## 3.6 Formale Eigenschaften des EM-Algorithmus

Die originale Absicht bei der Einführung des EM-Algorithmus war es, eine iterative Berechnung der Maximum-Likelihood-Estimierung vornehmen zu können (Dempster et al. 1977). Trotz seines langsamen Konvergenzverhaltens (siehe unten, Theorem 4 von Dempster et al. (1977)) wurde der EM-Algorithmus schon in den 70er Jahren zu einem sehr populären numerischen Verfahren innerhalb der Statistik. Im Gegensatz zu den damaligen Erfahrungen mit numerischen Optimierungsmethoden, war der EM-Algorithmus sehr einfach zu implementieren: Lösungen des M-Schritts existierten oft in geschlossener Form, oder konnten mit Standard-Programmpaketen aus der Statistik gelöst werden. Ein weiterer Vorteil des EM-Algorithmus war, dass er nur sehr wenig Speicherplatz benötigte. Während der letzte Aspekt in der Zwischenzeit an Bedeutung verloren hat, wird der EM-Algorithmus heute typischerweise für Probleme eingesetzt, von denen man den Eindruck hat, dass sie ohne Benutung des EM-Algorithmus nur schwer (bzw. garnicht) lösbar sind, oder wenn alternative Lösungsmethoden sehr hohe Kosten verursachen würden (beispielsweise weil sie manuell annotierte Daten benötigen). Aus diesem Grund ist der EM-Algorithmus nach wie vor in der Statistik sehr beliebt, während in der stochastischen Linguistik damit begonnen wird, ihn bewusst für aktuelle Probleme einzusetzen.

In diesem Abschnitt sollen die grundlegenden Eigenschaften des EM-Algorithmus vorgestellt werden, sowie die Frage beantwortet werden, ob der EM-Algorithmus tatsächlich seinen Job macht und eine iterative Maximum-Likelihood-Estimierung vornimmt. Hierzu werden die wichtigsten Theoreme in Dempster et al. (1977) und Wu (1983) vorgestellt werden. Diese Theoreme werden um einige neue Resultate ergänzt, die für sogenannte *injektive Parametrisierungen der Analysen* $X(y)$ in der Lage sind, die Korollare von Dempster et al. (1977) und Wu (1983) über Log-Likelihood-Funktionen mit eindeutigen Maxima zu ersetzen.

#### Einleitende Bemerkungen zur Maximum-Likelihood-Estimierung im Kontext des EM-Algorithmus

In Kapitel 2 wurde die Maximum-Likelihood-Estimierung als ein Optimierungsproblem, nämlich als das Problem der Maximierung der Korpuswahrscheinlichkeit (bzw. der parametrisierten Log-Likelihood) vorgestellt:

$$\begin{aligned}\hat{\Theta} &= \operatorname{argmax}_{\Theta\in\Omega}\ L(\Theta)\\ &= \operatorname{argmax}_{\Theta\in\Omega}\ \sum_{y\in\mathcal{Y}} \tilde{p}(y)\cdot\log\ p_\Theta(y)\ .\end{aligned}$$

Abbildung 3.15 skizziert das Problem. Auf der x-Achse werden die Parameter $\Theta \in \Omega$



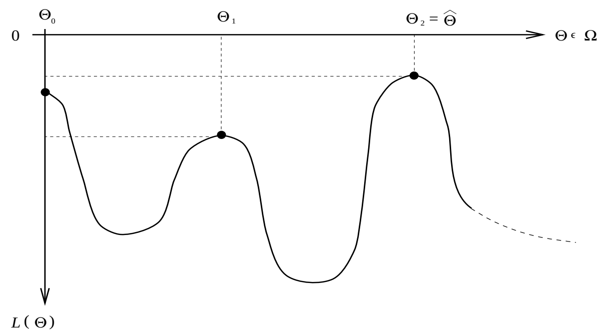


Abbildung 3.15: Maximum-Likelihood-Estimierung

abgetragen, auf der y-Achse die Werte der Log-Likelihood $L(\Theta)$. Da die Wahrscheinlichkeiten $p_\Theta(y)$ kleiner oder gleich 1 sind, sind die Terme log $p_\Theta(y)$ kleiner oder gleich 0. Die Log-Likelihood nimmt deshalb nur negative Werte an. Die globale Maximierung einer Funktion mit mehrdimensionalen Argumenten, welche eventuell gewisse Nebenbedingungen erfüllen sollen, ist ein sehr schwieriges mathematisches Problem. Es wird in der Regel nach folgendem Muster vorgenommen:

1. Die Paramter aller *lokalen Maxima* der Log-Likelihood werden bestimmt. In Abbildung 3.15 sind dies $\Theta_1$ und $\Theta_2$.

2. Der sogenannte *Rand* $\partial\Omega$ des Parameterraums $\Omega$ wird bestimmt. In Abbildung 3.15 sind dies die Parameter $\Theta_0$ und $+\infty$.

3. Der optimale Parameter $\hat{\Theta}$ der Maximum-Likelihood-Estimierung ist der Parameter, der zu dem grössten lokalen Maximum, bzw. zu dem grössten Randwert der Log-Likelihood gehört. In Abbildung 3.15 ist dies der Parameter $\Theta_2$.

Wie der EM-Algorithmus könnte auch die Maximum-Likelihood-Estimierung als ein Erwartungswert-Maximierungs-Verfahren bezeichnet werden, da nach der Berechnung gewisser, relevanter Erwartungswerte:

$$L:\ \Omega \to \mathbb{R},\quad \Theta \mapsto \tilde{p}\,[\ \log\ p_\Theta\ ]$$

eine entsprechende Maximierung erfolgt:

$$\begin{aligned}\hat{\Theta} &= \operatorname{argmax}_{\Theta\in\Omega}\ L(\Theta)\\ &= \operatorname{argmax}_{\Theta\in\Omega}\ \tilde{p}\,[\ \log\ p_\Theta\ ]\ .\end{aligned}$$

| | **Maximum-Likelihood-Estimierung** (auf unvollständigen Datentypen $\mathcal{Y}$) | **EM-Algorithmus** (auf vollständigen Datentypen $\mathcal{X}$ vermittels der symbolischen Analyse-Komponente $X(.)$) |
|---|---|---|
| $L(\Theta)$ | $\sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log\ p_\Theta(y)$ | $\sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log \sum_{x\in X(y)} p_\Theta(x)$ |
| $Q_\Phi(\Theta)$ | - | $\sum_{y\in\mathcal{Y}} \tilde{p}(y) \sum_{x\in X(y)} p_\Phi(x\|y) \cdot \log\ p_\Theta(x)$ |
| $\hat{\Theta}$ | $\text{argmax}_{\Theta\in\Omega}\quad L(\Theta)$ | $\text{argmax}_{\Theta\in\Omega}\quad Q_\Phi(\Theta)$ |

Abbildung 3.16: Der EM-Algorithmus als iteratives Lösungsverfahren der Maximum-Likelihood-Estimierung

In vielen Fällen ist eine Lösung dieses Optimierungsproblems bekannt. So wurde zum Beispiel in Abschnitt 2.4 gezeigt, dass die Parametrisierung $p_\Theta(y) = \Theta_y$ zu der naheliegenden Lösung $\hat{\Theta}_y = \tilde{p}(y)$ führt ($y \in \mathcal{Y}$). Eine einfache, aber sehr wichtige Schlüssel-Beobachtung ist, dass der EM-Algorithmus annimmt, dass sich die Wahrscheinlichkeiten der unvollständigen Datentypen $y \in \mathcal{Y}$ auf der Grundlage der Wahrscheinlichkeiten bestimmter vollständiger Datentypen $x \in X(y)$ erklären lassen. Für den EM-Algorithmus hat dies zur Folge, dass keine direkte Parametrisierung der unvollständigen Wahrscheinlichkeitsverteilungen $p(y) = \sum_{x\in X(y)} p(x)$ vorhanden ist, sondern sich aus der für die vollständigen Wahrscheinlichkeitsverteilungen $p(x)$ vorgenommenen Parametrisierung errechnet. Leider zeigt die nächste Formel, dass sich aus diesem Grund die Maximum-Likelihood-Estimierung für den EM-Algorithmus formal deutlich schwieriger gestaltet als im einleitenden Beispiel:

$$\begin{aligned}\hat{\Theta} &= \text{argmax}_{\Theta\in\Omega} \quad \sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log\ p_\Theta(y) \\ &= \text{argmax}_{\Theta\in\Omega} \quad \sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log \sum_{x\in X(y)} p_\Theta(x)\ .\end{aligned}$$

Obwohl die Maximierung des Logarithmus einer Summe nicht grundsätzlich unmöglich ist, stellt sie doch ein mathematisch anspruchsvolles Problem dar. Es ist interessant, dass der EM-Algorithmus als ein iteratives Lösungsverfahren für dieses Problem angesehen werden kann[7]. Abbildung 3.16 zeigt einen Vergleich der Maximum-Likelihood-Estimierung (zweite Spalte) mit dem EM-Algorithmus (dritte Spalte). Es werden jeweils die Log-Likelihood (zweite Zeile), die EM-Hilfsfunktion (dritte Zeile), und das Optimierungskriterium des estimierten Parameters (vierte Zeile) angegeben. Der Vergleich zeigt:

[7] mit der Konsequenz, dass der EM-Algorithmus nicht angewendet werden sollte, wenn man die Lösung der Maximum-Likelihood-Estimierung direkt finden kann.



