Inside-Outside Estimation Meets Dynamic EM — GOLD —

Detlef Prescher
DFKI Language Technology Lab
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
prescher@dfki.de

Abstract

It is an interesting fact that most of the stochastic models used by linguists can be interpreted as probabilistic context-free grammars (Prescher 2001). In this paper, this result will be accompanied by the formal proof that the inside-outside algorithm, the standard training method for probabilistic context-free grammars, can be regarded as a dynamic-programming variant of the EM algorithm. Even if this result is considered in isolation this means that most of the probabilistic models used by linguists are trained by a version of the EM algorithm. However, this result is even more interesting when considered in a theoretical context because the well-known convergence behavior of the inside-outside algorithm has been confirmed by many experiments but it seems that it never has been formally proved. Furthermore, being a version of the EM algorithm, the inside-outside algorithm also inherits the good convergence behavior of EM. We therefore contend that the as yet imperfect line of argumentation can be transformed into a coherent proof.

1 Introduction

The inside-outside algorithm is the standard training method for probabilistic context-free grammars and can effecteviley be applied to many NLP tasks. Carroll and Rooth (1998) present head-lexicalized probabilistic context-free grammar formalisms for English, as well as Beil et al. (1999) for German. Recently, Müller (2001) uses probabilistic context-free grammars for syllabification and grapheme-to-phoneme conversion. Prescher (2001) shows that many of the stochastic models used by linguists can be interpreted as probabilistic context-free grammars.
In literature it is often alleged that Baker (1979) or Lari and Young (1990) proved that probabilistic grammars when trained by the inside-outside algorithm are characterized by monotonously increasing log-likelihood values (corpus probabilities respectively) which converge towards a local maximum of the log-likelihood function. However, Baker has only intuitively generalized the forward-backward algorithm (Baum 1972), whereas Baum explicitly stated the forward-backward algorithm for hidden Markov models. Furthermore, Baker unfortunately studied

merely training corpora consisting of one single sentence. Lari and Young (1991) eliminated this weak point and generalized the inside-outside algorithm in order to apply it to any training corpus. However, this study also lacks a formal proof of monotonicity and convergence. It is worth noting that Lari and Young pointed out that certain counts of grammar rules used as in the inside-outside algorithm can be interpreted as expected frequencies. Unfortunately, the mathematical principles of EM theory (Dempster et al. 1977) were not applied to formally proof this intuitively conceived observation. To the best of our knowledge, this is the first time that the common wisdom saying the inside-outside algorithm is a dynamic-programming variant of the EM algorithm, is being formally proved. Moreover, being a dynamic-programming variant of the EM algorithm, the inside-outside algorithm inherits its good convergence behavior.
The paper is organized as follows. In Section 2 and 3 we briefly review the inside-outside and the EM algorithm, whereas in Section 4, inside-outside meets EM. In Section 5, we conclude.

2 Inside-Outside Estimation

The modern inside-outside algorithm was introduced by Lari and Young (1990) who reviewed an algorithm proposed by Baker (1979) and extended it to an iterative training method for probabilistic context-free grammars enabling the use of unrestricted free text. The re-estimation formulas proposed by Lari and Young are the core of many statistical parsers and are build upon so-called inside and outside probabilities. In contrast to widely used tree-bank training (Charniak 1996) or to partially-bracketed corpora training (Pereira and Schabes 1992), pure inside-outside estimation does not rely on manually annotated (thus relatively small) corpora.

Inside and outside probabilities

Following the lines of Lari and Young (1990), we define the inside probability $e(s,t,A)$ as the probability of the non-terminal symbol A generating the observation $w_s \dots w_t$ (see left-hand side of Figure 1), i.e. $e(s,t,A) := p(A \Rightarrow^* w_s \dots w_t)$. In determining a recursive procedure for calculating e, two cases must be considered for a grammar in Chomsky normal form:

- $(s = t)$: Only one observation is emitted and therefore a rule of the form $A \to w_s$ applies: $e(s,s,A) = \begin{cases} p(A \to w_s) & \text{if } (A \to w_s) \in G \\ 0 & \text{else} \end{cases}$
- $s < t$: In this case we know that rules of the form $A \to BC$ must apply since more than one observation is involved. Refering to the right-hand side of Figure 1, it is clear that $e(s,t,A)$ can be expressed as follows:

$$e(s,t,A) = \sum_{(A \to BC) \in G} \sum_{r=s}^{t-1} p\,(A \to BC) \cdot e(s,r,B) \cdot e(r+1,t,C) \ .$$

