3.3 Anwendung und Interpretation des EM-Algorithmus

In diesem Abschnitt werden zwei Ziele verfolgt:

1. In Abschnitt 3.2 wurde der EM-Algorithmus formal eingeführt. Um ein Gefühl für mögliche Anwendungsumgebungen des EM-Algorithmus zu vermitteln, sollen die dafür wesentlichen Elemente des EM-Algorithmus ein weiteres Mal vorgestellt werden. Hierbei wird die Methodik verfolgt, zuerst die vertrauten *symbolischen Komponenten* des EM-Algorithmus vorzustellen, und dann zu den vielleicht weniger vertrauten *stochastischen Basiskomponenten*, sowie den aus ihnen *abgeleiteten stochastischen Verteilungen und objektiven Funktionen* überzugehen.

2. Die stochastischen Komponenten des EM-Algorithmus, insbesondere aber die nützlichste stochastische Funktion des EM-Algorithmus, nämlich die EM-Hilfsfunktion, sollen interpretiert werden. Als Schlüsselschritt wird sich hierbei die Einführung der sogenannten *empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen* herausstellen. Mit dieser Wahrscheinlichkeitsverteilung über vollständigen Datentypen wird es gelingen die EM-Hilfsfunktion als eine *Log-Likelihood auf vollständigen Datentypen* zu interpretieren, sodass der EM-Algorithmus als eine iterative Maximum-Likelihood-Estimierung von schrittweise verbesserten Log-Likelihood-Funktionen interpretierbar ist. (Obwohl diese Interpretation sehr befriedigend ist, beantwortet sie zunächst noch nicht die Frage, ob das Ergebnis der iterativen Maximum-Likelihood-Estimierungen etwas mit der Maximum-Likelihood-Estimierung der Log-Likelihood des EM-Algorithmus zu tun hat. Dieser Komplex wird ausführlich in Abschnitt 3.6 behandelt werden.)

Anwendung des EM-Algorithmus

Der EM-Algorithmus besteht aus drei *symbolischen Komponenten*, zwei *stochastischen Basiskomponenten*, sowie einigen *abgeleiteten stochastischen Wahrscheinlichkeitsverteilungen und objektiven Funktionen*. Die drei symbolischen Komponenten des EM-Algorithmus, die in Abbildung 3.8 gezeigt werden, gehören zu einer Anwendungsumgebung des EM-Algorithmus, in der diese spezifiziert sein müssen. Es handelt sich um:

- die unvollständigen Datentypen $\mathcal{Y}$,
- die vollständigen Datentypen $\mathcal{X}$,
- sowie die *symbolische Analyse-Komponente* $X(.)$, die jeden unvollständigen Datentyp $y \in \mathcal{Y}$ *analysiert* und ihm als Ergebnis eine Menge $X(y) \subseteq \mathcal{X}$ von vollständigen

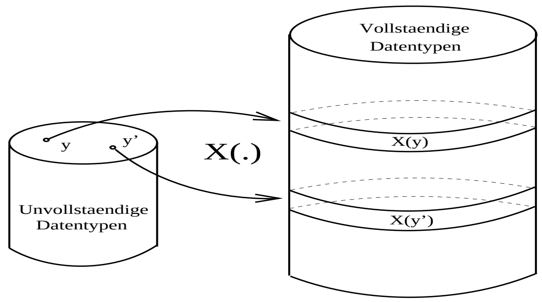

Abbildung 3.8: Symbolische Komponenten des EM-Algorithmus

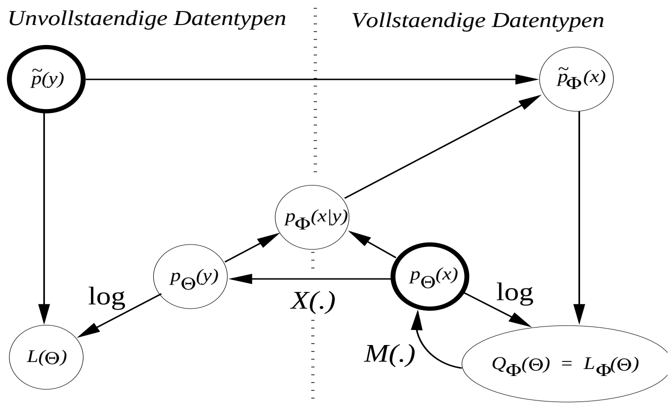

Abbildung 3.9: Stochastische Komponenten des EM-Algorithmus

Datentypen zuweist.

