to develop probabilistic versions of more expressive grammars.

Standard grammars in computational linguistics are attribute-value grammars of some variety. In this article, we will call attribute-value grammars feature based grammars. RGs and CFGs are two types of feature based grammars, but among the least expressive ones. The more expressive feature based grammars cope with context sensitivity by addressing features that contain non-local information of languages. Efforts have been made to develop general probabilistic feature based grammars (Mark et al. [7], Abney [1]). Invariably, all the probabilities proposed for feature based grammars take the form of Gibbs distribution. The argument for the Gibbs form is based on the "maximum entropy" principle (Jaynes [6]). In Mark et al. [7], a Gibbs distribution was derived for a simple case, where the probabilistic models are combinations of a PCFG and n-gram language models, by invoking maximum entropy estimation. Similar argument can be applied to more general cases to get the Gibbs distributions as discussed in Abney [1]. However, this was not pursued in either of the two articles. In §6.2, we will derive the Gibbs form of distributions on features based grammars and some of its variants.

The emphasis of this article is on the technical issues of parameter estimation. In §6.3 and §6.4, we will propose two schemes for estimation. Both schemes are easy to prove to be consistent. We will argue that the second scheme, which is a pseudo-likelihood type scheme for estimation, is efficient, if the goal of parameter estimation is to analyze sentences rather than sample sentences.

6.2 Gibbs Distributions for Feature Based Grammars

Given a grammar G, let Ω be the set of all parse trees allowed by G. Elements in Ω are denoted as ω. Because a natural language has only countably many sentences, and each sentence has only finitely many parse trees allowed by G, Ω is countable. Let $f_1(\omega), \dots, f_N(\omega)$ be N real functions, or "features", on Ω. Suppose under certain unknown distribution on Ω, the expectation of $f_1(\omega), \dots, f_N(\omega)$ are $\bar{f}_1, \dots, \bar{f}_N$, respectively. With only $\bar{f}_1, \dots, \bar{f}_N$ being known, we want to make a reasonable guess about the unknown distribution.

For this end, the maximum entropy principle suggests using the solution of the following constrained maximization problem,

$$p = \underset{\tilde{p} \text{ prop on } \Omega}{\arg\max} \left\{ -\sum_{\omega \in \Omega} \tilde{p}(\omega) \log \tilde{p}(\omega) \right\},$$

subject to

$$E_p(f_i(\omega)) = \sum_{\omega \in \Omega} f_i(\omega) p(\omega) = \bar{f}_i, \quad i = 1, \dots, N \tag{6.1}$$

and

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{6.2}$$

The philosophy for the above approximation is that while $p(\omega)$ satisfies the given constraints on f_i, it should be made as random (or un-informative) as possible in other unconstrained dimensions, i.e., $p(\omega)$ should represent information no more than what is available and in this sense, the maximum entropy principle is often called the minimum prejudice principle (Zhu et al. [5]).

By introducing the Lagrange multipliers λ_i, $i = 1, \ldots, N$, and β, the constrained maximization problem is changed to

$$\frac{\partial}{\partial p(\omega)} \left\{ -\sum_{\omega \in \Omega} p(\omega) \log p(\omega) + \sum_{i=1}^{N} \lambda_i \sum_{\omega \in \Omega} f_i(\omega) p(\omega) + \beta \sum_{\omega \in \Omega} p(\omega) \right\} = 0.$$

Solving this equation gives

$$p(\omega) = \frac{1}{Z(\lambda)} e^{\lambda \cdot f(\omega)}, \tag{6.3}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$ and $f(\omega) = (f_1, \ldots, f_N)$, and $Z(\lambda) = \sum e^{\lambda \cdot f(\omega)}$.

The maximum entropy principle can be generalized to the "minimum discriminant principle" (Mark et al. [7]). Suppose we have a distribution $\pi(\omega)$ on Ω, then the minimum discriminant principle requires the guess of the unknown distribution minimize the following quantity,

$$\sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{\pi(\omega)},$$

subject to (6.1) and (6.2). Then we get the solution with the form

$$p(\omega) = \frac{1}{Z(\lambda)} \pi(\omega) e^{\lambda \cdot f(\omega)} = \frac{1}{Z(\lambda)} e^{\lambda \cdot f(\omega) + \log \pi(\omega)}, \tag{6.4}$$

which is still a Gibbs form.