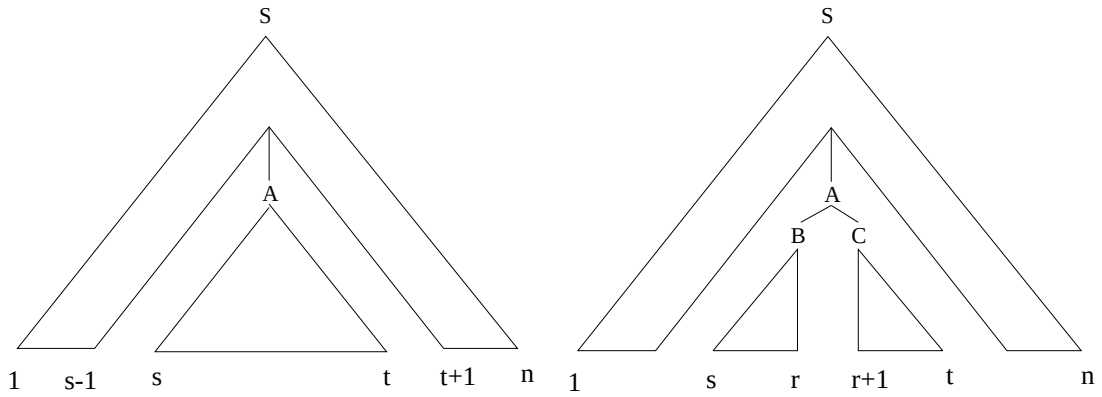

Figure 1: Definition of inside and outside probs (left), and calculation of inside probs (right)

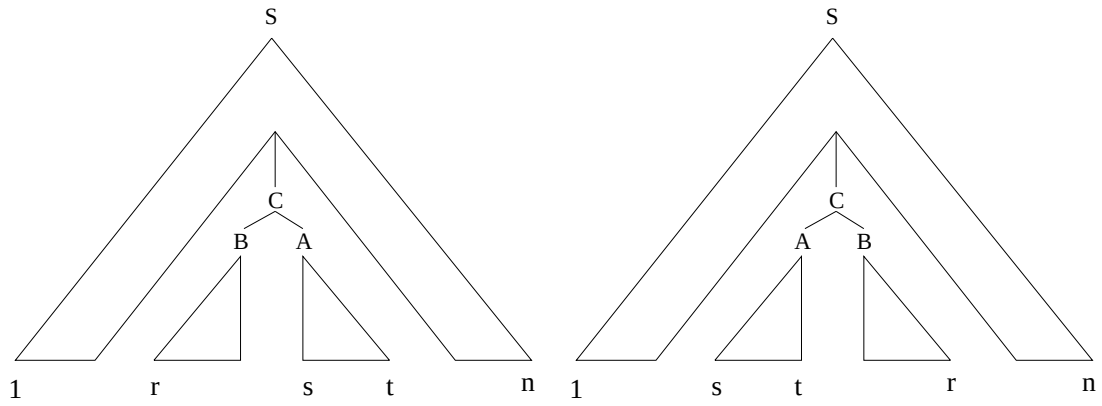

Figure 2: Calculation of outside probabilities

The quantity e can therefore be computed recursively by determining e for all sequences of length 1, then all sequences of length 2, and so on. Next, we define the outside probabilities as follows: $f(s,t,A) = p\left(S \Rightarrow^* w_1 \dots w_{s-1} A w_{t+1} \dots w_n\right)$. The quantity $f(s,t,A)$ may be thought of as the probability that A is generated in the re-write process and that the strings not dominated by it are $w_1 \dots w_{s-1}$ to the left and $w_{t+1} \dots w_n$ to the right (see left-hand side of Figure 1). In this case, the non-terminal A could be one of two possible settings $C \rightarrow B\ A$ or $C \rightarrow A\ B$ as shown in Figure 2, hence:

$$f(s,t,A) = \sum_{B,\ C \in G} \left(\sum_{r=1}^{s-1} f(r,t,C) \cdot p(C \rightarrow BA) \cdot e(r, s-1, B) + \sum_{r=t+1}^{n} f(s,r,C) \cdot p(C \rightarrow AB) \cdot e(t+1, r, B) \right)$$

and $f(s,t,A) = \begin{cases} 1 & \text{if } A = S \\ 0 & \text{else} \end{cases}$. After the inside probabilities have been computed bottom-up, the outside probabilities can therefore be computed top-down.

The inside-outside algorithm as introduced by Baker (1979)

It is well-known that Baker (1979) introduced the first training procedure for probabilistic context-free grammars, whereas it is less noticed that Baker presented his re-estimation formulas for a special case (of no relevance from the modern point of view), namely for a training corpus consisting of one single sentence. Bakers goal was to obtain a procedure for automatically training a stochastic grammar allowing an arbitrary degree of ambiguity. The procedure for estimating the parameters for a finite-state, hidden Markov process was well-established (Baum 1972) and Baker obtained the new re-estimation formulas as a result of an intuitive generalization process taking Baums formulas as a starting point. The forward-backward algorithm formalizes stochastic training of hidden Markov processes and uses particulary a training corpus of words. A simple generalization of this algorithm, not changing the training corpus itself, obviously yields re-estimation formulas which again rely on a sequence of words:

$$\hat{p}(A \to a) := \frac{\frac{1}{P} \sum_{1 \le t \le n,\ w_t = a} e(t,t,A) \cdot f(t,t,A)}{\frac{1}{P} \sum_{s=1}^{n} \sum_{t=s}^{n} e(s,t,A) \cdot f(s,t,A)} \ ,$$

and

$$\hat{p}(A \to BC) := \frac{\frac{1}{P} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} \sum_{r=s}^{t-1} p(A \to BC) e(s,r,B) e(r+1,t,C) f(s,t,A)}{\frac{1}{P} \sum_{s=1}^{n} \sum_{t=s}^{n} e(s,t,A) f(s,t,A)} \ .$$

