3.8 EM-Algorithmus für kontextfreie Grammatiken

In diesem Abschnitt wird der EM-Algorithmus auf probabilistische kontextfreie Grammatiken angewendet. Es wird sich zeigen, dass sich die *re-estimierten Grammatikregel- -Wahrscheinlichkeiten* berechnen, indem die *erwartete Häufigkeit einer Grammatikregel* durch die *erwartete Häufigkeit ihrer Mutterkategorie* dividiert wird. Der Erwartungswert muss hierbei auf dem *gegenwärtigen Korpus der vollständigen Datentypen* ermittelt werden. Mit Hilfe der in diesem Abschnitt vorgestellten Re-Estimierungsformeln wird in Abschnitt 3.9 gezeigt werden, dass der *Inside-Outside-Algorithmus* (Baker 1979) eine *dynamische Programmiervariante des EM-Algorithmus* ist. Die wichtigste Folgerung aus dieser Beziehung ist, dass der Inside-Outside-Algorithmus die *Konvergenzeigenschaften des EM-Algorithmus für kontextfreie Grammatiken* erbt. Dieser Beweis wurde bisher noch nie geführt, vielleicht weil häufig falsch zitiert wird, dass ein entsprechender Beweis bereits von Baker (1979) vorgenommen wurde[9].

In diesem Abschnitt werden daher neben den Re-Estimierungsformeln auch sehr genau die Konvergenzeigenschaften dieser Instanz des EM-Algorithmus vorgestellt. Es wird sich zeigen, dass (i) die Log-Likelihood während des Trainings monoton wächst, (ii) dieEM-Hilfsfunktion für die re-estimierten Grammatikregel-Wahrscheinlichkeiten ein *lokales Maximum* annehmen wird, wenn vorausgesetzt werden kann, dass alle Mutterkategorien eine *positive Vorkommenserwartung* aufweisen, und (iii) die Folge der re-estimierten Parameter gegen einen *stationären Punkt* der Log-Likelihood konvergiert. Falls vorausgesetzt werden kann, dass alle Mutterkategorien während der gesamten Re-Estimierung eine positive Vorkommenserwartung haben und die *erwartete Cross-Entropie aller Analysen* kontrollierbar ist, so liegt sogar Konvergenz gegen ein *lokales Maximum* der Log-Likelihood vor. Zusammenfassend kann man bemerken, dass die Verhältnisse sehr viel komplexer sind, als in der Literatur gemeinhin angenommen wird.

Zum Abschluss dieses Abschnitts wird gezeigt, dass das in Abschnitt 2.5 eingeführte *Baumbank-Trainingsverfahren* ein Spezialfall des EM-Algorithmus für annotierte Datentypen und probabilistische kontextfreie Grammatiken ist. Hierfür wird der *EM-Algorithmus für annotierte Datentypen* eingeführt und gezeigt, dass dieser Algorithmus den optimalen Parameter stets bereits nach einer einzigen Iteration findet. Angewandt auf eine Baumbank liefert er die bekannten Re-Estimierungsformeln (siehe Abschnitt 2.5), sowie einen sehr kurzen Beweis der ausgezeichneten Eigenschaft, dass die so estimierten Grammatikregel-Wahrscheinlichkeiten die Baumbank-Wahrscheinlichkeit global maximieren.

[9] Im Gegensatz zu Baker (1979) stellt Baum (1972) den *Forward-Backward-Algorithmus* für Hidden-Markov-Modelle mit Konvergenzaussage und -beweis vor.

Probabilistische kontextfreie Grammatiken

In dieser Arbeit wurden probabilistische kontextfreie Grammatiken und ihre Eigenschaften bereits mehrfach behandelt:

In Abschnitt 2.2 wurden *probabilistische kontextfreie Grammatiken* und die zentralen Begriffe *Grammatikregel-Wahrscheinlichkeit*, *Syntaxbaum-Wahrscheinlichkeit* und *Satz-Wahrscheinlichkeit* eingeführt. Es wurde motiviert, dass die Grammatikregel-Wahrscheinlichkeiten eine *Standard-Nebenbedingung* erfüllen sollten, damit Grammatiken *konsistent* sind, und es wurde bewiesen, dass z.B. *probabilistisches mehrdimensionales Clustering* mit einer probabilistischen kontexfreien Grammatik konsistent ist. In Abschnitt 2.5 wurde das *Baumbank-Trainingsverfahren* und seine *Evaluierung* diskutiert. Es wurden *Baumbanken* und *Baumbank-Grammatiken* an einem Beispiel vorgestellt. Ferner wurde ein neuer Beweis vorgeführt, der zeigt, dass die *Baumbank-Wahrscheinlichkeit* von der Baumbank-Grammatik unter Standard-Nebenbedingung maximiert wird. Der Schlüsselschritt dieses neuen Beweises war, dass die *Baumbank-Wahrscheinlichkeit* ein Wahrscheinlichkeits-Produkt gewisser *Teilkorpora* ist, von denen bekannt ist, dass sie durch ihre jeweiligen empirischen Wahrscheinlichkeitsverteilungen maximiert werden. In Abschnitt 2.6 wurden sehr ausführlich zwei symbolische Parsingverfahren (*Parsewald-* und *CKY-Algorithmus*), sowie drei stochastische Parsingverfahren (*Count-*, *Inside-*, und *Viterbi-Algorithmus*) vorgestellt. Für alle Algorithmen wurde sorgfältig herausgestellt, dass sie *dynamische Programmierung* verwenden und gezeigt, dass alle Algorithmen *effizient* sind, nämlich von der *Ordnung* $O(n^3)$, wobei n die Satzlänge ist. In Abschnitt 2.7 wurde der *Semiring-Parsing-Algorithmus* (Goodman 1998) vorgestellt, der alle in Abschnitt 2.6 eingeführten symbolischen und stochastischen Parsingverfahren verallgemeinert. Es wurde gezeigt, dass der Semiring-Parsing-Algorithmus effizient ist, wenn die Semiring-Operationen in konstanter Zeit vorgenommen werden können. In Abschnitt 2.8 wurden die drei wichtigsten Sprachmodelle vorgestellt: *n-gram-Modelle*, sowie *Markov-* und *Hidden Markov-Modelle*. Es wurde bewiesen, dass n-gram-Modelle weniger allgemein als Markov-Modelle, und diese wiederum weniger allgemein als Hidden Markov-Modelle sind. Es wurde gezeigt, dass alle drei Sprachmodelle durch probabilistische kontextfreie Grammatiken modelliert werden können und es wurde darauf hingewiesen, dass der bekannte *Forward-Backward-Algorithmus* (Baum 1972) sich deshalb als ein einfacher Spezialfall innerhalb der EM-Theorie erweisen wird.

Im Abschnitt 3.1 wurde der *Inside-Outside-Algorithmus* vorgestellt, mit dem es möglich ist, eine gegebene kontextfreie Grammatik auf unannotiertem Text (d.h. ohne Benutzung einer Baumbank) zu trainieren. Es wurde gezeigt, dass sich der Inside-Outside-Algorithmus als die iterative Hintereinanderausführung des *Inside-, Outside- und Conditional-Algorithmus* darstellt, wobei in einem einzelnen Schritt die *bedingten Grammatikregel-*

Wahrscheinlichkeiten für alle Sätze eines gegebenen Textkorpus akkumuliert werden und in neue (bezüglich Textkorpus und kontextfreie Grammatik besser angepasste) Grammatikregel-Wahrscheinlichkeiten umgerechnet werden. Der Inside-Outside-Algorithmus wurde erstmals von Baker (1979) vorgestellt. Leider wird oft falsch zitiert, dass sich in dieser Arbeit auch eine Konvergenzaussage für die Log-Likelihoodwerte der trainierten Grammatiken, und ein entsprechender Konvergenzbeweis finden lassen. Dies ist nicht der Fall, obwohl die probabilistischen kontextfreien Grammatiken tatsächlich gewisse Konvergenz-Eigenschaften besitzen, wie am Ende dieses Abschnittes gezeigt wird.

In Abschnitt 2.2 wurden die folgenden Bezeichnungen für die *symbolischen Komponenten einer probabilistischen kontextfreien Grammatik* G eingeführt[10]:

- $A \in G$ bezeichnet eine *Grammatikkategorie* A der Grammatik G,
- $r \in G$ bezeichnet eine *Grammatikregel* r der Grammtik G,
- $\text{lhs}(r)$ bezeichnet die *linke Seite* einer Grammatikregel $r \in G$,
- $r \in G_A$ bezeichnet eine Grammatikregel $r \in G$ mit $\text{lhs}(r) = A$,
- $x \in \mathcal{T}(G)$ bezeichnet einen *Syntaxbaum* x der Grammatik G,
- $r \in x$ bezeichnet eine *Grammatikregel* r aus dem Syntaxbaum $x \in \mathcal{T}(G)$,
- $y \in \mathcal{L}(G)$ bezeichnet einen *Satz* der Grammatik G,
- $x \in \mathcal{T}(y)$ bezeichnet einen *Syntaxbaum* des Satzes $y \in \mathcal{L}(G)$.