1. Der Term $\log\ \sum_{x\in X(y)} p_\Theta(x)$ (der Logarithmus einer Summe), der in der Log-Likelihood $L(\Theta)$ des EM-Algorithmus vorkommt und als Term

   $$\log\ p_\Theta(y)$$

   in der Log-Likelihood der Maximum-Likelihood-Estimierung, wird in der EM-Hilfsfunktion $Q_\Phi(\Theta)$ durch den Term $\sum_{x\in X(y)} p_\Phi(x|y) \cdot \log\ p_\Theta(x)$ (eine Summe von Logarithmen) ersetzt, der als bedingte Erwartung

   $$p_\Phi(.|y)\,[\ \log\ p_\Theta\ ]$$

   interpretierbar ist. Der gegenwärtige Parameter $\Phi$ sollte hierbei als eine Annäherung an den Parameter $\hat{\Theta}$ der Maximum-Likelihood-Estimierung angesehen werden.

2. In der Maximum-Likelihood-Estimierung ist die Maximierung der Log-Likelihood (d.h. der Korpus-Wahrscheinlichkeit) das Optimierungskriterium, während der EM-Algorithmus darauf abzielt, die EM-Hilfsfunktion (iterativ) zu maximieren. Diese unterscheidet sich lediglich um die sogenannte *erwartete Cross-Entropie aller Analysen* von der Log-Likelihood (Lemma, siehe unten).

3. Zusammenfassend erhält man das gewünschte Resultat, dass im EM-Algorithmus (iterativ) die empirische Erwartung

   $$\tilde{p}\,[\ p_\Phi(.|.)\,[\ \log\ p_\Theta\ ]\ ]$$

   zu maximieren ist, was (i) intuitiv annähernd das gleiche Ergebnis liefern könnte, wie die Maximierung der empirischen Erwartung

   $$\tilde{p}\,[\ \log\ p_\Theta\ ]$$

   in der Maximum-Likelihood-Estimierung, aber (ii) als Maximierung einer Summe von Logarithmen mathematisch sehr viel einfacher ist als die Maximierung des Logarithmus einer Summe.

**Ein formaler Vergleich der Log-Likelihood des EM-Algorithmus mit der EM-Hilfsfunktion**

Die nächsten Schritte haben das Ziel, den formalen Zusammenhang zwischen der Log-Likelihood des EM-Algorithmus und der EM-Hilfsfunktion aufzudecken. Die vorangegangene

Diskussion hat insbesondere gezeigt, dass die bedingten Wahrscheinlichkeitsverteilungen ($y \in \mathcal{Y}$):

$$p_\Phi(.|y): \; X(y) \to [0,1] \; ,$$

wegen der Erwartungen

$$p_\Phi(.|y)\,[\, \log \; p_\Theta \,]$$

im engeren Zentrum des Interesses stehen. Gemäss Abschnitt 2.1 kann jede Wahrscheinlichkeitsverteilung als ein Korpus interpretiert werden. Entsprechend gilt dies auch für die bedingten Wahrscheinlichkeitsverteilung $p_\Phi(.|y)$, die für jeden unvollständigen Datentyp $y \in \mathcal{Y}$ als *gegenwärtiger Korpus der Analysen* $X(y)$ interpretiert werden könnten. In Abschnitt 2.3 wurden Entropie und Cross-Entropie als wichtige Bewertungsgrössen eines Korpus eingeführt. Angewandt auf den gegenwärtigen Korpus der Analysen $X(y)$ liefert dies die folgende Definition:

**Definition (Entropie und Cross-Entropie der Analysen $X(y)$):** Sei $y \in \mathcal{Y}$ ein unvollständiger Datentyp und $\Phi,\ \Theta \in \Omega$ zwei Parameter.

$$\begin{aligned} H(\; p_\Phi(.|y)\;) &= -\, p_\Phi(.|y)\,[\, \log\; p_\Phi(.|y)\,] \\ &= -\sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\; p_\Phi(x|y) \end{aligned}$$

heisst *Entropie der Analysen* $X(y)$.

$$\begin{aligned} H_{\text{cross}}(\; p_\Theta(.|y);\; p_\Phi(.|y)\;) &= -\, p_\Phi(.|y)\,[\, \log\; p_\Theta(.|y)\,] \\ &= -\sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\; p_\Theta(x|y) \end{aligned}$$

heisst *Cross-Entropie der Analysen* $X(y)$.

Während die Cross-Entropie der Analysen $X(y)$ die erwartete Anzahl von Datentoken in *bits* (logarithmische Basis 2) oder *nats* (Basis $e$) angibt, die von der Wahrscheinlichkeitsverteilung $p_\Theta(.|y)$ an einer beliebigen (aber festen) Position des Korpus $p_\Phi(.|y)$ generiert werden, muss die Entropie der Analysen $X(y)$ als deren untere Schranke angesehen werden.

Das wichtigste Ergebnis, welches in Abschnitt 2.1 vorgestellt (und in Abschnitt 2.4 bewiesen) wurde, aber auch aus der tiefen *Informations-Ungleichung der Informationstheorie* gefolgert werden kann, besagt, dass die Cross-Entropie eines Korpus stets grösser als seine Entropie ist. Angewandt auf die Analysen $X(y)$ ergibt dies das folgende Resultat:



**Lemma (Entropie und Cross-Entropie der Analysen $X(y)$):** Sei $y \in \mathcal{Y}$ ein unvollständiger Datentyp und $\Phi,\ \Theta \in \Omega$ zwei Parameter. Dann gilt:

$$H_{\text{cross}}(\; p_\Theta(.|y);\; p_\Phi(.|y)\;) \geq H(\; p_\Phi(.|y)\;)\; .$$

Gleichheit gilt genau dann, wenn:

$$p_\Theta(.|y) \;=\; p_\Phi(.|y)\; .$$

Wenn auch völlig trivial, ist es doch ein Schlüsselschritt der folgenden Diskussion, dass sich dieses Ergebnis auf die empirische Erwartung überträgt.

**Definition (Erwartete Entropie und Cross-Entropie aller Analysen):**

$$\begin{aligned} H(\Phi) &= \tilde{p}\,[\; H(\; p_\Phi(.|.)\;)\;] \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot H(\; p_\Phi(.|y)\;) \\ &= -\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\; p_\Phi(x|y) \end{aligned}$$

heisst *erwartete Entropie aller Analysen.*

$$\begin{aligned} H_{\text{cross}}(\Theta;\Phi) &= \tilde{p}\,[\; H_{\text{cross}}(\; p_\Theta(.|.);\; p_\Phi(.|.)\;)\;] \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot H_{\text{cross}}(\; p_\Theta(.|y);\; p_\Phi(.|y)\;) \\ &= -\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log\; p_\Theta(x|y) \end{aligned}$$

heisst *erwartete Cross-Entropie aller Analysen.*

Damit folgt, wie obgen schon angekündigt:

**Lemma (Erwartete Entropie und Cross-Entropie aller Analysen):** Seien $\Phi,\ \Theta \in \Omega$ zwei Parameter. Dann gilt:

$$H_{\text{cross}}(\Theta;\; \Phi) \geq H(\Phi)\; .$$

Gleichheit gilt genau dann, wenn:

$$p_\Theta(.|y) \;=\; p_\Phi(.|y) \qquad \text{für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0\; .$$

**Beweis:** Gemäss Definition gilt:

$$\begin{aligned} &H_{\text{cross}}(\Theta;\; \Phi) - H(\Phi) \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot (H_{\text{cross}}(\; p_\Theta(.|y); p_\Phi(.|y)\;) - H(\; p_\Phi(.|y)\;)) \end{aligned}$$