The sentence probability $P := p(S \Rightarrow^* w_1 \ldots w_n)$ and the inside and outside probabilities can be computed in a pre-processing procedure using the current rule probabilities. It is interesting that Baker did not state any convergence behaviour of the sequence of re-estimated grammars. Presumably, Baker expected (correctly) that his generalization of the re-estimation formulas for hidden Markov models transferred the convergence properties of these models to probabilistic context-free grammars. However, Baker, as well as other researchers incorrectly assumed that there is no need to formally prove this result. First, training corpora usually consist of several thousand sentences not just of one single sentence. Therefore the common wisdom that the log-likelihood of probabilistic context-free grammars increases while training with inside-outside estimation on large corpora can not have its basis on Bakers work. Second, the formal proof of the convergence behaviour we will give in Section 4 is relatively complex and premises a good knowledge about both EM and dynamic programming methods. Thus, this proof is not so easy to be taken for granted.

The inside-outside algorithm as introduced by Lari and Young (1990)

The inside-outside algorithm as introduced by Baker (1979) was extended by Lari and Young (1990) to a training method on multiple observations, because in practice, a single observation is insufficient to accurately estimate the parameters of a probabilistic context-free grammar. The key step is the introduction of so-called *rule and category counts*. Using

$$C_w(A) \ := \ \frac{1}{P} \sum_{s=1}^{n} \sum_{t=s}^{n} e(s,t,A) \cdot f(s,t,A),$$

and

$$C_w(A \to a) \;:=\; \frac{1}{P} \sum_{1 \le t \le n,\ w_t = a} e(t,t,A) \cdot f(t,t,A)$$

as well as

$$C_w(A \to BC) \;:=\; \frac{1}{P} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} \sum_{r=s}^{t-1} p(A \to BC) e(s,r,B) e(r+1,t,C) f(s,t,A)$$

the re-estimation formulas of Baker have the following simple form:

$$\hat{p}(A \to a) \;=\; \frac{C_w(A \to a)}{C_w(A)}, \qquad \hat{p}(A \to BC) \;=\; \frac{C_w(A \to BC)}{C_w(A)}.$$

The idea of Lari and Young (1990), how to extend these formulas, is simple: compute the rule and category counts for each sentence of the given training corpus, sum and normalize them to get rule probabilties:

$$\hat{p}(A \to a) := \frac{\sum_{w=y_1}^{y_N} C_w(A \to a)}{\sum_{w=y_1}^{y_N} C_w(A)}, \text{ and} \qquad \hat{p}(A \to BC) := \frac{\sum_{w=y_1}^{y_N} C_w(A \to BC)}{\sum_{w=y_1}^{y_N} C_w(A)}\ .$$

Here, $y_1 \ldots y_N$ are the sentences of the training corpus and $C_w(A \to a),\ C_w(A \to BC),\ C_w(A)$ are computed for each sentence w with the current rule probabilities. Given these non-trivial re-estimation formulas which are applicable to training corpora of arbitrary size, it is astonishing that Lari and Young (1990) gave no formal convergence proofs at all. Possibly, experimental results made them very sure that the new re-estimation formulas increase the log-likelihood.

3 EM for Probabilistic Context-Free Grammars

In this Section, we present a brief review of both the EM algorithm and probabilistic context-free grammars, followed by the application of EM to probabilistic context-free grammars. This will yield the well-known result that a re-estimated rule probability can be computed by normalizing the expected rule frequency with the expected frequency of its mother category, where the expectations have to be computed on the *current corpus of complete data types.*

A Brief Review of the EM algorithm

Figure 3 displays the EM algorithm (implemented in pseudo code). The input consists of the *empirical distribution*[1] $\tilde{p}(y) := \frac{f(y)}{\sum_{y \in \mathcal{Y}} f(y)}$ of so-called *incomplete data types* $y \in \mathcal{Y}$. Obviously, this definition relies on the frequencies $f(y)$ gathered from a given training corpus of incomplete data types. Here, the term "incomplete" refers to a given *symbolic analysis component* which maps each incomplete data type y to a subset $X(y) \subseteq \mathcal{X}$ of so-called *complete data types*, i.e. the analyses of y, such that the space of all complete data types equals the disjoint

[1] Throughout this paper, the term "distribution" refers to a discrete probability distribution, i.e. a non-negative function summing up to 1.

1. **for each** $\Theta_0 \in \Omega_0$ **do** * Version with explicit variation of starting points *\
2. **for each** $i := 1, \ldots,$ *numberOfIterations* **do**
3. $\Phi := \Theta_{i-1}$;
4. **(E-step) compute** $Q_\Phi(\Theta) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\ p_\Theta(x)$;
5. **(M-step) compute** $\hat{\Theta} = \text{argmax}_{\Theta \in \Omega}\ \ Q_\Phi(\Theta)$;
6. $\Theta_i := \hat{\Theta}$;
7. **print** $\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots$;