Die Abbildung 3.8 zeigt, dass die symbolische Analyse-Komponente (i) einem unvollständigen Datentyp eine Menge vollständiger Datentypen zuweist, und (ii) verschiedene unvollständige Datentypen $y \neq y'$ auf disjunkte *Analysen* $X(y) \cap X(y') = \emptyset$ abbbildet.

Der symbolische Ansatz des EM-Algorithmus ist daher so allgemein, dass er auf eine Vielzahl von linguistischen Analyse- und Generierungswerkzeugen, wie Speech-Recognizer, Tagger, Silbifizierer, Morphological-Analyzer, Chunker, Parser (mit und ohne Semantikanalyse), Speech-Synthesizer, Generierer, Graphem-Phonem-Systeme, Text-To-Speech-Systeme, Übersetzungs-Systeme, u.s.w. zutreffen wird.

Abbildung 3.9 zeigt, dass der EM-Algorithmus eine Vielzahl von stochastischen Komponenten besitzt, aber nur zwei (**fett umkreiste**) stochastische Basiskomponenten:

- den durch die *empirische Wahrscheinlichkeitsverteilung* $\tilde{p}(y),\ y \in \mathcal{Y}$, gegebenen *Korpus von unvollständigen Datentypen*, sowie

- die *parametrisierte vollständige Wahrscheinlichkeitsverteilung* $p_\Theta(x),\ x \in \mathcal{X}$, welche gelegentlich auch als *Wahrscheinlichkeitsmodell* bezeichnet wird. Implizit sind mit der parametrisierten vollständigen Wahrscheinlichkeitsverteilung natürlich auch die Parameter $\Theta \in \Omega$ gegeben, sowie die zufälligen Startparameter $\Omega_0 \subseteq \Omega$.

Der Sinn ist der folgende: selbst bei der Entwicklung *symbolischer Analysewerkzeuge* müssen im allgemeinen gewisse *Heuristiken* benutzt werden, um dem Werkzeug, in Situationen in denen es vor der Wahl zwischen mehreren möglichen Alternativen der Problembehandlung steht, die Fähigkeit der Entscheidungsfindung zu verleihen. Im allgemeinen manifestieren sich Heuristiken durch die Belegung gewisser Werkzeug-interner Variablen mit bestimmten Werten, die durch Erfahrung und Intuition diktiert werden. Damit stark kontrastierend, wird bei der Entwicklung eines stochastischen Analysewerkzeugs zwar nicht auf Heuristiken und Parameter verzichtet, wohl aber auf die weitgehend unmotivierte Festlegung von Parameterwerten. Ein *stochastisches Analysewerkzeug*, wie es z.B. im EM-Algorithmus vorliegt, kann als ein Paar

$$< X(.),\ p_\Theta(.) >$$

aufgefasst werden und ist einerseits in der Lage, einen unvollständigen Datentyp $y \in \mathcal{Y}$ zu analysieren, d.h. in seine Analysen $x \in X(y) \subseteq \mathcal{X}$ zu zerlegen; andererseits kann es für zwei alternative Analysen $x \neq x'$ vermittels einem Wahrscheinlichkeitsvergleich

$$p_\Theta(x) > p_\Theta(x')\ ,$$

eine Entscheidung für die Analyse x herbeiführen. Der Index Θ deutet an, dass diese Entscheidungsfindung von den Werten gewisser Parameter

$$\Theta \in \Omega$$

abhängt. Die Belegung dieser Parameterwerte findet im EM-Algorithmus durch eine iterative Anwendung von wohldefinierten *E- und M-Schritten* statt, die in jeder Iteration für einen gegenwärtigen Parameter $\Phi \in \Omega$ die *EM-Hilfsfunktion*

