INSIDE-OUTSIDE ESTIMATION MEETS DYNAMIC EM

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

Abstract

We briefly review the inside-outside and EM algorithm for probabilistic context-free grammars. As a result, we formally prove that inside-outside estimation is a dynamic-programming variant of EM. This is interesting in its own right, but even more when considered in a theoretical context since the well-known convergence behavior of inside-outside estimation has been confirmed by many experiments but apparently has never been formally proved. However, being a version of EM, inside-outside estimation also inherits the good convergence behavior of EM. Therefore, the as yet imperfect line of argumentation can be transformed into a coherent proof.

1 Inside-Outside Estimation

The modern inside-outside algorithm was introduced by [4] who reviewed an algorithm proposed by [1] and extended it to an iterative training method for probabilistic context-free grammars enabling the use of unrestricted free text. In the following, \( y_1 \ldots y_N \) are numbered (but unannotated) sentences.

Definition: Inside-outside re-estimation formulas for probabilistic context-free grammars in Chomsky normal form are given by (see [4], but see also [1] for the special case \( N = 1 \)):

\[
p(A \rightarrow a) := \frac{\sum_{w=y_1}^{y_n} C_w(A \rightarrow a)}{\sum_{w=y_1}^{y_n} C_w(A)}, \quad p(A \rightarrow BC) := \frac{\sum_{w=y_1}^{y_n} C_w(A \rightarrow BC)}{\sum_{w=y_1}^{y_n} C_w(A)}.
\]

The key variables of this definition are so-called category and rule counts: \( C_w(A) := \frac{1}{n} \sum_{s=1}^{n} \sum_{t=s}^{n} e(s, t, A) \cdot f(s, t, A) \), \( C_w(A \rightarrow a) := \frac{1}{n} \sum_{1 \leq t \leq n, w_i = a} e(t, t, A) \cdot f(t, t, A) \), and \( C_w(A \rightarrow BC) := \frac{1}{n} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} \sum_{r=s}^{t-1} p(A \rightarrow BC)e(s, r, B)e(r+1, t, C)f(s, t, A) \) which are computed for each sentence \( w := w_1 \ldots w_n \) with so-called inside and outside probabilities: An inside probability is defined as the probability of category \( A \) generating observations \( w_s \ldots w_t \), i.e. \( e(s, t, A) := p(A \Rightarrow^* w_s \ldots w_t) \). In determining a recursive procedure for calculating \( e \), two cases must be considered:

- \((s = t)\): Only one observation is emitted and therefore a rule of the form \( A \rightarrow w_s \) applies: \( e(s, s, A) = p(A \rightarrow w_s) \), if \( A \rightarrow w_s \in G \) (and 0, otherwise).
- \((s < t)\): In this case we know that rules of the form \( A \Rightarrow BC \) must apply since more than one observation is involved. Thus, \( e(s, t, A) \) can be expressed as follows: \( e(s, t, A) = \sum_{(A \Rightarrow BC) \in G} \sum_{r=s}^{t-1} p(A \rightarrow BC) \cdot e(s, r, B) \cdot e(r+1, t, C) \).

The quantity \( e \) can therefore be computed recursively by determining \( e \) for all sequences of length 1, then 2, and so on. The sentence probability \( P := p(S \Rightarrow^* w) \) is a special inside probability. The outside probabilities are defined as follows: \( f(s, t, A) = p(S \Rightarrow^* w_1 \ldots w_{s-1}Aw_{s+1} \ldots w_n) \).


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 \( w_1 \ldots w_{s-1} \) to the left and \( w_{t+1} \ldots w_n \) to the right. In this case, the non-terminal \( A \) could be one of two possible settings \( C \rightarrow BA \) or \( C \rightarrow AB \), 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. Unfortunately, no convergence proofs of inside-outside estimation were given by [1] and [4].

## 2 EM for Probabilistic Context-Free Grammars

The EM algorithm was introduced by [3] as iterative maximum likelihood estimation for parameterized probability models \( p(y) \) using a sample \( \tilde{p}(y) \) of incomplete data types \( y \) which are defined via a symbolic analyzer \( X(y) \) dealing with complete data types \( x \). It is known, that EM generalizes ordinary maximum likelihood estimation and monotonically increases the log-likelihood \( L(p) := \sum_y \tilde{p}(y) \cdot \log \sum_{x \in X(y)} p(x) \). Furthermore, the limit point of a convergent parameter sequence is a stationary point (i.e. local minimum, saddle point or maximum) of the log likelihood [3]. Moreover, both the parameter sequence and the associated sequence of log likelihood values converge (in some cases to local maxima), if some weak conditions are fulfilled [3].