Figure 3: EM algorithm

union of all analyses, i.e. $\mathcal{X} = \sum_{y \in \mathcal{Y}} X(y)$. Further, so-called *parameterized distributions* $p_\Theta(x)$ of complete data types $x \in \mathcal{X}$ induce parameterized distributions $p_\Theta(y) := \sum_{x \in X(y)} p_\Theta(x)$ of incomplete data types $y \in \mathcal{Y}$ (using the properties of the symbolic analysis component), as well as parameterized *conditional distributions of analyses* $p_\Theta(x|y) := \frac{p_\Theta(x)}{p_\Theta(y)} = \frac{p_\Theta(x)}{\sum_{x \in X(y)} p_\Theta(x)}$ (given $x \in X(y),\ p_\Theta(y) \neq 0$). Further, a finite set $\Omega_0 \subseteq \Omega$ of starting points (out of the set Ω of all possible parameters), and a fixed number of iterations *numberOfIterations* ≥ 1 belong to the input. The EM algorithm was designed to maximize the log likelihood

$$L(\Theta) := \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log p_\Theta(y) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log \sum_{x \in X(y)} p_\Theta(x)$$

of the training data and EM tries to do this job by performing an iterative sequence of *E and M steps*. The *E step* computes the *current auxiliary function* (given the parameter Φ)

$$Q_\Phi(\Theta) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\ p_\Theta(x)\ ,$$

whereas the *M step* tries to maximize this function: $\hat{\Theta} = \text{argmax}_{\Theta \in \Omega}\ \ Q_\Phi(\Theta)$. It is an interesting question, whether EM really does its job. Dempster et al. (1977) introduced the EM algorithm and showed that the log likelihood increases monotonically. Thus, the output of the EM algorithm is a sequence of re-estimated parameters: $\Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots \in \Omega$, such that the associated sequence of log likelihood values of the training corpus is bounded, and monotonically increasing: $L(\Theta_1) \leq L(\Theta_2) \leq L(\Theta_3) \leq \ldots \in [L(\Theta_0),\ 0]$. Additionally, Dempster et al. (1977) showed that the limit point of a convergent parameter sequence (each parameter being a stationary point of its auxiliary function) is a stationary point of the log likelihood. More interesting results were presented by (Wu 1983), who showed that both the parameter sequence and the associated sequence of log likelihood values converge, if some weak conditions are fulfilled. Additionally, Wu investigated the conditions for convergence to local maxima. It is interesting

4.	**generate** corpus $\tilde{p}_\Phi(x)$ of complete data (given parameter Φ);
5.	**compute MLE** $\hat{\Theta}$ of models $p_\Theta(x)$ on corpus $\tilde{p}_\Phi(x)$;

Figure 4: Modified version of the EM algorithm (as a sequence of MLE steps)

that the EM algorithm equals a sequence of certain maximum likelihood estimation (MLE) steps. For this pupose, we define the *current complete empirical distribution*:

$$\tilde{p}_\Phi(x) := \tilde{p}(y) \cdot p_\Phi(x|y) \qquad \text{with } y \in \mathcal{Y} \text{ such that } x \in X(y) \ .$$

Note, that this setting is well-defined, because the incomplete data type y is unique (using the properties of the symbolic anlaysis component). Obviously, the current complete empirical distribution can be regarded as a corpus of complete data types. This corpus is finite (i.e. behaves like an ordinary corpus), if each incomplete data type has a finite set of analyses. Thus, the auxiliary function $Q_\Phi(\Theta)$ equals

$$L_\Phi(\Theta) := \sum_{x \in \mathcal{X}} \tilde{p}_\Phi(x) \cdot \log p_\Theta(x) \ ,$$

which is indeed a log-likelihood function, namely on the "current corpus" $\tilde{p}_\Phi$ of complete data types. We will use the presented definitions as a key step in the following sections. For now, we note that the procedure of the EM algorithm (see Figure 3) can be easily rewritten resulting in a sequence of MLE steps performed on complete data associated with $\tilde{p}_\Phi$ (see Figure 4).

Instantiation of the EM algorithm for probabilistic context-free grammars

We wish to introduce the following notions for the symbolic component of a probabilistic context-free grammar (see Hopcroft and Ullman (1979) for a more precise presentation):

- $A \in G$ denotes a category A of the grammar G, $r \in G$ denotes a rule r of G, $\text{lhs}(r)$ is the left-hand side of the rule $r \in G$, $r \in G_A$ is a rule $r \in G$ with $\text{lhs}(r) = A$, $x \in \mathcal{T}(G)$ denotes a syntax tree x of G, $r \in x$ is a rule $r \in G$ occuring in $x \in \mathcal{T}(G)$, $y \in \mathcal{L}(G)$ denotes a sentence of G, $x \in \mathcal{T}(y)$ is a syntax tree of the sentence $y \in \mathcal{L}(G)$.

Additionally, we use the following terms for the stochastic component of a context-free grammar:

- $p(r)$ denotes the probability of a rule $r \in G$, $f_r(x)$ is the frequency of rule $r \in G$ occuring in the syntax tree $x \in \mathcal{T}(G)$, $f_A(x)$ is the frequency of category $A \in G$ occuring in the syntax tree $x \in \mathcal{T}(G)$, i.e. $f_A(x) := \sum_{r \in G_A} f_r(x)$, the probability of a syntax tree $x \in \mathcal{T}(G)$ is defined as $p(x) := \prod_{r \in x} p(r)^{f_r(x)}$, the probability of a sentence $y \in \mathcal{L}(G)$ is defined as $p(y) := \sum_{x \in \mathcal{T}(y)} p(x)$.