If π is finite, an explanation for the constrained minimization is as follows. Write

$$\sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{\pi(\omega)} = -\log \pi(\Omega) + \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{\kappa(\omega)},$$

where $\kappa(\omega) = \pi(\omega)/\pi(\Omega)$ is a probability distribution on Ω. The minimization then finds the distribution which satisfies the constraints and is closest to the distribution κ in terms of Kullback-Leiber distance. However, this explanation does not apply to the case where $\pi(\Omega) = \infty$.

Because the set of all parses is infinite, both (6.3) and (6.4) have the possible problem that the partition number $Z(\lambda)$ might be infinity, which makes the distribution not well-defined. An alternative to the Gibbs forms (6.3) and (6.4) is the following distribution,

$$p(\omega) = \pi(Y(\omega)) \frac{e^{\lambda \cdot f(\omega)}}{\sum\limits_{Y(\omega')=Y(\omega)} e^{\lambda \cdot f(\omega')}}, \tag{6.5}$$

where $Y(\omega)$ is the "yield" of ω, which is the terminal string associated with the parse tree ω, and π is a probability distribution on the language.

From the information point of view, we can think of π as a description of the mechanism to generate sentences. It can be different from the Gibbs distribution with the potential function $\lambda \cdot f(\omega)$. On the other hand, the rules to analyze individual sentences, which are given by λ and f, with λ being the parameter, are uniform across all sentences.

The potential function $\lambda \cdot f(\omega)$ can be looked on as the first order expansion of a function $\varphi(f(\omega))$. Even when f gives all the information about Ω, i.e., the σ-algebra $\mathcal{F}(f)$ contains all the singleton sets $\{\omega\}$, the Gibbs distribution (6.3) can still be very far from the true distribution. As an example, suppose $\Omega = \mathbf{N}$ and $f(\omega)$ for $\omega \in \Omega$ is the numerical value of the element. If p is a distribution on Ω with $p(2) \gg p(\Omega \backslash \{2\})$, then the Gibbs distribution (6.3) can never get close to p.

A solution to this problem is to learn the function φ, on the set of all possible values of $f(\omega)$. To do this, one can approximate $\varphi(f(\omega))$ by a higher order expansion and estimate λ's in the expansion,

$$\sum_{i \leq k} \lambda_i (f(\omega))^i,$$

where $i = (i_1 \dots i_n)$ is a multiple index composed of non-negative integers. $i \leq k$ means $i_1 + \cdots + i_n \leq k$, and f^i means $f_1^{i_1} \cdots f_N^{i_N}$. For the example given just now, a second order expansion can do well enough in the sense that

$$\sum_{\omega} \left| \sqrt{p(\omega)} - \sqrt{p_\lambda(\omega)} \right|$$

is small, where $p_\lambda(\omega)$ is a Gibbs distribution with the potential function $\lambda_1 f(\omega) + \lambda_2 (f(\omega))^2$. It turns out that λ_1 and λ_2 should satisfy $\lambda_1 \approx -4\lambda_2$ and $\lambda_1 \gg 0$.

One can also learn $\varphi(f(\omega))$ by dividing the range of f into several bins $B_1, \dots, B_k$, and approximating $\varphi(f)$ by a function which is constant in each bin (Zhu et al. [5]). The potential function is then changed to

$$\sum_{i=1}^{k} \lambda_i \mathbf{1}_i (f(\omega)),$$

where $\mathbf{1}_i$ is the indicator function of the bin B_i.

Next we will consider how to estimate parameter λ of the Gibbs forms. From now on, we will always use N as the notation for the dimension of f.

6.3 Maximum-Likelihood (ML) Type Estimation of Parameters

Suppose we are given n i.i.d. samples. Let us consider two cases about the data.

Case One:

In this case, the n samples are fully observed parse trees $\omega_1, \ldots, \omega_n$. Under the assumption that the distribution of ω is given by (6.3) with parameter λ_0, if $|f|$ has finite mean and if n is large, then by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} f(\omega_i) \approx E_{\lambda_0}(f),$$

where $E_{\lambda_0}(f)$ is the expectation of $f(\omega)$ under the distribution $e^{\lambda_0 \cdot f(\omega)}/Z(\lambda_0)$.

Therefore, we take any solution to the following equation in λ as an estimate of λ_0,

$$E_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} f(\omega_i). \tag{6.6}$$

The estimation formulated by (6.6) is a maximum-likelihood type estimation. Indeed, *if* there is a solution to the following maximization problem,

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{\lambda \cdot f(\omega_i)}}{Z(\lambda)},$$

then the solution, $\hat{\lambda}$, is a solution to (6.6). However, because the set of all parse trees is infinite, we can not compute $Z(\lambda)$, therefore $E_\lambda(f)$ in (6.6) is unknown.