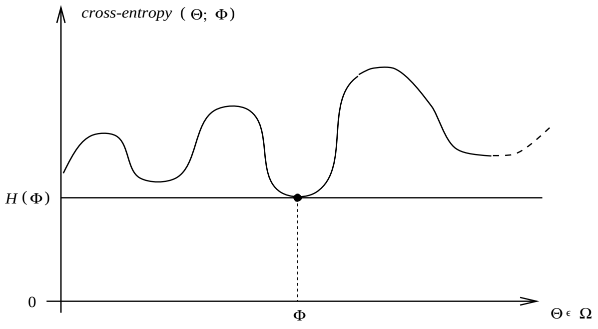


Abbildung 3.17: Erwartete Entropie und Cross-Entropie aller Analysen

Aus $\tilde{p}(y) \geq 0$ und dem ersten Lemma folgt, dass jeder Summand dieser Summe nicht-negativ ist. Hieraus ergibt sich der erste Teil der Behauptung. Ferner gilt, dass diese Summe genau dann Null ist, wenn alle ihre Summanden Null sind. Wieder mit dem ersten Lemma folgt hieraus der zweite Teil der Behauptung **q.e.d.**

Abbildung 3.17 skiziert den eben angegebenen Zusammenhang zwischen erwarteter Entropie und Cross-Entropie aller Analysen. Auf der x-Achse werden die Parameter $\Theta \in \Omega$ abgetragen, auf der y-Achse die Werte der erwarteten Cross-Entropie $H_{\text{cross}}(\Theta; \Phi)$. Die erwartete Entropie $H(\Phi)$ wurde als Konstante abgebildet, weil sie nicht von dem Parameter $\Theta \in \Omega$ abhängig ist. Da die Terme $\log\ p_\Theta(y)$ und $\log\ p_\Phi(y)$ kleiner oder gleich 0 sind, sind die Entropie und die Cross-Entropie der Analysen $X(y)$ nicht negativ. Dies überträgt sich auch auf ihre empirischen Erwartungswerte $H(\Phi)$ und $H_{\text{cross}}(\Theta; \Phi)$. Das Lemma besagt, dass die erwartete Cross-Entropie oberhalb der erwarteten Entropie liegt und wenigstens für den Parameter $\Theta = \Phi$ mit ihr übereinstimmt. Haben die parametrisierten Wahrscheinlichkeitsverteilungen $p_\Theta(.)$ und die symbolische Analyse-Komponente $X(.)$ die folgende Eigenschaft:

$$p_\Theta(.|y) = p_\Phi(.|y) \text{ für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \quad \Rightarrow \quad \Theta = \Phi\ ,$$

so ist $\Theta = \Phi$ tatsächlich auch der einzige Punkt, an dem die erwartete Cross-Entropie mit der erwarteten Entropie übereinstimmen kann. Da es unklar ist, ob man für den EM-Algorithmus fordern sollte, dass dessen Parametrisierung und Analyse-Komponente die angegebene Eigenschaft besitzen, weiss man nicht, ob es weitere Paramter $\Theta$ gibt, an denen die erwartete Cross-Entropie mit der erwarteten Entropie übereinstimmt. In Abbildung 3.17 wurde deshalb nur der eine Punkt $\Theta = \Phi$ eingezeichnet, von dem man dies sicher weiss.

Das nächste (sehr schöne) Resultat zeigt, dass sich die Log-Likelihood des EM-Algorithmus und die EM-Hilfsfunktion gerade um die erwartete Cross-Entropie aller Analysen unterscheiden.



**Lemma (Log-Likelihood und EM-Hilfsfunktion)**: Seien $\Phi,\ \Theta \in \Omega$ zwei Parameter. Dann gilt:

$$L(\Theta) \ = \ Q_\Phi(\Theta) \ + \ H_{\text{cross}}(\Theta;\ \Phi)\ .$$

**Beweis:** Für alle $y \in \mathcal{Y}$ und alle $\Phi \in \Omega$ gilt $\sum_{x \in X(y)} p_\Phi(x|y) = 1$ . Daher erhält man für die Log-Likelihood:

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

Da für jede Analyse $x \in X(y)$ gilt, dass:

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

folgt weiter:

$$\begin{aligned} L(\Theta) &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log \frac{p_\Theta(x)}{p_\Theta(x|y)} \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot (\log p_\Theta(x) - \log p_\Theta(x|y)) \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log p_\Theta(x) \\ &\quad - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log p_\Theta(x|y) \\ &= Q_\Phi(\Theta) + H_{\text{cross}}(\Theta; \Phi) \qquad \textbf{q.e.d.} \end{aligned}$$

Abbildung 3.18 greift eine Idee aus Abschnitt 3.3 auf und zeigt die Cross-Entropie im Kontext wichtiger stochastischer Komponenten des EM-Algorithmus (vergleiche auch Abbildung 3.9). Die Abbildung ist so organisiert, dass diejenigen Komponenten des EM-Algorithmus, die unvollständige Datentypen beschreiben, im linken Teil der Abbildung gezeigt werden, während die Komponenten, die vollständige Datentypen beschreiben, im rechten Teil der Abbildung zu finden sind. Somit muss die empirische Wahrscheinlichkeitsverteilung $\tilde{p}(y)$ und die auf ihr berechnete Log-Likelihood $L(\Theta)$ links, und die empirische Wahrscheinlichkeitsverteilung $\tilde{p}_\Phi(x)$ und die auf ihr berechnete Log-Likelihood $L_\Phi(\Theta)$ rechts

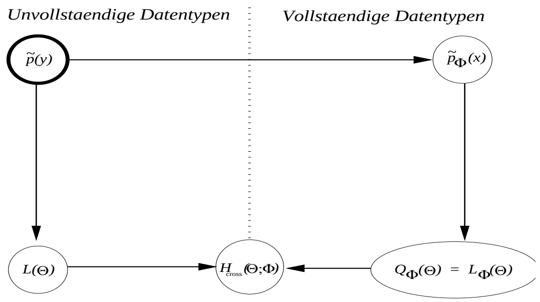


Abbildung 3.18: Die erwartete Cross-Entropie im Kontext der stochastischen Komponenten des EM-Algorithmus

dargestellt werden. Die Cross-Entropie, die sich als Differenz der beiden Log-Likelihood-Funktionen auf unvollständigen bzw. vollständigen Datentypen darstellen lässt, kann daher weder dem linken noch dem rechten Teil der Abbildung zugeordnet werden und wird deshalb in der Mitte dieser beiden wichtigen Funktionen angesiedelt.

Es folgt eine interessante Bemerkung, die darauf abzielt, die neu gewonnene Erkenntnis zur Bestimmung optimaler Modelle zu nutzen.

**Korollar**: Für alle Parameter $\Theta, \Phi \in \Omega$ gilt:

$$L(\Theta) \;\geq\; Q_{\Phi}(\Theta) \;+\; H(\Phi)\ .$$

Somit wird bei der Maximierung der EM-Hilfsfunktion zumindest eine untere Schranke der Log-Likelihood des EM-Algorithmus maximiert.

**Beweis:** Folgt aus dem letzten Lemma und dem Lemma über die erwartete Cross-Entropie und Entropie. **q.e.d.**

### Mathematische Grundlagen aus der Differentialrechnung für den EM-Algorithmus

Die vorangegangene Diskussion hat gezeigt, dass die Maximum-Likelihood-Estimierung mit der globalen Maximierung der Log-Likelihood und der EM-Algorithmus mit der iterativen Maximierung der EM-Hilfsfunktionen beschäftigt ist. Obwohl motiviert wurde, dass die Maximierung der EM-Hilfsfunktion die leichtere von beiden Aufgaben ist, handelt es sich bei der EM-Hilfsfunktion (wie bei der Log-Likelihood) im allgemeinen um eine Funktion mit mehrdimensionalen Parameter-Argumenten und es ist bekannt, dass



die Maximierung solcher Funktionen mathematisch anspruchsvoll ist.

In diesem Abschnitt sollen deshalb die wichtigsten Begriffe und Sätze aus der Differentialrechnung vorgestellt werden, die im weiteren Verlauf von Nutzen sein könnten. Einleitend wurde bereits darauf hingewiesen, dass die *globale Maximierung* einer Funktion ein dreistufiger Prozess ist: (i) der *lokalen Maximierung* folgt (ii) die Bestimmung der *Randwerte* und (iii) die Auswahl der *globalen Maxima* aus der Menge der lokalen Maxima und Randwerte (vergleiche Abbildung 3.15).