The standard probability model assumes that $\sum_{r \in G_A} p(r) = 1$ for all (productive and reachable) categories $A \in G$. Using the theorem of Booth and Thompson (1973) and a further presumption it can be shown that G is consistent: $\sum_{y \in \mathcal{L}(G)} p(y) = 1$. Moreover, Chi (1999) showed that probabilistic context-free grammars estimated via tree-bank training or pure inside-outside algorithm (e.g. without internal smoothing procedures) are always consistent. Using the introduced notation, the EM algorithm will be defined for a probabilistic context-free grammar as follows:

- $\mathcal{Y} = \mathcal{L}(G)$, the incomplete data types are the grammatical sentences of a given context-free grammar G. The empirical distribution $\tilde{p}(.)$ of incomplete data types will be computed using the grammatical sentences of a large training corpus of free text. The complete data types $\mathcal{X} = \mathcal{T}(G)$ are the syntax trees of G. The symbolic analysis component $X(y) = \mathcal{T}(y)$ is a context-free parser which produces for each grammatical sentence its analyses (i.e. its syntax trees) and for each ungrammatical sentence the empty set. The parameter space $\Omega = \{\ \Theta \in [0,1]^{|G|} \mid g_A(\Theta) = 0\ \}$ is constrained, where $|G|$ is the number of grammar rules. The parameterization $\Theta_r = p(r)$ yields the following constraints: $g_A(\Theta) = 1 - \sum_{r \in G_A} \Theta_r$ $(A \in G)$. The set Ω_0 of starting points will be randomly selected from the parameter space. The parametrization $\Theta_r = p(r)$ yields the following parameterized distributions of complete data types $p_\Theta(x) = \prod_{r \in G} \Theta_r^{f_r(x)}$.

Thus, the following well-known propostion is valid within this framework.
Lemma: The EM algorithm for probabilistic context-free grammars yields the following simple re-estimation formulas:

$$\hat{\Theta}_r = \frac{\tilde{p}_\Phi [\ f_r\ ]}{\tilde{p}_\Phi [\ f_A\ ]} \qquad (r \in G,\ A = \mathrm{lhs}(r)) \ .$$

Here, $\tilde{p}_\Phi [\ f_r\ ]$ and $\tilde{p}_\Phi [\ f_A\ ]$ denote the expected frequency of rule r, respectively the expected frequency of category A given the current corpus $\tilde{p}_\Phi$ of complete data types, i.e. $\tilde{p}_\Phi [\ f_r\ ] := \sum_{x \in \mathcal{X}} \tilde{p}_\Phi(x) \cdot f_r(x)$, and $\tilde{p}_\Phi [\ f_A\ ] := \sum_{x \in \mathcal{X}} \tilde{p}_\Phi(x) \cdot f_A(x)$.
Proof: Using expectations, the current log likelihood of the corpus $\tilde{p}_\Phi$ can be rewritten as $L_\Phi(\Theta) = \tilde{p}_\Phi [\ \log p_\Theta\ ]$. The stationary points of this function can be calculated using the first partial derivatives[2]. We compute these exploiting the "linear properties" of the expectation:

$$\partial_r L_\Phi(\Theta) = \partial_r \tilde{p}_\Phi [\ \log p_\Theta\ ] = \tilde{p}_\Phi [\ \partial_r \log p_\Theta\ ] = \tilde{p}_\Phi \left[\ \frac{\partial_r p_\Theta}{p_\Theta}\ \right]$$

Thus, it follows

$$\partial_r L_\Phi(\Theta) = \tilde{p}_\Phi \left[\ \frac{f_r}{\Theta_r}\ \right] = \frac{1}{\Theta_r} \cdot \tilde{p}_\Phi [\ f_r\ ] \ .$$

Unfortunately, the parameters $\Theta \in [0,1]^{|G|}$ are restricted by constraints

$$g_A(\Theta) = 0, \quad (A \in G) \ .$$

[2]Throughout this proof, we use the following notation: $\partial_r := \frac{\partial}{\partial \Theta_r}$

Thus, we introduce Lagrangian multipliers $\lambda_A \in \mathbb{R}$ yielding

$$\partial_r L_\Phi(\Theta)|_{\Theta=\hat{\Theta}} \; + \; \sum_{A \in G} \lambda_A \cdot \partial_r g_A(\Theta)|_{\Theta=\hat{\Theta}} \; = \; 0 \qquad (r \in G) \; .$$

This equation system can be simplified using $\partial_r g_A(\Theta) = \left\{ \begin{array}{ll} -1 & \text{if } r \in G_A \\ 0 & \text{else} \end{array} \right.$.