Als *stochastische Komponenten der probabilistischen kontextfreien Grammatik* kommen die folgenden Definitionen hinzu:

- $p(r)$ bezeichnet die Wahrscheinlichkeit einer Grammatikregel $r \in G$,
- $f_r(x)$ bezeichnet die Vorkommenshäufigkeit der Grammatikregel $r \in G$ in dem Syntaxbaum $x \in \mathcal{T}(G)$,
- $f_A(x)$ bezeichnet die Vorkommenshäufigkeit der Grammatikkategorie $A \in G$ in dem Syntaxbaum $x \in \mathcal{T}(G)$, genauer[11]
$$f_A(x) = \sum_{r \in G_A} f_r(x) \ ,$$

[10] In Abschnitt 2.2 wurden einige formale Details ausgelassen. Diese können jedem Standardwerk, wie Hopcroft und Ullman (1979) oder Aho und Ullman (1972) entnommen werden.

[11] Dies setzt voraus, dass x ein Syntaxbaum ist, in dem keine Grammatikkategorien, als *Terminalknoten* vorkommen. Offensichtlich ist dies aber für die Syntaxbäume $x \in \mathcal{T}(y)$ eines Satzes $y \in \mathcal{L}(G)$ der Fall.

- die Wahrscheinlichkeit des Syntaxbaums $x \in \mathcal{T}(G)$ ist
$$p(x) = \prod_{r \in x} p(r)^{f_r(x)} \ ,$$
- die Wahrscheinlichkeit des Satzes $y \in \mathcal{L}(G)$ ist
$$p(y) = \sum_{x \in \mathcal{T}(y)} p(x) \ .$$

Das *Standard-Wahrscheinlichkeitsmodell* für probabilistische kontextfreie Grammatiken macht für die *produktiven* und *erreichbaren* Grammatikkategorien $A \in G$ die folgenden Annahmen:
$$\sum_{r \in G_A} p(r) = 1 \ .$$
Mit dem Theorem von Booth und Thompson (1973) kann mit dieser und einer weiteren Voraussetzung (siehe Abschnitt 2.2) gezeigt werden, dass G *konsistent* ist:
$$\sum_{y \in \mathcal{L}(G)} p(y) = 1 \ .$$

Instantiierung des EM-Algorithmus für probabilistische kontextfreie Grammatiken

Mit diesen Vorgaben gelingt es relativ leicht, den EM-Algorithmus auf probabilistische kontextfreie Grammatiken anzuwenden. Die Abbildung 3.22 zeigt die dafür nötige Instantiierung der symbolischen und stochastischen Elemente des EM-Algorithmus. Die Anwendung der in Abschnitt 3.7 vorgestellten Vorgehensweise zur Bestimmung stationärer Punkte der Log-Likelihood und der EM-Hilfsfunktion führt für probabilistische kontextfreie Grammatiken zu sehr einfachen Re-Estimierungsformeln der Grammatikregel-Wahrscheinlichkeiten.

Lemma (Re-Estimierungsformeln): Der EM-Algorithmus führt für probabilistische kontextfreie Grammatiken zu den folgenden besonders einfachen *Re-Estimierungsformeln für Grammatikregel-Wahrscheinlichkeiten*:
$$\hat{\Theta}_r = \frac{\tilde{p}_\Phi [\ f_r\ ]}{\tilde{p}_\Phi [\ f_A\ ]} \qquad (r \in G,\ A = \text{lhs}(r)) \ .$$
Für einen Syntaxbaum $x \in \mathcal{X}$ bezeichnet $\tilde{p}_\Phi(x) = \tilde{p}(y) \cdot p_\Phi(x|y)$ die mit den *gegenwärtigen Grammatikregel-Wahrscheinlichkeiten* $\Phi \in \Omega$ berechnete *empirische Wahrscheinlichkeit*, während $f_A(x) = \sum_{r \in G_A} f_r(x)$ die Vorkommenshäufigkeit einer Grammatikkategorie $A \in G$ in x bezeichnet. Für die *re-estimierten Grammatikregel-Wahrscheinlichkeiten* $\hat{\Theta} \in \Omega$ und die EM-Parameter-Abbildung gilt also $\mathcal{M} :\ \Omega \to \Omega,\ \Phi \mapsto \hat{\Theta}$.