In order to get around this problem, we modify the estimation as follows. Let $Y_n = \{Y(\omega_1), \ldots Y(\omega_n)\}$. Then we replace (6.6) by

$$\frac{1}{n}\sum_{i=1}^{n} f(\omega_i) = \frac{\displaystyle\sum_{Y(\omega)\in Y_n} f(\omega) e^{\lambda \cdot f(\omega)}}{\displaystyle\sum_{Y(\omega)\in Y_n} e^{\lambda \cdot f(\omega)}} = E_\lambda[f(\omega)|Y(\omega) \in Y_n],$$

or

$$\bar{f} - E_\lambda[f(\omega)|Y(\omega) \in Y_n] = 0, \tag{6.7}$$

where $\bar{f}$ is the average of $f(\omega_1), \ldots f(\omega_n)$. It can be shown the left hand side of (6.7) is the gradient of the function

$$L_n(\lambda, \omega_1, \ldots, \omega_n) = \lambda \cdot \bar{f} - \log\left(\sum_{Y(\omega)\in Y_n} e^{\lambda \cdot f(\omega)}\right),$$

which is convex in λ. Note that since there can be multiple parse trees with the same yield, the set $\{\omega : Y(\omega) \in Y_n\}$ might be strictly larger than $\{\omega_1, \ldots, \omega_n\}$. Since $L_n(\lambda, \omega_1, \ldots, \omega_n)$ is convex in λ, any solution to the following maximization problem is a solution to (6.7), and vice versa,

$$\hat{\lambda}_n = \arg\max_{\lambda} \{L_n(\lambda, \omega_1, \ldots, \omega_n)\}. \tag{6.8}$$

In the remaining part of this section, we will use $L(\lambda)$ as short for $L_n(\lambda, \omega_1, \ldots, \omega_n)$. That the maximization problem (6.8) has a solution is not guaranteed. For example, suppose we have 3 ω's, ω_1, ω_2 and ω_3, and $f(\omega_1) = (0,0)$, $f(\omega_2) = (0,1)$, and $f(\omega_3) = (1,0)$. Suppose only ω_2 and ω_3 are observed, with each being observed once, and $Y(\omega_i)$, $i = 1,2,3$ are the same. Then $\bar{f}$ is the average of $f(\omega_2)$ and $f(\omega_3)$, i.e., $(1/2, 1/2)$, and

$$L(\lambda) = \lambda \cdot \bar{f} - \log\left(\sum_{i=1}^{3} e^{\lambda \cdot f(\omega_i)}\right) = \frac{\lambda_1 + \lambda_2}{2} - \log\left(1 + e^{\lambda_1} + e^{\lambda_2}\right).$$

The above function can not achieve its maximum. Indeed,

$$\nabla L(\lambda) = \left(\frac{1}{2} - \frac{e^{\lambda_1}}{1 + e^{\lambda_1} + e^{\lambda_2}}, \frac{1}{2} - \frac{e^{\lambda_2}}{1 + e^{\lambda_1} + e^{\lambda_2}}\right).$$

Since ∇L can never be 0, there are no extreme points for L.

In order to get the condition for the existence of solution to (6.8), let $\Omega_n = \{\omega : Y(\omega) \in Y_n\}$ and C be the convex closure of the set $\{f(\omega) : \omega \in \Omega_n\}$. The boundary of C is the union of all the facets of C and denoted as ∂C. The inner part of C is defined as $C \backslash \partial C$. Because $\bar{f}$ is the average of some of the $f(\omega)$'s with $\omega \in \Omega_n$, $\bar{f} \in C$.

Proposition 16. Suppose $f(\omega)$ are not all the same for $\omega \in \Omega_n$. Then the maximization problem (6.8) has a solution if and only if $\bar{f} \in C \backslash \partial C$.

Remark. If $f(\omega)$ are the same for all $\omega \in \Omega_n$, then $L(\lambda)$ is a constant.

Proof. What we need to show is that the function

$$L(\lambda) = \lambda \cdot \bar{f} - \log\left(\sum_{\omega \in \Omega_n} e^{\lambda \cdot f(\omega)}\right)$$

can achieve its maximum if and only if $\bar{f} \in C \backslash \partial C$.