Thus

$$\frac{1}{\hat{\Theta}_r} \cdot \tilde{p}_\Phi [\; f_r \;] \; - \; \lambda_A \; = \; 0 \;\; (r \in G, \; A = \text{lhs}(r))$$

which yields the solution:

$$\hat{\Theta}_r \; = \; \frac{\tilde{p}_\Phi [\; f_r \;]}{\lambda_A} \; (r \in G, \; A = \text{lhs}(r)) \; .$$

As a last step, we add up the solutions for all $r \in G_A$ and respect the associated constraint:

$$\lambda_A = \sum_{r \in G_A} \tilde{p}_\Phi [\; f_r \;] = \tilde{p}_\Phi \left[\; \sum_{r \in G_A} f_r \; \right] = \tilde{p}_\Phi [\; f_A \;] \; .$$

Combining the last two findings, we have our proposition. **q.e.d.**

4 Inside-Outside as Dynamic EM

In this Section, the well-known convergence properties of the inside-outside algorithm, which have been unfortunately omitted in the original literature (Baker (1979), Lari and Young (1990)), will be formally proven. For this purpose, we will show that the inside-outside algorithm is a dynamic-programming variant of the EM algorithm for context-free grammars. This property is also well-known in stochastic linguistics, but to the best of our knowledege all mentioned properties have not been formally proven till now. Moreover, the exact proof is more complicated than generally expected (Manning and Schütze 1999) and some insight into both the inside-outside algorithm and the EM algorithm is necessary.

Theorem: Let G be a context-free grammar in Chomsky normal form. Let $\hat{p}(r)$ the re-estimated probabilities resulting from one single step of the inside-outside algorithm using the current rule probabilities $p(r)$. Then the following propositions are valid: (i) The log likelihood $L(.)$ of the training corpus increases monotonically, i.e. $L(\hat{p}) \geq L(p)$. (ii) The limit points of a sequence of re-estimated probabilities are stationary points (i.e. maxima, minima or saddle points) of the log likelihood function. (iii) *The inside-outside algorithm is a dynamic-programming variant of the EM algorithm*, i.e. $\hat{p}(r)$ corresponds with $\hat{p}_{EM}(r)$ resulting from one single EM iteration.

Proof: (i) and (ii) follow using both (iii) and the convergence properties of the EM algorithm for context-free grammars. (iii): For each grammar rule $r \in G$ with left-hand side $A = \text{lhs}(r)$, we have: $\hat{p}_{EM}(r) = \frac{\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_r(x)}{\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_A(x)}$. If the empirical distribution $\tilde{p}(.)$ is given by the sequence $< y_1, \ldots, y_N >$ of sentences $y_i \in \mathcal{Y}$, it follows $\tilde{p}(y) = N^{-1} \cdot f(y)$ and:

$$\hat{p}_{EM}(r) = \frac{\sum_{y \in \mathcal{Y}} f(y) \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_r(x)}{\sum_{y \in \mathcal{Y}} f(y) \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_A(x)} = \frac{\sum_{y=y_1}^{y_N} \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_r(x)}{\sum_{y=y_1}^{y_N} \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_A(x)} \; .$$

Comparing these formulas with the re-estimation formulas presented by Lari and Young (1990), it follows $\hat{p}_{EM}(r) = \hat{p}(r)$, if for each sentence $y \in \{y_1 \dots y_N\}$, for each grammar rule $r \in G$ and each grammar category $A \in G$ the following propositions can be shown:

$$C_y(r) = \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_r(x), \text{ and } \quad C_y(A) = \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_A(x) \ .$$

This is the goal of the rest of the proof, which we split in two lemmas. The first lemma is probably due to Charniak, who at least used corresponding formulas to present the inside-outside algorithm in his famous book on statistical NLP (Charniak 1993). The lemma says that the category counts can be computed by summing the rule counts of all rules with the same left-hand side. Unfortunately, an explicit proof of this proposition was not given by Charniak (as well as a reference to the work of Lari and Young (1990)).

Lemma: $C_y(A) = \sum_{r \in G_A} C_y(r)$ for each sentence y and category A.

Proof: Assuming Chomsky normal form, and $y = w_1 \dots w_n$:

$$\begin{aligned}
\sum_{r \in G_A} C_y(r) &= \sum_a C_y(A \to a) \ + \sum_{B,C \in G} C_y(A \to B\ C) \\
&= \sum_a \frac{1}{P} \sum_{1 \le t \le n,\ w_t = a} e(t,t,A)\ f(t,t,A) \\
&\quad + \sum_{B,C \in G} \frac{1}{P} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} \sum_{r=s}^{t-1} p(A \to BC) e(s,r,B) e(r+1,t,C) f(s,t,A) \\
&= \frac{1}{P} \left(\sum_{1 \le t \le n} e(t,t,A)\ f(t,t,A) \right. \\
&\quad \left. + \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} f(s,t,A) \sum_{B,C \in G} \sum_{r=s}^{t-1} p(A \to BC) e(s,r,B) e(r+1,t,C) \right) \\
&= \frac{1}{P} \left(\sum_{1 \le t \le n} e(t,t,A)\ f(t,t,A) \ + \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} f(s,t,A)\ e(s,t,A) \right) \\
&= \frac{1}{P} \sum_{1 \le s \le t \le n} e(s,t,A)\ f(s,t,A) \ = \ C_y(A) \ .
\end{aligned}$$

In the fourth equation, we used the recursion formula of the inside probabilities.
q.e.d.