Die Elemente des EM-Algorithmus

$$< \mathcal{Y},\ \tilde{p}(.),\ \mathcal{X},\ X(.),\ \Omega,\ \Omega_0,\ \{p_\Theta(.)|\Theta \in \Omega\},\ \mathcal{M}(.) >$$

werden für eine probabilistische kontextfreie Grammatik G wie folgt instantiiert:

- $\mathcal{Y} = \mathcal{L}(G)$, die Menge der unvollständigen Datentypen besteht aus den Sätzen der Grammatik G,
- $\tilde{p}(.)$, die empirische Wahrscheinlichkeitsverteilung der unvollständigen Datentypen berechnet sich aus den grammatischen Sätzen eines grossen vorgegebenen Trainingskorpus von freiem Text,
- $\mathcal{X} = \mathcal{T}(G)$, die Menge der vollständigen Datentypen besteht aus den Syntaxbäumen der Grammatik G,
- $X(y) = \mathcal{T}(y)$, die symbolische Analyse-Komponente ist ein *Parser*, der zu jedem grammatischen Satz dessen Syntaxbäume liefert,
- $\Omega = \{\ \Theta \in [0,1]^{|G|}\ |\ g_A(\Theta) = 0\ \}$, ein *Parameterraum mit Nebenbedingungen*. Mit $|G|$ wird die Anzahl der Grammatikregeln $r \in G$ bezeichnet. Die Parametrisierung $\Theta_r = p(r)$ und die Annahme des Standard-Wahrscheinlichkeitsmodells ergeben die *Parameter-Nebenbedingungen*:
$$g_A(\Theta) = 1 - \sum_{r \in G_A} \Theta_r \qquad (A \in G)\ ,$$
- Ω_0, die Startparameter ergeben sich durch eine zufällige Auswahl aus dem eben spezifizierten Parameterraum,
- die parametrisierten Wahrscheinlichkeitsverteilungen der vollständigen Datentypen
$$p_\Theta(x) = \prod_{r \in G} \Theta_r^{f_r(x)}$$
ergeben sich mit der Parametrisierung $\Theta_r = p(r)$ aus dem Standard-Wahrscheinlichkeitsmodell,
- die EM-Parameter-Abbildung $\mathcal{M}:\ \Omega \to \Omega,\ \Phi \mapsto \hat{\Theta}$ maximiert die (interpretierte) EM-Hilfsfunktion:
$$\begin{aligned} \hat{\Theta} &= \operatorname{argmax}_{\Theta \in \Omega}\ L_\Phi(\Theta) \\ &= \operatorname{argmax}_{\Theta \in \Omega}\ \tilde{p}_\Phi\,[\ \log\ p_\Theta\ ]\ . \end{aligned}$$

Abbildung 3.22: Der EM-Algorithmus für probabilistische kontextfreie Grammatiken

Beweis: Gemäss Abbildung 3.22 gibt es zu jedem $r \in G$ ein $c \in \mathbb{R}$ mit:

$$p_\Theta(x) = c \cdot \Theta_r^{f_r(x)}\ ,$$

wobei c nicht von Θ_r abhängt. Mit den Rechenregeln des Δ-Operators (siehe Abschnitt 3.7) folgt daher:

$$\begin{aligned} \Delta_r p_\Theta(x) &= \Delta_r \left(c \cdot \Theta_r^{f_r(x)}\right) \\ &= \Delta_r\ \Theta_r^{f_r(x)} \\ &= f_r(x) \cdot \Delta_r\ \Theta_r \\ &= \frac{f_r(x)}{\Theta_r}\ . \end{aligned}$$

Gemäss Abschnitt 3.7 berechnen sich die stationären Punkte der interpretierten EM-Hilfsfunktion unter Benutzung von

$$\begin{aligned} \partial_r L_\Phi(\Theta) &= \tilde{p}_\Phi\,[\ \Delta_r p_\Theta\ ] \\ &= \tilde{p}_\Phi \left[\ \frac{f_r}{\Theta_r}\ \right] \\ &= \frac{1}{\Theta_r} \cdot \tilde{p}_\Phi\,[\ f_r\ ] \end{aligned}$$