Let k be the dimension of convex set C. Recall that N is the dimension of λ. Clearly, $k \leq N$. If $k < N$, then there is an N-dimensional vector $\beta \neq 0$ and a constant c, such that $\beta \cdot f(\omega) = c$ for all $\omega \in \Omega_n$. Without loss of generality, suppose the last component of β, $\beta_N \neq 0$. Then for all $\omega \in \Omega_n$,

$$f_N(\omega) = \frac{c}{\beta_N} - \frac{\beta_1}{\beta_N} f_1(\omega) - \cdots \frac{\beta_{N-1}}{\beta_N} f_{N-1}(\omega).$$

Then

$$L(\lambda) = \lambda' \cdot \bar{g} - \log\left(\sum_{\omega \in \Omega_n} e^{\lambda' \cdot g(\omega)}\right) \triangleq L'(\lambda'),$$

where

$$\lambda' = \left(\lambda_1 - \frac{\beta_1 \lambda_N}{\beta_N}, \ldots, \lambda_{N-1} - \frac{\beta_{N-1} \lambda_N}{\beta_N}\right),$$

is an $N-1$ dimensional vector, and

$$g(\omega) = (f_1(\omega), \ldots, f_{N-1}(\omega)) .$$

Let C' be the convex closure of $\{g(\omega) : \omega \in \Omega_n\}$. Then C' is still a k dimensional convex polygon but embedded in an $N-1$ dimensional space and $\bar{g} \in C' \backslash \partial C'$ if and only if $\bar{f} \in C \backslash \partial C$. Obviously, $L(\lambda)$ can get to its maximum if and only if $L'(\lambda')$ can. From the above procedure we see that we can reduce the dimension of λ until it equals k, without affecting the final conclusion.

In the remaining part of the proof we only consider the case where $k = N$. Let S be the $N-1$ dimensional unit sphere, which consists of all N dimensional vectors with $|v| = 1$. If $\bar{f} \in C \backslash \partial C$, then for any $v \in S$,

$$M(v) > v \cdot \bar{f},$$

where

$$M(v) = \max_{\omega \in \Omega_n} \{v \cdot f(\omega)\} .$$

The function $M(v) - v \cdot \bar{f}$ is continuous, therefore, by compactness of S, there is a constant $A > 0$ such that $M(v) - v \cdot \bar{f} > A$ for all $v \in S$.

For each $v \in S$, taking $L(tv)$ as a function in t, we have

$$L'(tv) = v \cdot \bar{f} - \frac{\displaystyle\sum_{\omega \in \Omega_n} v \cdot f(\omega) e^{tv \cdot f(\omega)}}{\displaystyle\sum_{\omega \in \Omega_n} e^{tv \cdot f(\omega)}} \to v \cdot \bar{f} - M(v) < -A, \quad t \to \infty,$$

and $L''(tv) < 0$. For each $t > 0$, let $B_t = \{v \in S : L'(tv) < 0\}$. From the above result we see $S \subset \cup_{t>0} B_t$. Because of the continuity of $L'(tv)$ in v, B_t is open. Because S is compact, $S \subset \cup B_{t_i}$ for some $t_1 < t_2 < \ldots < t_m$. For each $v \in S$, because $L(tv)$, when taken as a function in t, is concave, therefore, if $L'(t_0 v) < 0$, then for any $t > t_0$, $L'(tv) < 0$, which means $B_{t_i} \subset B_{t_m}$. This implies that for all $v \in S$ and $t > t_m$, $L'(tv) < 0$. Thus if $|\lambda| > t_m$, then $L(\lambda) < L(t_m v)$, where $v = \lambda/|\lambda|$. Therefore $L(\lambda)$ must get its maximum in the region $\{\lambda : |\lambda| \leq t_m\}$.

Conversely, assume $L(\lambda)$ achieves its maximum at some λ_0, then $\nabla L(\lambda_0) = 0$, and therefore

$$\bar{f} = \frac{\displaystyle\sum_{\omega \in \Omega_n} f(\omega) e^{\lambda_0 \cdot f(\omega)}}{\displaystyle\sum_{\omega \in \Omega_n} e^{\lambda_0 \cdot f(\omega)}} .$$

If the vertices of C are $v_1, \ldots, v_p$, then every $f(\omega)$ can be written as $a_1 v_1 + \ldots a_p v_p$, where $a_i \geq 0$ and $a_1 + \ldots a_p = 1$. Because each $e^{\lambda_0 \cdot f(\omega)} > 0$, from the above equality, $\bar{f} = b_1 v_1 + \ldots b_p v_p$ with each b_i being positive. Hence $\bar{f} \in C \backslash \partial C$. □