The lemma shows that the wanted identities for the category counts can be calculated (by summation over all rules with the same left-hand side) using the identities for the rule counts, because $C_y(A) = \sum_{A \to \alpha} C_y(A \to \alpha)$, and per definition $f_A(x) = \sum_{A \to \alpha} f_{A \to \alpha}(x)$. Thus, the proof of the theorem is finished, as soon as the following lemma has been proven. It says that the grammar counts of the inside-outside algorithm can be identified (not only intuitively but also formally) with the expected rule frequencies of the EM algorithm.

Lemma: For each sentence y and each grammar rule $r \in G$:

$$C_y(r) \ = \ \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_r(x) \ = \ p(.|y)\,[\ f_r\ ] \ .$$

Proof: The second equation is the definition of the expectation. Assuming Chomsky normal form, we make a fall differentiation:

First case: The grammar rule has the form $A \to B\ C$

For a given sentence $y = w_1 \ldots w_n$ and given numbers $1 \leq s \leq r < t \leq n$ we define

$$\begin{aligned} X_{(s,t,A)(s,r,B)(r+1,t,C)} \ := \ S \ &\Rightarrow^* \ w_1 \ldots w_{s-1}\ A\ w_{t+1} \ldots w_n \\ &\Rightarrow \ w_1 \ldots w_{s-1}\ B\ C\ w_{t+1} \ldots w_n \\ &\Rightarrow^* \ w_1 \ldots w_r\ C\ w_{t+1} \ldots w_n \\ &\Rightarrow^* \ w_1 \ldots w_n\ . \end{aligned}$$

This means that $X_{(s,t,A)(s,r,B)(r+1,t,C)}$ is the set of all syntax trees corresponding to the given derivation. The right-hand side of Figure 1 displays the three relevant *spans* (s, r, B), $(r+1, t, C)$, and (s, t, A). Let

$$f_{(s,t,A)(s,r,B)(r+1,t,C)}(x) \ := \ \begin{cases} 1 & \text{if} \quad x \in X_{(s,t,A)(s,r,B)(r+1,t,C)} \\ 0 & \text{else} \end{cases}$$

the *characteristic function* interpreting the parse forest $X_{(s,t,A)(s,r,B)(r+1,t,C)}$ as a simple subset of the set of all possible syntax trees $\mathcal{T}(y)$ of the sentence y. Thus, the frequency $f_{A \to BC}(x)$ of the rule $A \to B\ C$ occurring in the syntax tree $x \in \mathcal{T}(y)$ can be computed as follows:

$$f_{A \to BC}(x) \ = \ \sum_{1 \leq s \leq r < t \leq n} f_{(s,t,A)(s,r,B)(r+1,t,C)}(x)\ .$$

Using the linear properties of the expectation, it follows:

$$\begin{aligned} p(.|y)\,[\ f_{A \to BC}\ ] \ &= \ p(.|y) \left[\sum_{1 \leq s \leq r < t \leq n} f_{(s,t,A)(s,r,B)(r+1,t,C)} \right] \\ &= \ \sum_{1 \leq s \leq r < t \leq n} p(.|y) \left[\ f_{(s,t,A)(s,r,B)(r+1,t,C)}\ \right] \\ &= \ \sum_{1 \leq s \leq r < t \leq n} \ \sum_{x \in \mathcal{T}(y)} p(x|y) \cdot f_{(s,t,A)(s,r,B)(r+1,t,C)}(x) \\ &= \ \frac{1}{p(y)} \sum_{1 \leq s \leq r < t \leq n} \ \sum_{x \in \mathcal{T}(y)} p(x) \cdot f_{(s,t,A)(s,r,B)(r+1,t,C)}(x) \\ &= \ \frac{1}{p(y)} \sum_{1 \leq s \leq r < t \leq n} \ \sum_{x \in X_{(s,t,A)(s,r,B)(r+1,t,C)}} p(x) \\ &= \ \frac{1}{p(y)} \sum_{1 \leq s \leq r < t \leq n} p(X_{(s,t,A)(s,r,B)(r+1,t,C)}) \\ &= \ \frac{1}{P} \sum_{1 \leq s \leq r < t \leq n} f(s,t,A) \cdot p(A \to BC) \cdot e(s,r,B) \cdot e(r+1,t,C) \\ &= \ C_y(A \to B\ C)\ . \end{aligned}$$

Second case: The grammar rule has the form $A \to a$

Analogously to the first case, we define for a given sentence $y = w_1 \ldots w_n$ and a given span (t, t, A) a parse forest together with its characteristic function ($1 \leq t \leq n$):

$$\begin{aligned} X_{(t,t,A)} \;:=\; S \;&\Rightarrow^*\; w_1 \ldots w_{t-1}\; A\; w_{t+1} \ldots w_n \\ &\Rightarrow\; w_1 \ldots w_n \end{aligned}$$

and

$$f_{(t,t,A)}(x) \;:=\; \begin{cases} 1 & \text{if} \quad x \in X_{(t,t,A)} \\ 0 & \text{else} \end{cases}$$