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

als eine *objektive Funktion* von Parametern $\Theta \in \Omega$ berechnen (E-Schritt) und einen *re-estimierten* Parameter

$$\hat{\Theta} = \text{argmax}_{\Theta \in \Omega}\ \ Q_\Phi(\Theta)$$

liefern (M-Schritt). Hierbei sind die *bedingten Wahrscheinlichkeiten*

$$p_\Phi(x|y) = \frac{p_\Phi(x)}{p_\Phi(y)}$$

und die *Wahrscheinlichkeiten der unvollständigen Datentypen*

$$p_\Phi(y) = \sum_{x \in X(y)} p_\Phi(x)$$

abgeleitete Wahrscheinlichkeiten der Modell-Wahrscheinlichkeiten $p_\Theta(x)$. (Siehe Abbildung 3.9: ein Pfeil markiert, dass die Quelle in der Definition des Ziels benötigt wird.) Es ist mehr als interessant, dass diese iterative Prozedur mit *zufällig gewählten Startparametern* gestartet wird und in der Praxis keine geschickte Wahl von Startparametern bekannt ist, mit der sich bessere Resultate als mit zufälligen Startparametern erzielen lassen (ausser für den Spezialfall der *Log-linearen Modelle* (Riezler et al. 2000)).

Zusammenfassend kann festgestellt werden, dass im stochastischen Ansatz des EM-Algorithmus ein symbolisches Analysewerkzeug dadurch verbessert werden soll, dass eine durch Erfahrung und Intuition vorgenommene Belegung bestimmter heuristischer Parameterwerte vermieden wird und stattdessen eine mathematisch wohldefinierte iterative Methode zur Auffindung optimaler Parameterwerte benutzt wird.

Interpretation des EM-Algorithmus

In diesem Abschnitt sollen die stochastischen Komponenten des EM-Algorithmus, insbesondere aber die EM-Hilfsfunktion, interpretiert werden. Als Schlüsselschritt wird sich hierbei die Einführung der sogenannten *empirischen Wahrscheinlichkeitsverteilung der vollständigen Datentypen* herausstellen.

Definition: Sei $\Phi \in \Omega$ ein vorgegebener Parameter des EM-Algorithmus. Dann heisst:

$$\tilde{p}_\Phi :\ \mathcal{X} \to [0,1] \qquad x \mapsto \tilde{p}_\Phi(x)\ ,$$

mit

$$\tilde{p}_\Phi(x)\ =\ \tilde{p}(y) \cdot p_\Phi(x|y) \qquad (\text{wobei } y \in \mathcal{Y} \text{ sodass } x \in X(y))$$

Korpus (oder *empirische Wahrscheinlichkeitsverteilung*) *der vollständigen Datentypen (unter dem gegenwärtigen Parameter* Φ*).*

y	y_1			y_2		y_3			
$\tilde{p}(y)$	0.3			0.2		0.5			
$X(y)$	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{31}	x_{32}	x_{33}	x_{34}
$p_\Phi(x\|y)$	0.2	0.4	0.4	0.7	0.3	0.2	0.3	0.1	0.4
$\tilde{p}_\Phi(x)$	0.06	0.12	0.12	0.14	0.06	0.10	0.15	0.05	0.20

Abbildung 3.10: Beispiel einer empirischen Wahrscheinlichkeitsverteilung auf vollständigen Datentypen

Aus $0 \leq \tilde{p}(y),\ p_\Phi(x|y) \leq 1$ folgt offensichtlich $0 \leq \tilde{p}_\Phi(x) \leq 1$. Ferner gibt es, wegen der Bedingung

$$\mathcal{X} = \sum_{y \in \mathcal{Y}} X(y) \ ,$$

zu jedem vollständigen Datentyp $x \in \mathcal{X}$ genau ein $y \in \mathcal{Y}$ mit $x \in X(y)$. Die empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen ist daher wohldefiniert.

Beispiel: Da die Definition des *Korpus der vollständigen Datentypen (unter dem gegenwärtigen Parameter* Φ*)* nicht nur für den EM-Algorithmus, sondern auch für den SB-Algorithmus wichtig ist, wird in Abbildung 3.10 ein Beispiel angegeben. In der zweiten Zeile der Tabelle wird eine empirische Wahrscheinlichkeitsverteilung $\tilde{p}$ von unvollständigen Datentypen