As mentioned earlier, (6.8) may not have a solution. One way to handle this problem is to modify the maximum-likelihood estimation (6.8) to the following form,

$$\hat{\lambda}_n = \arg \max_{|\lambda| \leq n} \{L_n(\lambda, \omega_1, \ldots, \omega_n)\} . \tag{6.9}$$

Case Two:

In this case, only the yields of the parse trees are observed. Let $y_1, \ldots, y_n$ be the n sentences and let $Y_n = \{y_1, \ldots, y_n\}$. Under the assumption that the distribution of ω is given by (6.3) with parameter λ_0, if $|f|$ has finite mean and if n is large, then by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda_0}[f(\omega)|Y(\omega) = y_i] \approx \sum_{y\in Y} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y) = \sum_{\omega\in\Omega} E_{\lambda_0}(f(\omega)),$$

where $P_{\lambda_0}(y)$ is the sum of all $P_{\lambda_0}(\omega)$ with $Y(\omega) = y$, and $Y = \{Y(\omega) : \omega \in \Omega\}$. On the other hand, as n is large enough,

$$\sum_{y\in Y_n} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y|Y_n) \approx \sum_{y\in Y} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y),$$

hence

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda_0}[f(\omega)|Y(\omega) = y_i] \approx \sum_{y\in Y_n} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y|Y_n).$$

With a similar argument as in the first case, we take any solution to the following equation as an estimate of λ_0,

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda}[f(\omega)|Y(\omega) = y_i] - \sum_{y\in Y} E_{\lambda}[f(\omega)|Y(\omega) = y]P_{\lambda}(y|Y_n) = 0. \tag{6.10}$$

To transform (6.10) into an optimization problem, define the log-likelihood function in λ,

$$L_n(\lambda, y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} \log P_{\lambda}(y_i|Y_n).$$

Then

$$\nabla_{\lambda} L_n(\lambda, y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} E_{\lambda}[f(\omega)|Y(\omega) = y_i] - \sum_{y\in Y} E_{\lambda}[f(\omega)|Y(\omega) = y]P_{\lambda}(y|Y_n).$$

Therefore, any maximizer of $L_n(\lambda, y_1, \ldots, y_n)$ is a solution to (6.10).

A condition that $L_n(\lambda, y_1, \ldots, y_n)$ can reach its maximum is as follows. As in the first case, suppose the dimension of λ is N.

Proposition 17. Given a convex set C in R^N, a plane P is called a support of C if $\emptyset \neq P \cap C \subset \partial C$. For each sentence y, let $C(y)$ be the convex closure of the set $\{f(\omega) : y(\omega) = y\}$.

If there is no such a plane P that it is a common support of $C(y_1), \ldots, C(y_n)$ and all $C(y)$'s are on the same side of P, then L_n achieves its maximum.

Proof. As in the proof of Proposition 16, let S be the unit sphere in $\mathbf{R}^N$. By the assumption, for any $v \in S$,

$$\max_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} > \min_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\}.$$

The functions on both sides of the above inequality are continuous in v. Because S is compact, there is a constant $\delta > 0$, such that for any $v \in S$,

$$\max_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} - \min_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} > \delta$$

Based on this, using the same argument as Proposition 16, we can show that L_n achieves its maximum. □

If $L_n(\lambda, y_1, \ldots, y_n)$ can achieve its maximum, then the maximizers of the function are solutions to (6.10). However, unlike $L_n(\lambda, \omega_1, \ldots, \omega_n)$ in case one, $L_n(\lambda, y_1, \ldots, y_n)$ is not necessarily convex. Unless $y_1, \ldots y_n$ satisfy the condition of Proposition 17, $L_n(\lambda, y_1, \ldots, y_n)$ might not be able to achieve its maximum. As an alternative, we take

$$\hat{\lambda}_n = \arg \max_{|\lambda|<n} L_n(\lambda, y_1, \ldots, y_n), \tag{6.11}$$

as the estimate of λ_0.

For the estimation given by (6.11), we have the following consistency result.