Zur Bestimmung der lokalen Maxima gibt es eine Reihe von mathematischen Theoremen, in denen die *partiellen ersten und zweiten Ableitungen* eine wichtige Rolle spielen:

**Definition:** Für $\Theta = < t_1 \dots t_d > \in \Omega \subseteq \mathbb{R}^d$ wird definiert:

- Die *erste partielle Ableitung nach $t_r$* wird mit $\partial_r$ bezeichnet. D.h. es gilt:

$$\partial_r = \frac{\partial}{\partial t_r}\ .$$

- Der *Gradient $\nabla_\Theta$* ist der *Vektor aller ersten partiellen Ableitungen*. D.h. es gilt:

$$\nabla_\Theta \;= < \partial_r \mid r = 1 \dots d > \ .$$

- Die *zweite partielle Ableitung nach $t_r$ und nach $t_s$* wird mit $\partial^2_{r,s}$ bezeichnet. D.h. es gilt:

$$\partial^2_{r,s} \;=\; \partial_r\ \partial_s \;=\; \frac{\partial^2}{\partial t_r \partial t_s}\ .$$

- Die *Matrix aller zweiten partiellen Ableitungen* wird mit $\partial^2_\Theta$ bezeichnet. D.h. es gilt:

$$\partial^2_\Theta \;=\; \left(\partial^2_{r,s}\right)_{r,s=1\dots d} \;=\; \left(\frac{\partial^2}{\partial t_r \partial t_s}\right)_{r,s=1\dots d}\ .$$

Der folgende Satz beschreibt eine wichtige Beobachtung:

**Satz 61 (Erwe (1962), Seite 344)**: Wenn die Funktion $F(\Theta)$ partiell differenzierbar ist und in $\Theta^*$ ein *relatives Extremum* (Maximum oder Minimum) besitzt, so gilt:

$$\nabla_\Theta F(\Theta)|_{\Theta=\Theta^*} = 0\ .$$

Leider gilt die Umkehrung nicht. Abbildung 3.19 zeigt ein Gegenbeispiel. Während für den Parameter $\Phi$ tatsächlich ein relatives Extremum vorliegt, ist dies für $\Phi'$ nicht der Fall, obwohl in beiden Fällen:

$$\nabla_\Theta F(\Theta)|_{\Theta=\Phi} = \nabla_\Theta F(\Theta)|_{\Theta=\Phi'} = 0\ .$$

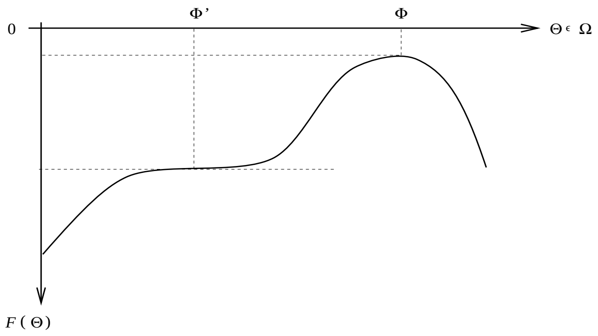


Abbildung 3.19: Maximum-Likelihood-Estimierung

**Definition:** Ein Parameter $\Theta^* \in \Omega$ heisst *stationärer Punkt der Funktion* $F(\Theta)$, falls der *Gradient von* $F(\Theta)$ an dieser Stelle Null ist:

$$\nabla_\Theta F(\Theta)|_{\Theta=\Theta^*} \;=\; 0 \;.$$

Ein Wert $F^* \in \mathbb{R}$ heisst *stationärer Wert der Funktion* $F$, falls

$$F^* = F(\Theta^*)$$

für einen stationären Punkt $\Theta^* \in \Omega$.

Satz 61 (Erwe 1962) besagt, dass an einem stationären Punkt möglicherweise ein *relativer Extremwert (Maximum oder Minimum) der Funktion F* vorliegt. Abbildung 3.19 zeigt jedoch, dass eventuell auch ein sogenannter *Sattelpunkt* (in keinem Fall aber ein beliebiger Funktionswert) vorliegen kann. Der folgende Satz gibt hinreichende Bedingungen für *relative Extrema* an.

**Satz 62 (Erwe (1962), Seite 345)**: Sei die Funktion $F(\Theta)$ zweimal stetig differenzierbar ist. Wenn

$$\nabla_\Theta F(\Theta)|_{\Theta=\Theta^*} = 0 \;,$$

und

$$\partial^2_\Theta F(\Theta)\Big|_{\Theta=\Theta^*}$$

*negativ definit* (bzw. *positiv definit*) ist, so ist $\Theta^*$ ein strenges *relatives Maximum* (bzw. *relatives Minimum*).

Die Begriffe *positiv/negativ definit* und der damit zusammenhängenden Begriff des *Eigenwerts* wurden bereits in Kapitel 2 eingeführt.



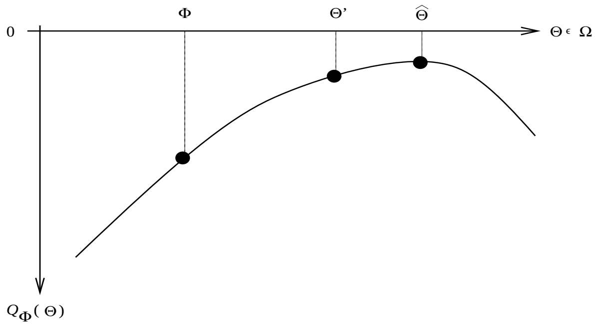


Abbildung 3.20: GEM-Algorithmus und EM-Algorithmus

**Einführung des EM-Algorithmus gemäss der originalen Arbeit von Dempster, Laird und Rubin (1977)**

Es ist interessant, dass seit der Einführung des EM-Algorithmus durch Dempster et al. (1977) alle wichtigen Aussagen über den EM-Algorithmus sogar für den sogenannten *generalisierten EM-Algorithmus* bewiesen wurden.

**Definition (Dempster et al. 1977):** Statt der EM-Parameter-Abbildung:

$$\mathcal{M}:\; \Omega \to \Omega, \quad \Phi \mapsto \mathrm{argmax}_{\Theta\in\Omega} \;\; Q_\Phi(\Theta) \;,$$

die fordert, dass:

$$Q_\Phi(\mathcal{M}(\Phi)) \geq Q_\Phi(\Theta) \qquad \text{für alle } \Phi,\; \Theta \in \Omega \;,$$

wird in einem *generalisierten EM-Algorithmus (GEM-Algorithmus)* von der Parameter-Abbildung $\mathcal{M}(.)$ lediglich verlangt, dass:

$$Q_\Phi(\mathcal{M}(\Phi)) \geq Q_\Phi(\Phi) \qquad \text{für alle } \Phi \in \Omega \;.$$

Abbildung 3.20 skizziert den Unterschied zwischen einen GEM-Algorithmus und einen EM-Algorithmus. Auf der x-Achse werden die Parameter $\Theta \in \Omega$ abgetragen, auf der y-Achse die Werte der EM-Hilfsfunktion $Q_\Phi(\Theta)$. Offensichtlich kann die EM-Hilfsfunktion nur negative Werte anehmen. Obwohl in der Abbildung eine *konkave* EM-Hilfsfunktion gezeigt wird, deren globales Maximum $\hat{\Theta}$ ein lokales Maximum ist, ist es sehr zweifelhaft (obschon wünschenswert), dass beliebige EM-Hilfsfunktionen diese Eigenschaft besitzen. Der Unterschied zwischen einem GEM-Algorithmus und einem EM-Algorithmus liegt in deren Parameter-Abbildungen. Ein EM-Algorithmus verlangt, dass der gegenwärtige Parameter $\Phi$ auf das globale Maximum $\hat{\Theta}$ der EM-Hilfsfunktion $Q_\Phi(.)$ abgebildet wird:

$$\mathcal{M}(\Phi) = \hat{\Theta} \; .$$

Ein GEM-Algorithmus fordert lediglich, dass der gegenwärtige Parameter $\Phi$ auf einen Parameter $\Theta'$ abgebildet wird, der zu einem grösseren Wert der EM-Hilfsfunktion $Q_\Phi(.)$ führt:

$$\mathcal{M}(\Phi) = \Theta' \qquad \text{mit } Q_\Phi(\Theta') \geq Q_\Phi(\Phi) \; .$$

Offensichtlich ist daher jeder EM-Algorithmus ein GEM-Algorithmus. In der Arbeit von Dempster et al. (1977) finden sich vier Theoreme zum GEM-Algorithmus, die alle die folgenden Annahmen machen, ohne sie jeweils explizit aufzuführen.

**Voraussetzungen (Dempster et al. 1977)**:

- $\Omega \subseteq \mathbb{R}^d$ , d.h. der Parameterraum ist eine Teilmenge eines mehrdimensionalen euklidischen Raumes,
- Die Log-Likelihood $L$, die EM-Hilfsfunktionen $Q_\Phi$, die erwartete Cross-Entropie $H_{\text{cross}}$ aller Analysen und die EM-Parameter-Abbildung $\mathcal{M}$ sind "genügend oft" *stetig differenzierbar*,
- Differenzierung und Summation sind vertauschbar.