Da die Parameter $\Theta \in [0,1]^{|G|}$ die Parameter-Nebenbedingungen $g_A(\Theta) = 0$ erfüllen sollen, reicht es leider nicht aus, die lokalen Maxima $\hat{\Theta} \in \Omega$ der interpretierten EM-Hilfsfunktion mit dem einfachen

$$\partial_r L_\Phi(\Theta)|_{\Theta=\hat{\Theta}} = 0$$

zu bestimmen. Stattdessen müssen sie mit der tiefen *Multiplikatorregel* von **J.L. Lagrange** berechnet werden (siehe Abschnitt 2.4), die hier besagt, dass es zu jeder Nebenbedingung $g_A(\Theta) = 0$ einen *Multiplikator* $\lambda_A \in \mathbb{R}$ gibt, sodass das Gleichungssystem

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

erfüllt ist. Unter Benutzung der bereits berechneten ersten partiellen Ableitungen von $L_\Phi(\Theta)$ und

$$\partial_r\ g_A(\Theta) = \begin{cases} -1 & \text{für } r \in G_A \\ 0 & \text{sonst} \end{cases}$$

kann dieses Gleichungssystem vereinfacht werden:

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

Dies liefert die Lösung:

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

Aus der Parameter-Nebenbedingung $g_A(\hat{\Theta}) = 0$ ergibt sich hieraus noch (per Summation über die $r \in G_A$):

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

Die letzte Umformung ist eine Rechenregel des Erwartungswerts. Die letzten beiden Gleichungen ergeben zusammen die Behauptung. **q.e.d.**

Eigenschaften des EM-Algorithmus für probabilistische kontextfreie Grammatiken

Das nächste Theorem ist neu und stellt die wichtigsten Eigenschaften des EM-Algorithmus für probabilistische kontextfreie Grammatiken vor.

Theorem: Sei G eine probabilistische kontextfreie Grammatik. Seien $\hat{\Theta} \in \Omega$ die mit dem EM-Algorithmus aus den gegenwärtigen Grammatikregel-Wahrscheinlichkeiten $\Phi \in \Omega$ re-estimierten Grammatikregel-Wahrscheinlichkeiten (siehe Lemma). Dann gelten die folgenden Aussagen:

(i) Die Log-Likelihood $L(\Theta)$ des EM-Algorithmus wächst während der Re-Estimierung monoton:

$$L(\hat{\Theta}) \geq L(\Phi)\;,$$

(ii) Die interpretierte EM-Hilfsfunktion $L_\Phi(\Theta)$ hat an der Stelle $\Theta = \hat{\Theta}$ ein *lokales Maximum*, wenn für alle Grammatikkategorien $A \in G$ gilt:

$$\tilde{p}_\Phi\,[\; f_A \;] > 0\;,$$

(iii) Ist Θ^* der Grenzwert einer durch iterative Re-Estimierung erhaltenen Parameterfolge des EM-Algorithmus, so hat die Log-Likelihood an der Stelle Θ^* einen *stationären Wert* (ein lokales Minimum oder Maximum oder einen Sattelpunkt). Existiert zu jeder Grammatikkategorie $A \in G$ eine Konstante $\lambda \in \mathbb{R}$ mit:

$$\tilde{p}_\Phi\,[\; f_A \;] \;>\; \lambda \;>\; 0 \qquad (\text{für alle } \Phi \text{ der Parameterfolge})\;,$$

so hat die Log-Likelihood $L(\Theta)$ des EM-Algorithmus an der Stelle $\Theta = \Theta^*$ ein *lokales Maximum*, falls ($r, s \in G$)

$$\partial^2_{r,s} H_{\mathrm{cross}}(\Theta;\; \Theta^*)\Big|_{\Theta=\Theta^*} \;=\; 0\;.$$

Beweis: Der re-estimierte Parameter $\hat{\Theta}$ ist ein stationärer Punkt der interpretierten EM-Hilfsfunktion (siehe Beweis des letzten Lemmas). Somit folgt (i) aus Theorem 1 (Dempster et al. 1977), und (iii) folgt aus (ii) und dem Theorem aus Abschnitt 3.7. Es bleibt (ii) zu zeigen.