$$\{\ y_1,\ y_2,\ y_3\ \} \subseteq \mathcal{Y}$$

angegeben, die ihrerseits in der ersten Zeile der Tabelle gezeigt werden. Diese Zahlen bedeuten, dass ein Korpus von unvollständigen Datentypen gegeben ist, und z.B. der unvollständige Datentyp y_1 mit 30% aller Vorkommen in ihm vertreten ist. In der dritten Zeile wird für jeden unvollständigen Datentyp $y \in \mathcal{Y}$ die Menge $X(y)$ aller seiner symbolischen Analysen angegeben. Z.B. hat der unvollständige Datentyp y_1 die drei vollständigen Analysen $x_{11},\ x_{12},\ x_{13}$. In der vierten Zeile wird eine (hier fiktive[5]) bedingte Wahrscheinlichkeitsverteilung $p_\Phi(x|y)$ von vollständigen Datentypen $x \in X(y)$, gegeben ihrem unvollständigen Datentyp $y \in \mathcal{Y}$, gezeigt. Auch diese Zahlen können als Vorkommenshäufigkeiten in einem Korpus gedeutet werden, wobei man allerdings davon ausgehen muss, dass in diesem Korpus nur ein einziger unvollständiger Datentyp y mit allen seinen Analysen $X(y)$ beobachtet wurde. Der Korpus des unvollständigen Datentyps y_1 und seiner Analysen $X(y_1)$ wird dann zu 40% aus der Analyse x_{13} bestehen. Aus diesen Vorgaben ergibt sich per Definition der Korpus $\tilde{p}_\Phi$. Er besteht höchstens aus den vollständigen Datentypen

[5] Im EM-Algorithmus wird der Parameter Φ nur zu Beginn zufällig gesetzt, später mit einem mathematisch wohldefiniertem Kriterium estimiert. Die bedingte Wahrscheinlichkeitsverteilung $p_\Phi(x|y)$ ist daher nicht *fiktiv*, sondern wird mit dem Parameter Φ aus dem Modell (der parametrisierten Wahrscheinlichkeitsverteilungen) *berechnet*.

$$\{\ x_{11},\ x_{12},\ x_{13},\ x_{21},\ x_{22},\ x_{31},\ x_{32},\ x_{33},\ x_{34}\ \} \subseteq \mathcal{X}$$

und wird in der fünften Zeile angegeben. In dem Korpus $\tilde{p}_\Phi$ kommt etwa der unvollständige Datentyp x_{13} mit 12% (= 30% $\cdot$ 40%) aller Vorkommen vor, während der unvollständige Datentyp x_{34} mit 20% aller Vorkommen am häufigsten, und der unvollständige Datentyp x_{33} mit 5% aller Vorkommen am seltensten ist.

Das Beispiel zeigt weiter, dass $\tilde{p}_\Phi$ nur auf der Menge

$$\sum_{y \in \mathcal{Y},\ \tilde{p}(y) > 0} X(y)$$

Werte ungleich Null annehmen wird. Sind die Analysen $X(y)$ der unvollständigen Datentypen sämtlich endlich, so folgt, dass der Korpus der vollständigen Datentypen wie der Korpus der unvollständigen Datentypen ein endlicher Korpus sein wird.

Die *empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen* wird in Abbildung 3.9 in Relation zu anderen stochastischen Komponenten des EM-Algorithmus gezeigt. Die Abbildung ist so organisiert, dass die Wahrscheinlichkeitsverteilungen (Korpora) und stochastischen objektiven Funktionen getrennt dargestellt werden, je nachdem, ob sie unvollständige Datentypen (linker Teil der Abbildung) oder vollständige Datentypen (rechter Teil) beschreiben. Die Abbildung ist völlig symmetrisch aufgebaut: man entnimmt ihr zum Beispiel, dass $p_\Theta(y)$ für die unvollständigen Datentypen dieselbe Rolle spielt, wie $p_\Theta(x)$ für die vollständigen Datentypen. Ferner wird sich erweisen, dass $\tilde{p}_\Phi$ für die vollständigen Datentypen dieselbe Rolle spielt wie $\tilde{p}$ für die unvollständigen Datentypen. Insofern ist es gerechtfertigt, $\tilde{p}_\Phi$ als *Korpus* oder *empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen* zu bezeichnen, obwohl doch ganz offensichtlich die vollständigen Datentypen im Gegensatz zu den unvollständigen Datentypen in keinem realen, physischen Korpus vorliegen werden (was in vielen Fällen bewirkt, dass der EM-Algorithmus gegenüber überwachten Trainingsverfahren zu bevorzugen ist). Das folgende Lemma klärt das formale Verhältnis der beiden empirischen Wahrscheinlichkeitsverteilungen zueinander:

Lemma: Sei $\tilde{p}_\Phi$ die empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen. Dann gilt unabhängig vom gegenwärtigen Parameter $\Phi \in \Omega$:

$$\tilde{p}(y) \ = \ \sum_{x \in X(y)} \tilde{p}_\Phi(x) \qquad (y \in \mathcal{Y}).$$

Beweis:

$$\sum_{x\in X(y)} \tilde{p}_\Phi(x) \;=\; \sum_{x\in X(y)} \tilde{p}(y)\cdot p_\Phi(x|y) \;=\; \tilde{p}(y) \sum_{x\in X(y)} p_\Phi(x|y) \;=\; \tilde{p}(y)\ .$$

q.e.d.

Das Lemma zeigt, dass die empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen nicht nur dem Namen nach eine Wahrscheinlichkeitsverteilung ist:

$$\sum_{x\in\mathcal{X}} \tilde{p}_\Phi(x) \;=\; \sum_{y\in\mathcal{Y}}\sum_{x\in X(y)} \tilde{p}_\Phi(x) \;=\; \sum_{y\in\mathcal{Y}} \tilde{p}(y) \;=\; 1\ .$$

Das nächste Lemma ist von zentraler Bedeutung für die Interpretation des EM-Algorithmus. Mit Hilfe der empirischen Wahrscheinlichkeitsverteilung $\tilde{p}_\Phi$ über den vollständigen Datentypen gelingt es, die EM-Hilfsfunktion als eine *Log-Likelihood auf dem Korpus* $\tilde{p}_\Phi$ zu deuten, sowie den EM-Algorithmus als eine iterative Maximum-Likelihood-Estimierung von *Log-Likelihood-Funktionen auf vollständigen Datentypen* zu identifizieren.

Lemma: Sei $\tilde{p}_\Phi$ die empirische Wahrscheinlichkeitsverteilung der vollständigen Datentypen und L_Θ die Log-Likelihood der vollständigen Datentypen auf dem Korpus $\tilde{p}_\Phi$, d.h.

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

Dann gilt für alle $\Theta \in \Omega$:

$$Q_\Phi(\Theta) \;=\; L_\Phi(\Theta)\ .$$

Beweis:

$$\begin{aligned} L_\Phi(\Theta) &= \sum_{x\in\mathcal{X}} \tilde{p}_\Phi(x)\cdot \log\ p_\Theta(x) \\ &= \sum_{y\in\mathcal{Y}}\sum_{x\in X(y)} \tilde{p}_\Phi(x)\cdot \log\ p_\Theta(x) \\ &= \sum_{y\in\mathcal{Y}}\sum_{x\in X(y)} (\tilde{p}(y)\cdot p_\Phi(x|y))\cdot \log\ p_\Theta(x) \\ &= \sum_{y\in\mathcal{Y}} \tilde{p}(y) \sum_{x\in X(y)} p_\Phi(x|y)\cdot \log\ p_\Theta(x) \\ &= Q_\Phi(\Theta) \end{aligned}$$

q.e.d.

Abbildung 3.9 zeigt den im Lemma vorgestellten Zusammenhang in grösserer Klarheit:

- Für die unvollständigen Datentypen (linker Teil der Abbildung) gilt: Der Korpus $\tilde{p}(y)$ und die (Logarithmen der) Wahrscheinlichkeitsverteilung $p_\Theta(y)$ ergeben die Log-Likelihood $L(\Theta)$ des EM-Algorithmus. Der EM-Algorithmus wurde *konzipiert*, um diese Log-Likelihood als Funktion der Parameter Θ zu maximieren (Dempster et al. 1977).