Obwohl zwei der genannten vier Theoreme (zur sogenannten *Parameter-Konvergenz*) relativ rasch durch Gegenbeispiele (Boyles 1983) widerlegt wurden, erwies sich das folgende Theorem von fundamentaler Bedeutung für die EM-Theorie. Hierin wird gezeigt, dass eine iterative Anwendung der EM-Parameter-Abbildung in eine Folge von Parametern resultiert, deren Log-Likelihood-Werte monoton wachsen (siehe Abbildung 3.7).

**Theorem (Theorem 1 aus Dempster et al. (1977))**: In einem GEM-Algorithmus gilt:

$$L(\mathcal{M}(\Phi)) \geq L(\Phi) \qquad \text{für alle } \Phi \in \Omega \; .$$

Gleichheit liegt genau dann vor, wenn die folgenden beiden Gleichungen erfüllt sind:

$$\text{(i)} \quad Q_\Phi(\mathcal{M}(\Phi)) = Q_\Phi(\Phi)$$
$$\text{(ii)} \quad p_{\mathcal{M}(\Phi)}(.|y) = p_\Phi(.|y) \qquad \text{für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \; .$$

**Beweis:** Mit dem Lemma über die Log-Likelihood und die EM-Hilfsfunktion folgt:

$$\begin{aligned} L(\mathcal{M}(\Phi)) - L(\Phi) &= Q_\Phi(\mathcal{M}(\Phi)) + H_{\text{cross}}(\mathcal{M}(\Phi);\ \Phi) \\ &\quad - Q_\Phi(\Phi) - H_{\text{cross}}(\Phi;\ \Phi) \\ &= Q_\Phi(\mathcal{M}(\Phi)) - Q_\Phi(\Phi) \\ &\quad + H_{\text{cross}}(\mathcal{M}(\Phi);\ \Phi) - H(\Phi) \; . \end{aligned}$$



Für jeden GEM-Algorithmus gilt:

$$Q_\Phi(\mathcal{M}(\Phi)) - Q_\Phi(\Phi) \geq 0 \; .$$

Gemäss dem Lemma über die erwartete Cross-Entropie und Entropie gilt:

$$H_{\text{cross}}(\mathcal{M}(\Phi);\ \Phi) - H(\Phi) \geq 0 \; .$$

Es folgt $L(\mathcal{M}(\Phi)) - L(\Phi) \geq 0$ . Gleichheit tritt genau dann ein, wenn sowohl

$$Q_\Phi(\mathcal{M}(\Phi)) - Q_\Phi(\Phi) = 0 \; ,$$

als auch $H_{\text{cross}}(\mathcal{M}(\Phi);\ \Phi) - H(\Phi) = 0$ . Gemäss dem Lemma über die erwartete Cross-Entropie und Entropie der Analysen tritt das letztere genau dann ein, wenn

$$p_{\mathcal{M}(\Phi)}(.|y) = p_\Phi(.|y) \quad \text{für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \qquad \textbf{q.e.d.}$$

Es ist äusserst interessant, dass Dempster et al. (1977) in einer Randnote bemerken, dass das Lemma über die erwartete Cross-Entropie und Entropie der Analysen, sowie das Theorem 1 bereits von Baum et al. (1970) für den Spezialfall der Hidden-Markov-Modelle vorgestellt und bewiesen wurden.

Die nächsten zwei Korollare geben notwendige (jedoch nicht hinreichende) Bedingungen dafür an, dass eine (iterative) Anwendung der EM-Parameter-Abbildung eine Maximum-Likelihood-Estimierung $\hat{\Theta}$ ausgibt.

**Korollar (Korollar 1 aus Dempster et al. (1977))**: Der mit einem GEM-Algorithmus re-estimierte Parameter $\mathcal{M}(\hat{\Theta})$ einer Maximum-Likelihood-Estimierung $\hat{\Theta}$ (d.h. $L(\hat{\Theta}) \geq L(\Theta)$ für alle $\Theta \in \Omega$) hat die folgenden drei Eigenschaften:
(i) die Log-Likelihood ist *invariant*:

$$L(\mathcal{M}(\hat{\Theta})) = L(\hat{\Theta}) \; ,$$

(ii) die EM-Hilfsfunktion ist *invariant*:

$$Q_{\hat{\Theta}}(\mathcal{M}(\hat{\Theta})) = Q_{\hat{\Theta}}(\hat{\Theta})$$

(iii) die bedingten Wahrscheinlichkeitsverteilungen über die Analysen $X(y)$ sind *invariant*:

$$p_{\mathcal{M}(\hat{\Theta})}(.|y) = p_{\hat{\Theta}}(.|y) \quad \text{für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \; .$$

**Beweis:** Das Theorem 1 aus Dempster et al. (1977) liefert:

$$L(\mathcal{M}(\hat{\Theta})) \geq L(\hat{\Theta}) \ .$$

Da $\hat{\Theta}$ eine Maximum-Likelihood-Estimierung ist, folgt andererseits auch:

$$L(\hat{\Theta}) \geq L(\mathcal{M}(\hat{\Theta})) \ .$$

Damit ist Eigenschaft (i) bewiesen, aus der mit Theorem 1 aus Dempster et al. (1977) die Eigenschaften (ii) und (iii) folgen. **q.e.d.**

**Korollar (Korollar 2 aus Dempster et al. (1977)):** Der mit einem GEM-Algorithmus re-estimierte Parameter $\mathcal{M}(\hat{\Theta})$ einer eindeutigen Maximum-Likelihood-Estimierung $\hat{\Theta}$ (d.h. $L(\hat{\Theta}) > L(\Theta)$ für alle $\Theta \in \Omega \setminus \{\hat{\Theta}\}$) ist *Parameter-invariant*:

$$\mathcal{M}(\hat{\Theta}) = \hat{\Theta} \ .$$

**Beweis:** Korollar 1 aus Dempster et al. (1977) liefert: $L(\mathcal{M}(\hat{\Theta})) = L(\hat{\Theta})$. Da $\hat{\Theta}$ eine eindeutige Maximum-Likelihood-Estimierung ist, folgt die Behauptung. **q.e.d.**

Das folgende Lemma enthälte eine für die EM-Theorie grundlegende Beobachtung.

**Lemma (Lemma 2 aus Dempster et al. (1977)):** Sei $\Phi \in \Omega$. (i) Es gilt:

$$\nabla_\Theta \ H_{\text{cross}}(\Theta, \Phi)|_{\Theta=\Phi} \ = \ 0 \ .$$

D.h. die erwartete Cross-Entropie aller Analysen hat an der Stelle $\Theta = \Phi$ einen *stationären Punkt*. (ii) Ist die Matrix $\partial^2_\Theta \ H_{\text{cross}}(\Theta, \Phi)|_{\Theta=\Phi}$ *positiv definit*, so hat die erwartete Cross-Entropie aller Analysen an der Stelle $\Theta = \Phi$ ein *strenges lokales Minimum*. Unter Benutzung von $\Delta_r p_\Theta(x|y) = \partial_r \log \ p_\Theta(x|y)$ gilt:

$$\begin{aligned}
\partial^2_{r,s} H_{\text{cross}}(\Theta, \Phi)|_{\Theta=\Phi} &= \tilde{p} \left[ \, p_\Phi(.|.) \left[ \, \Delta_r p_\Theta(.|.)|_{\Theta=\Phi} \cdot \Delta_s p_\Theta(.|.)|_{\Theta=\Phi} \, \right] \right] \\
&= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \Delta_r p_\Theta(x|y)|_{\Theta=\Phi} \cdot \Delta_s p_\Theta(x|y)|_{\Theta=\Phi} \ .
\end{aligned}$$

**Beweis:** Erster Beweis zu (i): Dem Lemma über die erwartete Entropie und Cross-Entropie aller Analysen entnimmt man, dass $H_{\text{cross}}(\Theta, \Phi)$ als Funktion von $\Theta$ an der Stelle $\Theta = \Phi$ ein relatives Minimum besitzt. Aus der Differential- und Integralrechnung ist bekannt (Satz 61, Erwe (1962)), dass die relativen Extrema (Minima oder Maxima) von differenzierbaren Funktionen stationäre Punkte sein müssen[8].

[8] Gemäss Satz 63 (Erwe 1962) folgt zusätzlich, dass $\partial^2_\Theta \ H_{\text{cross}}(\Theta, \Phi)|_{\Theta=\Phi}$ *positiv semidefinit* ist.