Proposition 18. Assume the distribution on Ω is given by

$$P_{\lambda_0}(\omega) = \frac{e^{\lambda_0 \cdot f(\omega)}}{Z_{\lambda_0}}.$$

Suppose $\omega_1, \ldots, \omega_n$ are i.i.d. samples from P_{λ_0}. Let $y_i = Y(\omega_i)$, $i = 1, \ldots, n$, and $Y_n = \{y_1, \ldots, y_n\}$. Define $\Omega_n = \{\omega \in \Omega : Y(\omega) \in Y_n\}$. Let $\hat{\lambda}_n$ be the estimates given by (6.9) or (6.11). Define the distribution P_n on Ω such that

$$P_n(\omega) = \begin{cases} \dfrac{e^{\hat{\lambda}_n \cdot f(\omega)}}{\sum\limits_{\omega' \in \Omega_n} e^{\hat{\lambda}_n \cdot f(\omega')}} & \text{if } \omega \in \Omega_n \\ 0 & \text{otherwise} \end{cases}$$

(1) If $\hat{\lambda}_n$ are given by (6.9) and if $H = -\sum\limits_{\omega \in \Omega} P_{\lambda_0}(\omega) \log P_{\lambda_0}(\omega) < \infty$, then with probability 1, as $n \to \infty$, P_n weakly converges to P_{λ_0} on Ω, i.e.,

$$P_n(\omega) \to P_{\lambda_0}(\omega), \quad \text{for any } \omega \in \Omega.$$

(2) If $\hat{\lambda}_n$ are given by (6.11) and if $H = -\sum\limits_{y \in Y} P_{\lambda_0}(y) \log P_{\lambda_0}(y) < \infty$, then with probability 1, as $n \to \infty$, P_n weakly converge to P_{λ_0} on Y, i.e.,

$$P_n(y) \to P_{\lambda_0}(y), \quad \text{for any } y \in Y.$$

Proof. We only prove (2). The proof of (1) is very similar to the proof of (2).

Write $L_n(\lambda)$ for $L_n(\lambda, y_1, \ldots, y_n)$. For any integer $n > |\lambda_0|$, by (6.11), $L_n(\lambda_n) \geq L_n(\lambda_0)$. But

$$L_n(\lambda_0) = \frac{1}{n}\sum_{i=1}^{n} \log P_{\lambda_0}(y_i) + \log Z(\lambda_0) - \log\left(\sum_{\omega\in\Omega_n} e^{\lambda_0\cdot f(\omega)}\right).$$

With probability 1, $L_n(\lambda_0) \to H$, hence

$$\liminf \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \geq H.$$

Let $I_n(y)$ denote the empirical probability of y, i.e.,

$$I_n(y) = \frac{|\{i:\ y_i = y\}|}{n}.$$

Then

$$\begin{aligned} L_n(\lambda_n) &= \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \\ &= \sum_{y\in Y_n} I_n(y) \log P_n(y) \\ &\leq \sum_{y\in Y_n} I_n(y) \log I_n(y). \qquad (6.12) \end{aligned}$$

Fix $\epsilon > 0$, there is a finite $Y' \subset Y$, such that

$$\sum_{y\in Y'} P_{\lambda_0}(y) \log P_{\lambda_0}(y) \leq \sum_{y\in Y} P_{\lambda_0}(y) \log P_{\lambda_0}(y) + \epsilon. \qquad (6.13)$$

With probability 1, when n is large enough, $Y_n \supset Y'$, then

$$\sum_{y\in Y_n} I_n(y) \log I_n(y) \leq \sum_{y\in Y'} I_n(y) \log I_n(y). \qquad (6.14)$$

Letting $n \to \infty$, with probability 1,

$$\sum_{y\in Y'} I_n(y) \log I_n(y) \to \sum_{y\in Y'} P_{\lambda_0}(y) \log P_{\lambda_0}(y). \qquad (6.15)$$

By (6.12)-(6.15),

$$\limsup \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \leq H.$$

Therefore,

$$\lim \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) = H.$$

The above arguments also show that

$$\lim \frac{1}{n} \sum_{i=1}^{n} \log I_n(y_i) = H.$$

Then for large n,

$$\sum_{y \in Y_n} I_n(y) \log P_n(y) \geq \sum_{y \in Y_n} I_n(y) \log I_n(y) - \epsilon.$$

Since $\{P_n\}$ is a sequence of probability measures on the countable set Y, it contains convergent subsequences. Let $\tilde{P}$ be the limit of a convergent subsequence $\{P_{n_i}\}$. Then $\tilde{P}$ is a measure on Y with $\sum \tilde{P}(y) \leq 1$. From

$$\sum_{y \in Y_{n_i}} I_{n_i}(y) \log P_{n_i}(y) \geq \sum_{y \in Y_{n_i}} I_{n_i}(y) \log I_{n_i}(y) - \epsilon,$$

we get

$$\sum_{y \in Y} P_{\lambda_0}(y) \log \tilde{P}(y) \geq \sum_{y \in Y} P_{\lambda_0}(y) \log P_{\lambda_0}(y),$$

which can happen only if $\tilde{P} = P_{\lambda_0}$. Therefore any convergent subsequence of $\{P_n\}$ converges to P_{λ_0}. Therefore $P_n \to P_{\lambda_0}$. □