Hierzu werden die zweiten partiellen Ableitungen der interpretierten EM-Hilfsfunktion $L_\Phi(\Theta)$ an der Stelle $\Theta = \hat{\Theta}$ bestimmt. Es gilt ($r, s \in G$):

$$\partial^2_{s,r} L_\Phi(\Theta) \;=\; \tilde{p}_\Phi\,[\; \partial_s \Delta_r p_\Theta \;]\;.$$

Es wurde bereits gezeigt, dass:

$$\Delta_r p_\Theta(x) \;=\; \frac{f_r(x)}{\Theta_r}\;.$$

Somit kann die noch wichtig werdende Beobachtung getroffen werden, dass:

$$\partial_s \Delta_r p_\Theta(x) = 0 \text{ für } s \neq r\;.$$

Für $s = r$ folgt:

$$\partial_s \Delta_r p_\Theta(x) = -\frac{1}{\Theta_r^2} \cdot f_r(x) \ .$$

Zusammenfassend:

$$\partial^2_{s,r} L_\Phi(\Theta)\Big|_{\Theta=\hat\Theta} \ = \ \begin{cases} 0 & \text{falls } s \neq r \\ -\frac{1}{\hat\Theta_r^2} \cdot \tilde{p}_\Phi[\ f_r\ ] & \text{falls } s = r \end{cases}$$

Unter Benutzung der Re-Estimierungsformeln folgt ($A = \text{lhs}(r)$):

$$\partial^2_{s,r} L_\Phi(\Theta)\Big|_{\Theta=\hat\Theta} \ = \ \begin{cases} 0 & \text{falls } s \neq r \\ -\tilde{p}_\Phi[\ f_A\ ] \cdot \frac{\tilde{p}_\Phi[\ f_A\ ]}{\tilde{p}_\Phi[\ f_r\ ]} & \text{falls } s = r \end{cases}$$

Aus $\frac{\tilde{p}_\Phi[\ f_A\ ]}{\tilde{p}_\Phi[\ f_r\ ]} \geq 1$ und der Voraussetztung $\tilde{p}_\Phi[\ f_A\ ] > 0$ folgt, dass die Matrix

$$\partial^2_\Theta L_\Phi(\Theta)\Big|_{\Theta=\hat\Theta}$$

nur aus Diagonalelementen besteht, deren Werte alle wohldefiniert und echt kleiner als Null sind. Somit ist diese Matrix *negativ definit*. **q.e.d.**

Der Beweis des Theorems ist relativ einfach. Dies liegt vor allem daran, dass die parametrisierten Wahrscheinlichkeitsverteilungen $p_\Theta(x)$ der vollständigen Datentypen in dieser Instanz des EM-Algorithmus die Eigenschaft besitzen, dass $\partial_s \Delta_r p_\Theta(x)$ Null ist, wenn s und r verschiedene Grammatikregeln sind. Mit dieser Eigenschaft kann die Matrix der zweiten partiellen Ableitungen der EM-Hilfsfunktion als Diagonalmatrix dargestellt werden, sodass leicht entscheidbar ist, ob die interpretierte EM-Hilfsfunktion $L_\Phi(\Theta)$ an der Stelle $\Theta = \hat\Theta$ ein *lokales Maximum* besitzt und eine Parameterfolge des EM-Algorithmus gegen ein lokales Maximum der Log-Likelihood konvergiert. Dies könnte ein sehr guter Hinweis sein, in welchen Fällen der EM-Algorithmus prinzipiell erfolgreich bzw. nicht erfolgreich angewendet werden kann. Die folgende Definition benennt daher solche Wahrscheinlichkeitsverteilungen.

Definition (Parameterseparierbarkeit): Die parametrisierte Wahrscheinlichkeitsverteilung $p_\Theta(x)$ der vollständigen Datentypen heisst *parameterseparierbar*, falls

$$\partial_s \Delta_r p_\Theta(x) \ = \ 0 \qquad \text{für } s \neq r \ .$$

Bemerkung:

1. Die Bedingung

$$\tilde{p}_\Phi[\ f_A\ ] > 0$$

ist sehr schwach. Sie ist erfüllt, wenn es zu jeder Grammatikkategorie $A \in G$ einen Satz $y \in \mathcal{Y}$ im Trainingskorpus gibt ($\tilde{p}(y) > 0$), sodass (i) ein Syntaxbaum $x \in X(y)$ dieses Satzes die Grammatikkategorie enthält und (ii) dieser Syntaxbaum eine positive Modellwahrscheinlichkeit besitzt ($p_\Phi(x) > 0$ bzw. daraus folgend $p_\Phi(x|y) > 0$):

$$\tilde{p}_\Phi[\ f_A\ ] \ \geq \ \tilde{p}_\Phi(x) \cdot f_A(x) \ = \ \tilde{p}(y) \cdot p_\Phi(x|y) \cdot f_A(x) \ > \ 0 \ .$$

2. Vermutlich ist auch die Bedingung

$$\partial^2_{r,s} H_{\text{cross}}(\Theta;\ \Theta^*)\Big|_{\Theta=\Theta^*} \ = \ 0$$

nicht allzu stark, wie die Bemerkungen zum Theorem aus Abschnitt 3.7 vermuten lassen. Für das Training von probabilistischen kontexfreien Grammatiken bietet es sich daher an, in jeder Iteration nicht nur die Log-Likelihoodwerte, sondern auch die Werte der zweiten partiellen Ableitungen der erwarteten Cross-Entropie aller Analysen zu überwachen. Die entsprechenden Berechnungsformeln können zu diesem Zweck dem Theorem aus Abschnitt 3.7 entnommen werden.