Zweiter Beweis zu (i) (Dempster et al. (1977)): Per Definition gilt:

$$H_{\text{cross}}(\Theta; \Phi) \ = \ - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \log \ p_\Theta(x|y)$$

Es folgt:

$$\begin{aligned}
\partial_r H_{\text{cross}}(\Theta; \Phi) &= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \partial_r \log \ p_\Theta(x|y) \\
&= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \frac{\partial_r p_\Theta(x|y)}{p_\Theta(x|y)} \ .
\end{aligned}$$

Somit:

$$\begin{aligned}
\partial_r H_{\text{cross}}(\Theta; \Phi)|_{\Theta=\Phi} &= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \frac{\partial_r p_\Theta(x|y)|_{\Theta=\Phi}}{p_\Phi(x|y)} \\
&= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \frac{\partial_r p_\Theta(x|y)|_{\Theta=\Phi}}{p_\Phi(x|y)} \\
&= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} \partial_r p_\Theta(x|y)|_{\Theta=\Phi} \\
&= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \partial_r \left( \sum_{x \in X(y)} p_\Theta(x|y) \right) \Bigg|_{\Theta=\Phi} \\
&= 0 \ ,
\end{aligned}$$

denn die Summe $\sum_{x \in X(y)} p_\Theta(x|y)$ ist für alle Parameter $\Theta \in \Omega$ konstant 1, hat also nur erste partielle Ableitungen, die gleich 0 sind. Hieraus folgt die Behauptung.

zu (ii): Aus dem im Beweis zu (i) hergeleiteten Term für $\partial_r H_{\text{cross}}(\Theta; \Phi)$ folgt durch erneute Differentierung:

$$\partial_s \partial_r H_{\text{cross}}(\Theta; \Phi) = - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \partial_s \left( \frac{\partial_r p_\Theta(x|y)}{p_\Theta(x|y)} \right)$$

Unter Benutzung der *Quotientenregel der Differentialrechnung* folgt:

$$\begin{aligned}
\partial_s \left( \frac{\partial_r p_\Theta(x|y)}{p_\Theta(x|y)} \right) &= \frac{(\partial_s \partial_r p_\Theta(x|y)) \cdot p_\Theta(x|y) - (\partial_r p_\Theta(x|y)) \cdot (\partial_s p_\Theta(x|y))}{p^2_\Theta(x|y)} \\
&= \frac{\partial_s \partial_r p_\Theta(x|y)}{p_\Theta(x|y)} - \frac{\partial_r p_\Theta(x|y)}{p_\Theta(x|y)} \cdot \frac{\partial_s p_\Theta(x|y)}{p_\Theta(x|y)} \\
&= \frac{\partial_s \partial_r p_\Theta(x|y)}{p_\Theta(x|y)} - \Delta_r p_\Theta(x|y) \cdot \Delta_s p_\Theta(x|y)
\end{aligned}$$

Somit:

$$\begin{aligned}
\partial_s \partial_r H_{\text{cross}}(\Theta; \Phi) &= - \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \frac{\partial_s \partial_r p_\Theta(x|y)}{p_\Theta(x|y)} \\
&\quad + \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\Phi(x|y) \cdot \Delta_r p_\Theta(x|y) \cdot \Delta_s p_\Theta(x|y) \ .
\end{aligned}$$

Es folgt:

$$\partial^2_{r,s} H_{\text{cross}}(\Theta;\Phi)\Big|_{\Theta=\Phi} = -\sum_{y\in\mathcal{Y}} \tilde{p}(y) \sum_{x\in X(y)} \partial_s\partial_r p_\Theta(x|y)|_{\Theta=\Phi} + \sum_{y\in\mathcal{Y}} \tilde{p}(y) \sum_{x\in X(y)} p_\Phi(x|y)\cdot \Delta_r p_\Theta(x|y)|_{\Theta=\Phi}\cdot \Delta_s p_\Theta(x|y)|_{\Theta=\Phi} \ .$$

Während die erste Summe verschwindet (wobei man das selbe Argument benutzen kann, wie im Beweis zu Teil (ii)), ist die zweite Summe das gewünschte Ergebnis. **q.e.d.**

Der Sachverhalt wird in Abbildung 3.17 wiedergegeben. Es ist unklar, ob die erwartete Cross-Entropie der Analysen ausser an der Stelle $\Theta = \Phi$ noch weitere stationäre Punkte besitzt (was aber vermutet werden könnte). Der erste Teil dieses Lemmas ist der Schlüssel zu Theorem 4 (Dempster et al. 1977).

Die Theoreme 2 und 3 (Dempster et al. 1977) adressierten die Frage, ob die von einem GEM-Algorithmus ausgegebene Parameterfolge konvergiert. Leider wurden beide Theoreme widerlegt. Das folgende Theorem zeigt lediglich, dass eine konvergente Parameterfolge (i) einen *stationären Punkt der Log-Likelihood* als Grenzwert $\Theta^*$ besitzen muss (vorausgesetzt die Parameterfolge besteht aus stationären Punkten der EM-Hilfsfunktion, was wohl in der Praxis stets der Fall sein wird) und (ii) eine *Konvergenzrate* $\nabla_\Theta \mathcal{M}(\Theta)|_{\Theta=\Theta^*}$ aufweisen wird, die sich mit genau zwei seperaten Informationen über unvollständige bzw. vollständige Datentypen berechnen lässt.



**Theorem (Theorem 4 aus Dempster et al. (1977)):** (i) Sei

$$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots$$

die Parameterfolge eines GEM-Algorithmus, sodass:

- Die Parameterfolge konvergiert:

$$\Theta_i \to \Theta^* \qquad (i \to \infty)\ .$$

- Der Parameter $\Theta_i$ ist ein *stationärer Punkt* der EM-Hilfsfunktion $Q_{\Theta_{i-1}}$, d.h.

$$\nabla_\Theta Q_{\Theta_{i-1}}(\Theta)\Big|_{\Theta=\Theta_i} = 0\ .$$

Dann gilt:

- Der Grenzwert $\Theta^*$ ist ein *stationärer Punkt* der Log-Likelihood, d.h.

$$\nabla_\Theta L(\Theta)|_{\Theta=\Theta^*} = 0\ .$$

(ii) Gilt ferner, dass die *Eigenwerte* der *Matrix der zweiten partiellen Ableitungen* der EM-Hilfsfunktion $Q_{\Theta_{i-1}}$ an der Stelle $\Theta_i$

$$\partial^2_\Theta Q_{\Theta_{i-1}}(\Theta)\Big|_{\Theta=\Theta_i}$$

kleiner oder gleich einem (für alle $i = 1, 2, 3, \ldots$ fixierten) $\lambda < 0$ sind, so folgt nicht nur, dass die EM-Hilfsfunktionen $Q_{\Theta_{i-1}}$ an der Stelle $\Theta_i$ *negativ definit* sind, sondern auch, dass die *Matrix der partiellen zweiten Ableitungen* der EM-Hilfsfunktion $Q_{\Theta^*}$ an der Stelle $\Theta^*$

$$\partial^2_\Theta Q_{\Theta^*}(\Theta)\Big|_{\Theta=\Theta^*}$$

*negativ definit* ist, sodass für die Parameter-Abbildung des GEM-Algorithmus an der Stelle $\Theta^*$ gilt, dass

$$\nabla_\Theta \mathcal{M}(\Theta)|_{\Theta=\Theta^*} = \partial^2_\Theta H_{\text{cross}}(\Theta;\ \Theta^*)\Big|_{\Theta=\Theta^*} \cdot \left(\partial^2_\Theta Q_{\Theta^*}(\Theta)\Big|_{\Theta=\Theta^*}\right)^{-1}$$

**Beweis:** Für den Beweis von (ii) wird auf Dempster et al. (1977) verwiesen, weil im weiteren Verlauf dieser Arbeit nur der erste Teil des Theorems benutzt wird. Zu (i): Mit dem Lemma über die Log-Likelihood und die EM-Hilfsfunktion folgt durch partielle Differenzierung:

$$\nabla_\Theta L(\Theta) = \nabla_\Theta Q_\Phi(\Theta) + \nabla_\Theta H_{\text{cross}}(\Theta;\Phi)\ .$$

Es folgt für $\Phi = \Theta_{i-1}$, dass an der Stelle $\Theta = \Theta_i$ gilt:

$$\nabla_\Theta L(\Theta)|_{\Theta=\Theta_i} = \nabla_\Theta Q_{\Theta_{i-1}}(\Theta)|_{\Theta=\Theta_i} + \nabla_\Theta H_{\text{cross}}(\Theta, \Theta_{i-1})|_{\Theta=\Theta_i} .$$

Wegen der zweiten Voraussetzung an die Parameterfolge wird dies zu:

$$\nabla_\Theta L(\Theta)|_{\Theta=\Theta_i} = \nabla_\Theta H_{\text{cross}}(\Theta, \Theta_{i-1})|_{\Theta=\Theta_i} .$$