Corollary 5. If λ_0 is identifiable, i.e., for any $\lambda \neq \lambda_0$, $P_\lambda \neq P_{\lambda_0}$, then the estimation (6.9) is consistent, which means with probability 1, $\hat{\lambda}_n \to \lambda_0$ as $n \to \infty$. □

6.4 Pseudo-Likelihood (PL) Type Estimation of Parameters

The estimation procedures given in §6.3 are basically of maximum-likelihood type. They estimate the "global" distribution, i.e., the distribution on the set of all parse trees or the distribution on the set of all sentences. In the context of parsing, however, global distributions are irrelevant. What is really relevant for efficient parsing is that, given a sentence, all the possible parses of the sentence are properly assigned *conditional probabilities* so that the correct parses to the sentence are preferred in the sense that they have higher conditional probabilities. This observation suggests using the pseudo-likelihood (PL) type procedure for parameter estimation (Besag [2], [3]).

The idea for the PL estimation is as follows. Let (Ω, P) be a space. Suppose Ω is partitioned into disjoint subsets Ω_α. Then for each $\omega \in \Omega$, there is a unique Ω_α, denoted as $\Omega(\omega)$ such that $\omega \in \Omega(\omega)$. If we are given a parametric family of probability distributions $\{P_\theta\}_{\theta \in \Theta}$ and $P = P_{\theta_0}$, then for i.i.d. samples $\omega_1, \ldots, \omega_N$ from P, the PL estimate for θ_0 is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{N} \{P_\theta(\omega_i | \Omega(\omega_i))\}.$$

Now let Ω be the set of all parses. In the context of parsing, we are interested in the comparison of all the parses for each single sentence, but not the comparison of parses for different sentences. Therefore, the partition we choose is such that, for $\omega \in \Omega$,

$$\Omega(\omega) = \{\omega' \in \Omega : Y(\omega') = Y(\omega)\}.$$

If $Y(\omega) = y$, then clearly, for any distribution P on Ω,

$$P(\omega|\Omega(\omega)) = P(\omega|Y(\omega) = y) = \frac{P(\omega)}{\sum\limits_{Y(\omega')=y} P(\omega')}$$

The global distribution of sentences is irrelevant for parsing, and we assume it to be $\pi(y)$, which might be unknown. The conditional probability distribution of parses, given a sentence y, is assumed to be a Gibbs distribution. In certain sense, the Gibbs distribution of parses, given y, should depend on y, i.e.,

$$P(\omega|Y(\omega) = y) = \frac{e^{\lambda_y \cdot f_y(\omega)}}{\sum\limits_{Y(\omega')=y} e^{\lambda_y \cdot f_y(\omega')}},$$

where λ_y are parameters depending on y and f_y are features depending on y. However, it is reasonable to assume that across all the sentences, the parsing rules are the same. Therefore, we suppose the conditional distributions have the same λ and the same f, for all y.

The distribution of all the parses then takes the form given by (6.5). Given i.i.d. samples $\omega_1, \ldots, \omega_n$, let $y_i = Y(\omega_i)$. The PL estimate is

$$\hat{\lambda}_n = \arg\max_{\lambda} \left\{ \prod_{i=1}^{n} P_\lambda(\omega_i|\Omega(\omega_i)) \right\} = \arg\max_{\lambda} \left\{ \prod_{i=1}^{n} \frac{e^{\lambda \cdot f(\omega)}}{\sum\limits_{Y(\omega)=y_i} e^{\lambda \cdot f(\omega)}} \right\},$$

or, using the notion of log-likelihood,

$$\hat{\lambda}_n = \arg\max_{\lambda} \left\{ \lambda \cdot \bar{f} - \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{Y(\omega)=y_i} e^{\lambda \cdot f(\omega)} \right) \right\}. \tag{6.16}$$

Let $PL(\lambda, \omega_1, \ldots, \omega_n)$ be the function being maximized on the right hand side of (6.16). If the maximization has a solution $\hat{\lambda}_n$, then $\nabla PL(\lambda_0, \omega_1, \ldots, \omega_n) = 0$, i.e.,

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} E_\lambda \left[f(\omega) | Y(\omega) = y_i \right]. \tag{6.17}$$

The formula (6.17) has an explanation which has nothing to do with the Gibbs form. If $|f(\omega)|$ has finite mean, then by the law of large numbers, as $n \to \infty$, with probability one, $\bar{f} \to E(f)$. On the other hand, for any sentence y,

$$\frac{|\{i : \ y_i = y\}|}{n} \to \pi(y),$$

and therefore,

$$\frac{1}{n}\sum_{i=1}^{n} E\left[f(\omega)|Y(\omega)=y_i\right] = \sum_{y\in Y}\frac{|\{i:\ y_i=y\}|}{n}E\left[f(\omega)|Y(\omega)=y_i\right] \to E(f).$$

In the above formula we omit the subscript of E_λ to make it clear that the distribution considered here is not necessarily given by (6.5).