Das Baumbank-Trainingsverfahren als Spezialfall des EM-Algorithmus für annotierte Datentypen und probabilistische kontextfreie Grammatiken

Das Ziel dieses Abschnitts ist es, zu zeigen, dass das *Baumbank-Trainingsverfahren* ein Spezialfall des *EM-Algorithmus für annotierte Datentypen* und probabilistische kontextfreie Grammatiken ist.

In Abschnitt 2.5 wurde das *Baumbank-Trainingsverfahren* vorgestellt und ein neuer Beweis vorgeführt, der zeigt, dass die *Baumbank-Wahrscheinlichkeit* von der Baumbank-Grammatik unter Standard-Nebenbedingung maximiert wird. Der Schlüsselschritt dieses neuen Beweises war, dass die *Baumbank-Wahrscheinlichkeit* ein Wahrscheinlichkeits-Produkt gewisser *Teilkorpora* ist, von denen bekannt ist, dass sie durch ihre jeweiligen empirischen Wahrscheinlichkeitsverteilungen maximiert werden.

In diesem Abschnitt soll der Spezialfall des *EM-Algorithmus für annotierte Datentypen* und probabilistische kontextfreie Grammatiken vorgestellt werden. Das Lemma über die Re-Estimierungsformeln der Grammatikregel-Wahrscheinlichkeiten wird dann die in Abschnitt 2.5 aus einer Baumbank erhaltenen Grammatikregel-Wahrscheinlichkeiten liefern, während das letzte Theorem die Aussage liefern wird, dass gerade diese Grammatikregel--Wahrscheinlichkeiten die Baumbank-Wahrscheinlichkeit maximieren. Damit werden die Ergebnisse aus Abschnitt 2.5 ein zweites Mal, aber in einem völlig verschiedenen, und vor allem sehr viel allgemeineren Rahmen bewiesen. Dieses Ergebnis zeigt ein erstes Mal, dass die Theoreme rund um den EM-Algorithmus zwar sehr allgemeingültig formuliert sind, aber doch in der Anwendung zu tiefen Aussagen führen, für deren Beweis sonst ein nicht unerheblicher Aufwand nötig wäre.

In der folgenden Definition wird zunächst der EM-Algorithmus für annotierte Datentypen eingeführt.

Definition (EM-Algorithmus für annotierte Datentypen): Ein EM-Algorithmus ist ein *EM-Algorithmus für annotierte Datentypen*, falls:

- Die Menge der unvollständigen Datentypen fällt mit der Menge der vollständigen Datentypen zusammen:
$$\mathcal{Y} = \mathcal{X} \ .$$
- Es ist ein Trainingskorpus von vollständigen Datentypen gegeben, der durch eine empirische Wahrscheinlichkeitsverteilung repräsentiert wird
$$\tilde{p}(x), \ x \in \mathcal{X} \ .$$
- Die annotierten Datentypen $x \in \mathcal{X}$ sind bereits vollständig annotiert, d.h. $X(x) = \{x\}$. *Im eigentlichen Sinn existiert keine symbolische Analyse-Komponente*, bzw. sie ist trivial:
$$X: \ \mathcal{X} \to \mathcal{P}(\mathcal{X}), \quad x \mapsto \{x\} \ .$$

Man spricht dann von *annotierten Datentypen* $x \in \mathcal{X}$ und dem *Trainingskorpus* $\tilde{p}(x)$ *von annotierten Datentypen.*

Jeder linguistische Korpus, dessen Datentoken in irgendeiner Weise systematisch erstellte linguistische Annotationen tragen, wird diese drei Kriterien erfüllen. Insbesondere gilt dies auch für Baumbanken, die daher auch von dem EM-Algorithmus zum Training einer probabilistischen kontextfreien Grammatik benutzt werden können. Jedes Datentoken dieser Baumbank ist eine (manuell erstellte) linguistische Analyse eines Satzes, die keiner weiteren Annotation bedarf, also ein (manuell disambiguierter) kontextfreier Syntaxbaum der vorgegebenen kontextfreien Grammatik.