Grenzübergang $i \to \infty$ liefert wegen der ersten Voraussetzung an die Parameterfolge, dass:

$$\nabla_\Theta L(\Theta)|_{\Theta=\Theta^*} = \nabla_\Theta H_{\text{cross}}(\Theta, \Theta^*)|_{\Theta=\Theta^*} .$$

Mit dem vorangegangenen Lemma folgt hieraus die Behauptung:

$$\nabla_\Theta L(\Theta)|_{\Theta=\Theta^*} = 0 .$$

**q.e.d.**

**Verbesserte Aussagen zum Konvergenzverhalten des EM-Algorithmus durch Wu (1983)**

Die Arbeit von Wu (1983) wurde durch die folgende Ausgangslage motiviert:

- Obwohl es die ursprüngliche Absicht bei Einführung des EM-Algorithmus war, eine iterative Berechnungsmethode für die Maximum-Likelihood-Estimierung zu etablieren, war es kurze Zeit später bekannt, dass der EM-Algorithmus nicht unter allen Umständen garantiert, dass
$$L^* = \lim_{i\to\infty} L(\Theta_i)$$
ein *globales Maximum* oder zumindest ein *lokales Maximum* der Log-Likelihood ergibt. Vielmehr zeigte der EM-Algorithmus für viele Anwendungen das Verhalten, je nach gewähltem *Startparameter* in ein lokales Maximum oder in einen stationären Log-Likelihoodwert zu münden.

- Obwohl es aus numerischer/computationeller Sicht nicht wichtig erscheint, ob eine EM-Parameterfolge konvergiert, war diese Fragestellung damals wenigstens theoretisch interessant, da die zwei fehlerhaften Theoreme von Dempster et al. (1977) über die *Parameter-Konvergenz* des EM-Algorithmus oft zitiert wurden.

In einer Serie von sechs Theoremen zeigte Wu (1983), dass es relativ milde Bedingungen gibt, die einen GEM-Algorithmus dazu zwingen können, (i) dass die Log-Likelihoodwerte seiner Parameterfolge gegen ein *lokales Maximum* der Log-Likelihood konvergiert, und (ii)



dass die Parameterfolge selbst konvergiert. Alle Theoreme machen die folgenden Annahmen, ohne sie jeweils explizit aufzuführen.

**Voraussetzungen (Wu (1983))**:

- $\Omega \subseteq \mathbb{R}^d$ , d.h. der Parameterraum ist eine Teilmenge eines mehrdimensionalen euklidischen Raumes,
- $\{ \Theta \in \Omega \mid L(\Theta) \geq L(\Theta_0) \}$ ist für alle $\Theta_0 \in \Omega$ mit $L(\Theta) > -\infty$ *kompakt* (d.h. *abgeschlossen* und *beschränkt*),
- Die Log-Likelihood $L(.)$ ist in $\Omega$ stetig und im *Inneren* $\Omega \setminus \partial\Omega$ *des Parameterraums* differenzierbar.
- Die Parameterfolge eines EM-Algorithmus oder eines GEM-Algorithmus besteht aus lauter *inneren Punkten des Parameterraums*. Dies ist z.B. erfüllt, falls für jeden (Start-)Parameter gilt, dass:
$$\{ \Theta \in \Omega \mid L(\Theta) \geq L(\Theta_0) \} \subseteq \Omega \setminus \partial\Omega .$$

Die folgende Definition ist zentral, da auf ihr Theorem 1 (Wu 1983) beruht.

**Definition:** Sei $\mathcal{A} \subseteq \mathbb{R}^d$. Die Abbildung $\mathcal{M}: \mathcal{A} \to \mathcal{P}(\mathcal{A})$ heisst *abgeschlossen* an der Stelle $a \in \mathcal{A}$, falls aus

$$a_i \to a \quad (i \to \infty)$$

und

$$b_i \to b \quad (i \to \infty)$$

mit $a_i \in \mathcal{A}$ und $b_i \in \mathcal{M}(a_i)$ folgt, dass:

$$b \in \mathcal{M}(a) .$$

Das folgende Theorem ist die allgemeinste bekannte Aussage für einen EM-Algorithmus oder einen GEM-Algorithmus. Es zeigt, dass Log-Likelihoodwerte der erzeugten Parameterfolge unter milden Voraussetzungen gegen einen stationären Wert der Log-Likelihood konvergieren und unter strengeren Voraussetzungen sogar gegen ein lokales Maximum.

**Theorem (Theorem 1 aus Wu (1983)):** Sei $\mathcal{S}$ (bzw. $\mathcal{M}ax$) die Menge der stationären Punkte (bzw. lokalen Maxima) der Log-Likelihood, die im Inneren $\Omega \setminus \partial\Omega$ des Parameterraums liegen. Sei

$$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \dots$$

die Parameterfolge eines GEM-Algorithmus, sodass:

$$\Theta_i \in \mathcal{M}(\Theta_{i-1}) \qquad (i = 1, 2, 3, \dots)\ .$$

Hierbei sei (i) $\mathcal{M}(.)$ eine *abgeschlossene* Abbildung über dem Komplement von $\mathcal{S}$ (bzw. $\mathcal{M}ax$), und (ii)

$$L(\Theta_i) > L(\Theta_{i-1}) \qquad \text{für alle } \Theta_{i-1} \notin \mathcal{S} \text{ (bzw. } \mathcal{M}ax)\ .$$

Dann gilt:

- Jeder *Häufigkeitspunkt* (Grenzwert einer konvergeten Teilfolge) der Parameterfolge

  $$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \dots$$

  ist ein *stationärer Punkt* (bzw. lokales Maximum) der Log-Likelihood.

- Die zu der Parameterfolge gehörige Folge der Log-Likelihoodwerte

  $$L(\Theta_0),\ L(\Theta_1),\ L(\Theta_2),\ L(\Theta_3),\ \dots$$

  konvergiert monoton wachsend gegen ein:

  $$L^* = L(\Theta^*)$$

  mit $\Theta^* \in \mathcal{S}$ (bzw. $\mathcal{M}ax$).

**Beweis:** Anwendung eines Theorems von Zangwill (1969), Seite 91.

Für einen EM-Algorithmus (nicht aber für einen GEM-Algorithmus) liegt Konvergenz gegen stationäre Punkte (nicht aber gegen lokale Maxima) schon unter sehr schwachen Bedingungen vor. Um dies zu zeigen, benötigt man den folgenden Schlüsselschritt:

**Lemma**: Für alle Parameter $\Phi \notin \mathcal{S}$ gilt in einem EM-Algorithmus:

$$L(\mathcal{M}(\Phi)) > L(\Phi)\ .$$



**Beweis:** Es reicht zu zeigen, dass eine der beiden Eigenschaften in Theorem 1 aus Dempster et al. (1977) verletzt sind. Sei hierzu $\Theta \in \Omega$ zunächst ein beliebiger Parameter. Mit dem Lemma über die Log-Likelihood und die EM-Hilfsfunktion folgt durch Differenzieren, sowie durch Anwendung von Lemma 2 aus Dempster et al. (1977):

$$\nabla_\Theta L(\Theta)|_{\Theta=\Phi} \ = \ \nabla_\Theta Q_\Phi(\Theta)|_{\Theta=\Phi}\ .$$

Für $\Phi \notin \mathcal{S}$ folgt hieraus (und per Definintion von $\mathcal{S}$):

$$\nabla_\Theta Q_\Phi(\Theta)|_{\Theta=\Phi} \neq 0\ .$$

Ein allgemeiner Satz aus der Differentialrechnung mehrerer Veränderlicher (Erwe 1962) besagt daher, dass $\Phi$ kein *lokales Extremum*, d.h. insbesondere kein lokales Maximum, der EM-Hilfsfunktion $Q_\Phi$ sein kann. Als *innerer Punkt* des Parameterraums $\Omega$ liegt $\Phi$ auch nicht auf dem Parameterrand $\partial\Omega$. Da in einem EM-Algorithmus (per Definition der EM-Parameter-Abbildung) genau eine dieser zwei Eigenschaften auf $\mathcal{M}(\Phi)$ zutrifft, folgt:

$$\mathcal{M}(\Phi) \neq \Phi$$

und:

$$Q_\Phi(\mathcal{M}(\Phi)) > Q_\Phi(\Phi)\ .$$

Also ist die Eigenschaft (ii) in Theorem 1 aus Dempster et al. (1977) verletzt. Es folgt die Behauptung. **q.e.d.**

**Theorem (Theorem 2 aus Wu (1983)):** Die EM-Hilfsfunktion $Q_\Phi(\Theta)$ sei sowohl *stetig in* $\Theta$, als auch *stetig in* $\Phi$. Dann gilt für einen EM-Algorithmus:

- Jeder *Häufigkeitspunkt* der Parameterfolge

  $$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \dots$$

  ist ein *stationärer Punkt* der Log-Likelihood.