Eingabe: wie in Abbildung 3.7

Ausgabe: dito

Prozedur: * Modifizierte 5. und 6. Zeile des EM-Algorithmus *\

5a. **generate** corpus $\tilde{p}_\Phi(x)$ of complete data (given parameter Φ)

5b. **compute** log-likelihood $L_\Phi(\Theta)$ of model $p_\Theta(x)$ on corpus $\tilde{p}_\Phi(x)$

6. **compute** $\hat{\Theta} = \text{argmax}_{\Theta\in\Omega}\ \ L_\Phi(\Theta)$

Abbildung 3.11: Der EM-Algorithmus unter Benutzung des Korpus und der Log-Likelihood der vollständigen Datentypen

Eingabe: wie in Abbildung 3.7

Ausgabe: dito

Prozedur: * Modifizierte 5. und 6. Zeile des EM-Algorithmus *\

5. **generate** corpus $\tilde{p}_\Phi(x)$ of complete data (given parameter Φ)

6. **compute** maximum-likelihood estimate $\hat{\Theta}$ of model $p_\Theta(x)$ on corpus $\tilde{p}_\Phi(x)$

Abbildung 3.12: Der EM-Algorithmus als iterative Maximum-Likelihood-Estimierung auf dem Korpus der vollständigen Datentypen

- Für die vollständigen Datentypen (rechter Teil der Abbildung) gilt analog: Der Korpus $\tilde{p}_\Phi(x)$ und die (Logarithmen der) Wahrscheinlichkeitsverteilung $p_\Theta(x)$ ergeben die Log-Likelihood $L_\Phi(\Theta)$. Der EM-Algorithmus wurde *realisiert*, indem er bei der Abarbeitung eines kombinierten E- und M-Schritts (bzw. der Ausführung der EM--Parameter-Abbildung $\mathcal{M}(.)$) die EM-Hilfsfunktion $Q_\Phi(\Theta)$ als Funktion der Parameter Θ maximiert. Da die EM-Hilfsfunktion und die Log-Likelihood $L_\Phi(\Theta)$ gemäss Lemma übereinstimmen, wird in jedem Schritt des EM-Algorithmus also eine Maximum-Likelihood-Estimierung auf vollständigen Datentypen ausgeführt (weitere Details siehe unten).

Dieses Lemma hat drei Auswirkungen, von denen die ersten zwei in diesem Abschnitt und die letzte und folgenreichste im nächsten Abschnitt besprochen werden.

(i) Abbildung 3.11 zeigt eine leicht veränderte Implementierung des EM-Algorithmus (vergleiche Abbildung 3.7), wobei allerdings nur die modifizierten Zeilen gezeigt werden. Da die EM-Hilfsfunktion mit der Log-Likelihood des Modells $p_\Theta(x)$ auf dem Korpus $\tilde{p}_\Phi(x)$ der vollständigen Datentypen übereinstimmt, darf die 5. Zeile der Prozedur des EM-Algorithmus in zwei seperate Arbeitsgänge aufgeteilt werden: zunächst wird der Korpus $\tilde{p}_\Phi(x)$ generiert (Zeile 5a), anschliessend wird auf ihm die Log-Likelihood $L_\Phi(\Theta)$ des Modells $p_\Theta(x)$ berechnet (Zeile 5b). Natürlich muss nach diesen Veränderungen auch die 6. Zeile des EM-Algorithmus modifiziert werden, sodass die Log-Likelihood $L_\Phi(\Theta)$ anstelle der EM-Hilfsfunktion $Q_\Phi(\Theta)$ maximiert wird.

(ii) Abbildung 3.12 zeigt eine weitere leichte, allerdings entscheidende Modifikation der Implementierung des EM-Algorithmus. Wieder werden nur die veränderten Zeilen gezeigt. Während die Zeile 5a des modifizierten Algorithmus in Abbildung 3.11 unverändert als neue Zeile 5 übernommen wird, ist die Schlüssel-Beobachtung, dass die Zeilen 5b und 6 eine Maximum-Likelihood-Estimierung darstellen. Diese zwei Zeilen können also zu der neuen Zeile 6 zusammengefasst werden.

Als sehr schönes Resultat gewinnt man die Erkenntnis, dass sich der EM-Algorithmus als eine iterative Anwendung von Maximum-Likelihood-Estimierungen darstellen lässt, wobei es sehr angenehm erscheint, dass jede dieser Maximum-Likelihood-Estimierungen auf einem Korpus $\tilde{p}_\Phi$ von vollständigen Datentypen ausgeführt wird.