Applying EM to probabilistic context-free grammars, the grammatical sentences \( y \) are viewed as incomplete and their syntax trees \( x \) as complete. The required symbolic analyzer is given by a parser computing all trees \( x \in T(y) \) for a sentence \( y \). Via these non-probabilistic EM components, the probability model for the sentence is defined as \( p(y) := \sum_{x \in T(y)} p(x) := \sum_{x \in T(y)} \prod_r p(r)^{f_r(x)} \), where \( f_r(x) \) is the frequency of rule \( r \) occurring in \( x \), and parameterization is given by rule probabilities \( p(r) \). The key variables of EM re-estimation are conditional expected frequencies (relying on the conditional probability \( p(x|y) := \frac{\mu(x)}{\tilde{p}(y)} \)) for rules \( r \) and categories \( A \): \( p(\cdot|y) \mid f_r \) := \( \sum_{x \in T(y)} p(x|y) \cdot f_r(x) \) and \( p(\cdot|y) \mid f_A \) := \( \sum_{x \in T(y)} p(x|y) \cdot f_A(x) \), where \( f_A(x) := \sum_{r \in G_A} f_r(x) \) is the frequency of category \( A \) occurring in \( x \), and \( G_A \) is the set of grammar rules with left-hand side \( A \). See e.g. [3]:

**Lemma:** EM re-estimation formulas for probabilistic context-free grammars are given by:

\[
\tilde{p}(r) = \frac{\tilde{p} \left[ \frac{p(\cdot|y) \mid f_r}{p(\cdot|y) \mid f_A} \right]}{\sum_y \tilde{p}(y) \cdot p(\cdot|y) \mid f_r} = \frac{\sum_y \tilde{p}(y) \cdot p(\cdot|y) \mid f_r}{\sum_y \tilde{p}(y) \cdot p(\cdot|y) \mid f_A} \quad (r \in G, \ A = \text{lhs}(r)).
\]

## 3 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 ([1], [3]), 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 knowledge all mentioned properties have not been formally proven till now.

**Theorem:** For a context-free grammar in Chomsky normal form, let \( \tilde{p}(r) \) be re-estimated rule probabilities resulting from one single step of the inside-outside algorithm using the current rule probabilities \( p(r) \). Then: (i) The log likelihood \( L(.) \) of the training corpus increases monotonically, i.e. \( L(\tilde{p}) \geq L(p) \). (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 to \( \hat{p}_{EM}(r) \) resulting from one single EM iteration (using also \( p(r) \) as current rule probabilities).

**Proof:** (i) and (ii) follow using both (iii) and the convergence properties of EM. (iii): The empirical distribution of the sentences is defined as \( \hat{p}(y) = \frac{f(y)}{N} \), where \( f(y) \) is the frequency of \( y \) occurring in the corpus \( y_1 \ldots y_N \). Thus, for each rule \( r \) with left-hand side \( A \): \( \hat{p}_{EM}(r) = \frac{\sum_{y \in y_1 \ldots y_N} \sum_{x \in T(y)} p(x|y) \cdot f_r(x)}{\sum_{y \in y_1 \ldots y_N} \sum_{x \in T(y)} p(x|y) \cdot f_A(x)} \).

Comparing these formulas with the re-estimation formulas presented by \[2\], it follows \( \hat{p}(r) \), if for each sentence \( y \), for each rule \( r \) and each category \( A \) the following propositions can be shown:

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

This is the goal of the rest of the proof, which we split in two lemmas. The first lemma is probably due to \[2\], where corresponding formulas are used, but not explicitly proven, to present inside-outside estimation. The lemma says that category counts can be computed by summing certain rule counts.

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

**Proof:** Assuming Chomsky normal form, and \( y = w_1 \ldots w_n \):

\[
\sum_{r \in G_A} C_y(r) = \sum_{A} C_y(A \to a) + \sum_{B,C \in G} C_y(A \to B \ C)
\]

\[
= \sum_{A} \frac{1}{P} \sum_{1 \leq t \leq n, w_t = a} e(t, t, A) f(t, t, A)
\]

\[
+ \sum_{B,C \in G} \frac{1}{P} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} \sum_{s=1}^{t-1} p(A \to BC)e(s, r, B)e(r + 1, t, C)f(s, t, A)
\]

\[
= \frac{1}{P} \left( \sum_{1 \leq t \leq n} e(t, t, A) f(t, t, A) \right)
\]

\[
+ \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} f(s, t, A) \sum_{B,C \in G} p(A \to BC)e(s, r, B)e(r + 1, t, C)
\]

\[
= \frac{1}{P} \left( \sum_{1 \leq t \leq n} e(t, t, A) f(t, t, A) + \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} f(s, t, A) e(s, t, A) \right)
\]

\[
= \frac{1}{P} \sum_{1 \leq s \leq t \leq n} e(s, t, A) f(s, t, A) = C_y(A).
\]