If the true distribution belongs to a parametric family $\{P_\theta\}$ and its parameter is θ_0, then as n is large,

$$\bar{f} \approx \frac{1}{n}\sum_{i=1}^{n} E_{\theta_0}\left[f(\omega)|Y(\omega)=y_i\right],$$

and it is reasonable to let (any) solution of

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} E_{\theta}\left[f(\omega)|Y(\omega)=y_i\right]$$

be an estimate of θ_0.

The estimation given by (6.16) is consistent in the following sense.

Proposition 19. Let $\{P_\lambda\}$ be a parametric family of probability distributions on Ω, such that for each λ,

$$P_\lambda(\omega) = \pi(Y(\omega))\frac{e^{\lambda\cdot f(\omega)}}{\sum\limits_{\omega'\in\Omega(\omega)} e^{\lambda\cdot f(\omega')}}.$$

Assume $\omega_1,\ldots,\omega_n$ are i.i.d. samples from P_{λ_0}. Let $\hat{\lambda}_n$ be the estimates given by (6.16). For each n, let $P_n = P_{\hat{\lambda}_n}$. If

$$-\sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|Y(\omega)=y)\log P_{\lambda_0}(\omega|Y(\omega)=y) < \infty,$$

then with probability 1, for each sentence y, and for each ω with $Y(\omega)=y$,

$$P_n(\omega|Y(\omega)=y) \to P_{\lambda_0}(\omega|Y(\omega)=y),$$

Proof. With the similar arguments as in Proposition 18, it can be shown that with probability 1,

$$\frac{1}{n}\sum_{i=1}^{n}\log P_n(\omega_i|\Omega(\omega_i)) \to \sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|\Omega(\omega))\log P_{\lambda_0}(\omega|\Omega(\omega))$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log I_n(\omega_i|\Omega(\omega_i)) \to \sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|\Omega(\omega))\log P_{\lambda_0}(\omega|\Omega(\omega)).$$

But

$$\frac{1}{n}\sum_{i=1}^{n}\log P_n(\omega_i|\Omega(\omega_i)) = \sum_{y\in Y} I_n(y) \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)),$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log I_n(\omega_i|\Omega(\omega_i)) = \sum_{y\in Y} I_n(y) \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)).$$

For each y, since $I_n(y) \to \pi(y)$ and

$$\sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)) \leq \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)),$$

we conclude that

$$\sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)) - \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)) \to 0.$$

Now since the set $\{\omega :\ y(\omega) = y\}$ is finite, we get

$$P_n(\omega|\Omega(\omega)) \to P_{\lambda_0}(\omega|\Omega(\omega)).$$

The proof is complete. □

Corollary 6. If for each $\lambda \neq \lambda_0$, there is a y and an ω with $Y(\omega) = y$, such that $P_\lambda(\omega|Y(\omega) = y) \neq P_{\lambda_0}(\omega|Y(\omega) = y)$, then with probability one, $\hat{\lambda}_n \to \lambda_0$. □

Bibliography

[1] S. P. Abney. Stochastic Attribute-Value Grammars. *Computational Linguistics*. Accepted for publication.

[2] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion). *J. Roy. Statist. Soc. Ser. B* 36, 192-236. 1974.

[3] J. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician* 24, 179-195. 1975.

[4] P. A. Chou. Recognition of Equations Using a Two-Dimensional Stochastic Context-Free Grammar. *Visual Communications and Image Processing IV*. SPIE – The International Society for Optical Engineering. November, 1989.

[5] M. Johnson. NSF Grant Proposal. Department of Cognitive and Linguistic Sciences, Brown University. 1998.

[6] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review* 106, 620-630. 1957.

[7] K. E. Mark. Markov Random Field Models for Natural Languages. PhD thesis. Department of Electrical Engineering, Washington University. May, 1997.

[8] S. C. Zhu, Y. N. Wu, and D. B. Mumford. Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation* 9, 1627-1660. 1997.