Das folgende Lemma ist sehr nützlich. Es hat zur Folge, dass in einem *EM-Algorithmus für annotierte Datentypen* die interpretierte EM-Hilfsfunktion $L_\Phi(\Theta)$ (unabhängig vom gegenwärtigen Parameter $\Phi \in \Omega$) mit der Log-Likelihood $L(\Theta)$ übereinstimmt und daher (i) der optimale Parameter bereits nach einer einzigen Re-Estimierung gefunden wird und (ii) dieser die Log-Likelihood des *Korpus der annotierten Datentypen* maximiert.

Lemma: Ein *EM-Algorithmus für annotierte Datentypen* erfüllt (unabhängig vom gegenwärtigen Parameter $\Phi \in \Omega$):
$$\tilde{p}_\Phi(x) = \tilde{p}(x) \qquad (x \in \mathcal{X}) \ .$$

Beweis: Folgt aus $p_\Phi(x|y) = p_\Phi(x|x) = 1$. **q.e.d.**

Angewandt auf probabilistische kontextfreie Grammatiken liefert dieses Lemma das folgende Korollar, welches überraschend einfach mit Hilfe der allgemeinen EM-Theorie bewiesen werden kann und die Resultate aus Abschnitt 2.5 bestätigt.

Korollar: Sei G eine kontextfreie Grammatik und $\mathcal{X}$ eine Baumbank. Dann führt der *EM-Algorithmus für annotierte Datentypen* zu den folgenden Re-Estimierungsformeln ($r \in G$):
$$\hat{\Theta}_r = \frac{f(r)}{\sum_{r \in G_A} f(r)} \qquad (A = \mathrm{lhs}(r)) \ .$$
Hierbei bezeichnet $f(r) = \sum_{x \in \mathcal{X}} f(x) \cdot f_r(x)$ die *Vorkommenshäufigkeit der Grammatikregel r in der Baumbank $\mathcal{X}$*. Ferner gilt, dass an der Stelle $\Theta = \hat{\Theta}$ ein *globales Maximum* der Log-Likelihood $L(\Theta)$ vorhanden ist:
$$L(\hat{\Theta}) \geq L(\Theta) \qquad (\Theta \in \Omega) \ .$$

Beweis: Unabhängig vom Startparameter $\Phi \in \Omega$ liefert das letzte Lemma mit dem Lemma über die Re-Estimierungsformeln für probabilistische kontextfreie Grammatiken ($r \in G, \ A = \mathrm{lhs}(r)$):
$$\hat{\Theta}_r = \frac{\tilde{p}_\Phi[\ f_r\ ]}{\tilde{p}_\Phi[\ f_A\ ]} = \frac{\tilde{p}[\ f_r\ ]}{\tilde{p}[\ f_A\ ]} \ .$$
Bezeichnet $|f|$ die Anzahl aller Syntaxbäume in der Baumbank $\mathcal{X}$, so folgt:
$$|f| \cdot \tilde{p}[\ f_r\ ] \ = \ |f| \cdot \sum_{x \in \mathcal{X}} \tilde{p}(x) \cdot f_r(x) \ = \ \sum_{x \in \mathcal{X}} f(x) \cdot f_r(x) \ = \ f(r) \ .$$
Dies liefert (mit den Rechenregeln des Erwartungswerts):
$$|f| \cdot \tilde{p}[\ f_A\ ] \ = \ |f| \cdot \tilde{p}\left[\sum_{r \in G_A} f_r\right] \ = \ \sum_{r \in G_A} |f| \cdot \tilde{p}[\ f_r\ ] \ = \ \sum_{r \in G_A} f(r) \ .$$
Alles zusammenfassend erhält man die im Korollar angegebene Re-Estimierungsformel. Es bleibt zu zeigen, dass die Log-Likelihood $L(\Theta)$ an der Stelle $\Theta = \hat{\Theta}$ ein globales Maximum besitzt. Überraschenderweise kann dies hier schon durch die Anwendung des eher simplen Teil (i) des Theorems gezeigt werden, der insbesondere für die Startparameter liefert, dass
$$L(\hat{\Theta}) \geq L(\Phi) \ .$$

Weil die vorher hergeleitete Re-Estimierungsformel aber zeigt, dass $\hat{\Theta}$ von der speziellen Wahl des Startparameters Φ unabhängig ist, folgt die Behauptung. **q.e.d.**