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

It follows that the desired 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, since \( C_y(A) = \sum_{A \to \alpha} C_y(A \to \alpha) \), and per definition \( f_A(x) = \sum_{A \to \alpha} f_{A \to \alpha}(x) \). Thus, the proof of the theorem is completed, as once as the following central lemma has been proven. It states that the counts of the inside-outside algorithm can be identified with the expected rule frequencies of the EM algorithm.

**Lemma:** For each sentence \( y \) and each rule \( r \): \( C_y(r) = \sum_{x \in T(y)} p(x|y) \cdot f_r(x) = p(\cdot|y) [ f_r ] \).

**Proof:** The second equation is simply the definition of the expectation. Assuming Chomsky normal form, two cases must be considered. First, the rule has the form \( A \to B \ C \):

For a given sentence \( y = w_1 \ldots w_n \) and given three spans \( (s, r, B), (r + 1, t, C), (s, t, A) \) with \( 1 \leq s \leq r < t \leq n \), let \( X_{(s,t,A)(s,r,B)(r+1,t,C)} \) be the parse forest corresponding to the following derivation: \( S \Rightarrow^* w_1 \ldots w_{s-1} A w_{s+1} \ldots w_t \Rightarrow w_1 \ldots w_{s-1} B C w_{t+1} \ldots w_n \Rightarrow^* w_1 \ldots w_r C w_{r+1} \ldots w_n \Rightarrow^* \)
Let \( f(s,t,A)(s,r,B)(r+1,t,C)(x) := \begin{cases} 1 \quad \text{if} \quad x \in X(s,t,A)(s,r,B)(r+1,t,C) \\ 0 \quad \text{else} \end{cases} \) be the characteristic function interpreting \( X(s,t,A)(s,r,B)(r+1,t,C) \) as a simple subset of the set of all possible syntax trees \( T(y) \) of the sentence \( y \). Thus, the frequency \( f_{A \rightarrow BC}(x) \) of the rule \( A \rightarrow BC \) occurring in the syntax tree \( x \in T(y) \) can be computed as follows:

\[
f_{A \rightarrow BC}(x) = \sum_{1 \leq s \leq r < t \leq n} f(s,t,A)(s,r,B)(r+1,t,C)(x).
\]

Using the linear properties of the expected frequencies \( p(.|y)[. .] \), it follows:

\[
p(.|y)[f_{A \rightarrow BC}] = p(.|y) \left[ \sum_{1 \leq s \leq r < t \leq n} f(s,t,A)(s,r,B)(r+1,t,C) \right]
\]

\[
= \sum_{1 \leq s \leq r < t \leq n} p(.|y) [f(s,t,A)(s,r,B)(r+1,t,C)]
\]

\[
= \sum_{1 \leq s \leq r < t \leq n} \sum_{x \in T(y)} p(x|y) \cdot f(s,t,A)(s,r,B)(r+1,t,C)(x)
\]

\[
= \frac{1}{p(y)} \sum_{1 \leq s \leq r < t \leq n} \sum_{x \in T(y)} p(x) \cdot f(s,t,A)(s,r,B)(r+1,t,C)(x)
\]

\[
= \frac{1}{p(y)} \sum_{1 \leq s \leq r < t \leq n} p(X(s,t,A)(s,r,B)(r+1,t,C))
\]

\[
= \frac{1}{p} \sum_{1 \leq s \leq r < t \leq n} f(s, t, A) \cdot p(A \rightarrow BC) \cdot e(s, r, B) \cdot e(r + 1, t, C)
\]

\[
= Cy(A \rightarrow BC).
\]

The second case, for rules of the form \( A \rightarrow a \), follows analogously with spans \((s, s, A)\) and \((s, s, a)\). Here, the details are omitted, but see [6] q.e.d.

References

[1] James K. Baker. Trainable grammars for speech recognition. In D. Klatt and J. Wolf, editors, *Speech Communication Papers for ASA’97*, pages 547–550, 1979.

[2] Eugene Charniak. *Statistical Language Learning*. M.I.T. Press, Cambridge, MA, 1993.

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

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

[5] Detlef Prescher. Inside-outside estimation meets dynamic EM — GOLD. Technical report, DFKI GmbH, Language Technology Lab, 2001.

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