- Die zu einer Parameterfolge gehörige Folge der Log-Likelihoodwerte

  $$L(\Theta_0),\ L(\Theta_1),\ L(\Theta_2),\ L(\Theta_3),\ \dots$$

  konvergiert monoton wachsend gegen einen *stationären Wert der Log-Likelihood*

  $$L^* = L(\Theta^*)\ .$$

**Beweis:** Während Voraussetzung (i) in Theorem 1 aus Wu (1983) aus der Stetigkeit der EM-Hilfsfunktionen folgt, ist Voraussetzung (ii) für nicht-stationäre Punkte wegen des vorangegangenen Lemmas erfüllt. **q.e.d.**

Leider ist es (ohne zusätzliche Bedingungen zu benutzen) nicht möglich, einen analogen Beweis für lokale Maxima zu führen. Es ist bemerkenswert, dass die Resultate aus Theorem 2 bereits von Baum et al. (1970) für spezielle Modelle vorgestellt wurden. Die Resultate von Theorem 1 wurden hingegen nicht erfasst, da in Baum et al. (1970) schärfer vorausgesetzt wurde, dass $\mathcal{M}(.)$ stetig ist.

Theorem 4 ist ein einfacher Spezialfall des folgenden Theorems und wird deshalb hier nicht vorgestellt. Theorem 5 stellt für einen GEM-Algorithmus wahrscheinlich immer noch die einzige bekannte Aussage zur Parameter-Konvergenz gegen ein lokales Maximum der Log-Likelihooddar.

**Theorem (Theorem 5 aus Wu (1983)):** Sei

$$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots$$

die Parameterfolge eines GEM-Algorithmus, welche die Voraussetzungen (i) und (ii) aus dem Theorem 1 von Wu (1983) erfüllt. Dann gilt

$$\Theta_i \to \Theta^* \qquad (i \to \infty)\ ,$$

wobei

$$L(\Theta_i) \to L^* \qquad (i \to \infty)\ ,$$

und

$$\Theta^* \in \mathcal{S}(L^*) \ \ (\text{ bzw. } \mathcal{M}\text{ax}(L^*)\ )\ \ ,$$

vorausgesetzt, dass

$$||\Theta_i - \Theta_{i-1}|| \to 0 \qquad (i \to \infty)$$

und

$$\mathcal{S}(L^*) = \{\ \Theta \in \mathcal{S}\ |\ L(\Theta) = L^*\ \}\ ,$$

$$(\text{ bzw. } \quad \mathcal{M}\text{ax}(L^*) = \{\ \Theta \in \mathcal{M}\text{ax}\ |\ L(\Theta) = L^*\ \}\ )$$

ist eine *diskrete Menge.*

**Beweis:** Da eine allgemeine Annahme ist, dass $\{\ \Theta \in \Omega\ |\ L(\Theta) \geq L(\Theta_0)\ \}$ *kompakt* ist, folgt insbesondere, dass die Parameterfolge

$$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots$$

beschränkt ist. Gemäss Theorem 28.1 (Ostrowski 1966) bilden die Häufigkeitspunkte einer beschränkten Folge mit

$$||\Theta_i - \Theta_{i-1}|| \to 0 \qquad (i \to \infty)$$



eine kompakte und zusammenhängende Menge. Gemäss Theorem 1 (Wu 1983) liegen diese Häufungspunkte in $\mathcal{S}(L^*)$ (bzw. $\mathcal{M}\text{ax}(L^*)$). Da weiter angenommen wird, dass diese Menge diskret ist, kann es nur einen einzigen Häufungspunkt geben, d.h. die Parameterfolge konvergiert. **q.e.d.**

Analog zu Theorem 2 benötigt man für die Parameter-Konvergenz gegen stationäre Punkte nur relativ milde Voraussetzungen.

**Theorem (Theorem 6 aus Wu (1983)):** Die EM-Hilfsfunktion $Q_\Phi(\Theta)$ sei *differenzierbar in* $\Theta$, und $\nabla_\Theta\ Q_\Phi(\Theta)$ sei sowohl *stetig in* $\Theta$, als auch *stetig in* $\Phi$. Sei

$$\Theta_0,\ \Theta_1,\ \Theta_2,\ \Theta_3,\ \ldots$$

die Parameterfolge eines GEM-Algorithmus, sodass jeder Parameter $\Theta_i$ ein *stationärer Punkt* der EM-Hilfsfunktion $Q_{\Theta_{i-1}}$ ist:

$$\nabla_\Theta Q_{\Theta_{i-1}}(\Theta)\big|_{\Theta=\Theta_i} = 0\ .$$

Dann gibt es einen stationären Punkt $\Theta^*$ mit

$$\Theta_i \to \Theta^* \qquad (i \to \infty)\ ,$$

$$L(\Theta_i) \to L^* \qquad (i \to \infty)\ ,$$

vorausgesetzt

$$||\Theta_i - \Theta_{i-1}|| \to 0 \qquad (i \to \infty)$$

und

$$\mathcal{L}(L^*) = \{\ \Theta \in \Omega\ |\ L(\Theta) = L^*\ \}$$

ist eine *diskrete Menge.*

*Beweis:* Mit dem Theorem von Ostrowski (1966) kann wie im Beweis zu Theorem 5 (Wu 1983) gezeigt werden, dass:

$$\Theta_i \to \Theta^* \qquad (i \to \infty)\ ,$$

Die Stationarität von $\Theta^*$ folgt aus Behauptung (i) des Theorem 4 (Dempster et al. 1977). **q.e.d.**

**Neue Ergebnisse zum EM-Algorithmus**

**Definition (Injektive Parametrisierung der Analysen)**: Die Parametrisierung $p_\Theta(.|y)$ der Analysen $X(y)$ heisst *injektiv*, falls:

$$p_\Theta(.|y) = p_\Phi(.|y) \text{ für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \quad \Rightarrow \quad \Theta = \Phi \ .$$

**Korollar**: Der mit einem GEM-Algorithmus und einer *injektiven Parametrisierung der Analysen* re-estimierte Parameter $\mathcal{M}(\Theta)$ erfüllt

$$\mathcal{M}(\Theta) = \Theta \ .$$

oder

$$L(\mathcal{M}(\Theta)) > L(\Theta)$$

**Beweis:** Gemäss Theorem 1 (Dempster et al. 1977) gilt: $L(\mathcal{M}(\Theta)) \geq L(\Theta)$ . Für $L(\mathcal{M}(\Theta)) > L(\Theta)$ sind wir fertig. Sei also $L(\mathcal{M}(\Theta)) = L(\Theta)$ . Erneut mit Theorem 1 (Dempster et al. 1977) folgt

$$p_{\mathcal{M}(\Theta)}(.|y) = p_\Theta(.|y) \ \text{ für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \ .$$

Da angenommen wird, dass eine injektive Parametrisierung der Analysen vorliegt, folgt: $\mathcal{M}(\Theta) = \Theta$ . **q.e.d.**

**Korollar**: Sei eine *injektive Parametrisierung* $p_\Theta(.|y)$ *der Analysen* $X(y)$ gegeben. Dann gilt für alle Parameter $\Theta \neq \Phi \in \Omega$:

$$H_{\text{cross}}(\Theta;\ \Phi) > H(\Phi) \ .$$

D.h. die erwartete Cross-Entropie der Analysen $H_{\text{cross}}(\Theta;\ \Phi)$ besitzt als Funktion von $\Theta$ an der Stelle $\Theta = \Phi$ ein *strenges relatives Minimum.*

**Beweis:** Folgt aus dem Lemma über die erwartete Cross-Entropie und Entropie aller Analysen. **q.e.d.**

Es ist sehr interessant, dass die Injektivität der Parametrisierung der Analysen dasselbe Resultat liefern kann, wie die Eindeutigkeit der Maximum-Likelihood-Estimierung:



**Korollar**: Der mit einem GEM-Algorithmus und einer *injektiven Parametrisierung der Analysen* re-estimierte Parameter $\mathcal{M}(\hat{\Theta})$ einer Maximum-Likelihood-Estimierung $\hat{\Theta}$ ist *Parameter-invariant*:

$$\mathcal{M}(\hat{\Theta}) = \hat{\Theta} \ .$$

**Beweis:** Korollar 1 aus Dempster et al. (1977) liefert:

$$p_{\mathcal{M}(\hat{\Theta})}(.|y) = p_{\hat{\Theta}}(.|y) \ \text{ für alle } y \in \mathcal{Y} \text{ mit } \tilde{p}(y) > 0 \ .$$

Da angenommen wird, dass eine injektive Parametrisierung der Analysen vorliegt, folgt die Behauptung. **q.e.d.**