Thus, the frequency $f_{A\to a}(x)$ of the rule $A \to a$ occurring in the syntax tree $x \in \mathcal{T}(y)$ can be computed as follows:

$$f_{A\to a}(x) \;=\; \sum_{1\leq t\leq n,\; w_t=a} f_{(t,t,A)}(x)\ .$$

Therefore:

$$\begin{aligned} p(.|y)\,[\,f_{A\to a}\,] \;&=\; p(.|y) \left[\sum_{1\leq t\leq n,\; w_t=a} f_{(t,t,A)} \right] \\ &=\; \sum_{1\leq t\leq n,\; w_t=a} p(.|y) \left[\, f_{(t,t,A)} \,\right] \\ &=\; \sum_{1\leq t\leq n,\; w_t=a} \sum_{x\in\mathcal{T}(y)} p(x|y) \cdot f_{(t,t,A)}(x) \\ &=\; \frac{1}{p(y)} \sum_{1\leq t\leq n,\; w_t=a} \sum_{x\in\mathcal{T}(y)} p(x) \cdot f_{(t,t,A)}(x) \\ &=\; \frac{1}{p(y)} \sum_{1\leq t\leq n,\; w_t=a} \sum_{x\in X_{(t,t,A)}} p(x) \\ &=\; \frac{1}{p(y)} \sum_{1\leq t\leq n,\; w_t=a} p(X_{(t,t,A)}) \\ &=\; \frac{1}{P} \sum_{1\leq t\leq n,\; w_t=a} e(t,t,A) \cdot f(t,t,A) \\ &=\; C_y(A \to a) \qquad\qquad \textbf{q.e.d.} \end{aligned}$$

Reviewing the complete proof of the theorem, which is fairly straight-forward, the key step is to identify the rule and category counts of the inside-outside algorithm with the corresponding conditional expected frequencies used by EM. In consequence of the first lemma, the desired identities for the grammar categories arise as once as the identities for the grammar rules have been proven. These identities are the subject of the important second lemma. Here, the key steps are: (i) to express the frequency for a grammar rule as a sum over corresponding spans, (ii) to use the linear properties of the expectation in order to compute the expectation of this sum as a sum of expectations, and (iii) to identify the expectations inside the sum as probabilities of certain parse forests which can be easily calculated in terms of inside and outside probabilities.

5 Conclusion

We have presented a brief review of the inside-outside algorithm for context-free grammars which follows the lines of the original work of Baker (1979) and Lari and Young (1990). Baker presented the inside-outside algorithm for a training corpus of size 1, whereas Lari and Young introduced the inside-outside algorithm in its modern form. Unfortunately, no convergence proofs were given by these pioneers, which is probably due to the fact, that Baker intuitively generalized the forward-backward algorithm for hidden Markov models (Baum 1972), where an explicit proof of the convergence properties exists.
Furthermore, we briefly reviewed the standard EM algorithm. We presented an implementation in pseudo code, discussed its input, as well as its output focussing on the well-known convergence behaviour described by Dempster et al. (1977) and Wu (1983). As a nice side effect, we presented the EM algorithm as a sequence of maximum likelihood estimations on so-called current corpora of complete data types. We applied the EM algorithm to probabilistic context-free grammars and derived the EM-based re-estimation formulas for rule probabilities confirming the well-known fact that re-estimated rule probabilities rely on expected rule and category frequencies.
As the central goal of this paper, we have shown that the re-estimation formulas given by inside-outside can be transformed to EM re-estimation formulas. Even if this result is considered in isolation it is very interesting, because it seems to have never been formally proven. Moreover, it has the desired effect that the well-known convergence behaviour of the inside-outside algorithm, which has been confirmed by many experiments, is a consequence of the convergence behaviour of the EM algorithm, and has now been formally proven.

References

Baker, J. K. (1979). Trainable grammars for speech recognition. In D. Klatt and J. Wolf (Eds.), *Speech Communication Papers for ASA'97*, pp. 547–550.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities III*, 1–8.

Beil, F., G. Carroll, D. Prescher, S. Riezler, and M. Rooth (1999). Inside-outside estimation of a lexicalized PCFG for German. In *Proceedings of ACL'99*, College Park, MD.

Booth, T. L. and R. A. Thompson (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers C-22*(5), 442–450.

Carroll, G. and M. Rooth (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of EMNLP-3*, Granada.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: M.I.T. Press.

Charniak, E. (1996). Tree-bank grammars. Technical Report CS-96-02, Brown University.

Chi, Z. (1999). Statistical properties of probabilistic context-free grammars. *Computational Linguistics 25*(1).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statist. Soc. 39*(B), 1–38.

Hopcroft, J. E. and J. D. Ullman (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Lari, K. and S. J. Young (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language 4*, 35–56.

Lari, K. and S. J. Young (1991). Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language 5*, 237–257.

Manning, C. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Müller, K. (2001). Probabilistic context-free grammars for syllabification and grapheme-to-phoneme conversion. In *Proceedings of EMNLP'01*, Pittsburgh.

Pereira, F. and Y. Schabes (1992). Inside-outside reestimation from partially bracketed corpora. In *Proceedings of ACL'92*, Newark, Delaware.

Prescher, D. (to appear 2001). *EM-basierte maschinelle Lernverfahren für natürliche Sprachen*. Ph. D. thesis, IMS, University of Stuttgart.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics 11*(1), 95–103.