2.4 Estimierungsverfahren und empirische Wahrscheinlichkeiten

In Abschnitt 2.1 wurde ausgeführt, dass der wesentliche Aspekt einer Wahrscheinlichkeitsverteilung darin besteht, dass sie für jeden Datentyp die Information bereithält, mit welcher Wahrscheinlichkeit dieser Datentyp aus einer Grundmenge von Datentypen heraus erzeugt wird. Ferner wurde erwähnt, dass mit diesem Aspekt die Frage, wie oft ein linguistischer Datentyp in einem gegebenem Korpus vorkommt, sehr eng verbunden ist. In diesem Abschnitt wird angerissen, wie diese Verbindung möglicherweise aussehen könnte. Die vorsichtige Wortwahl soll darauf hinweisen, dass sich die Stochastiker in der Beantwortung dieser Frage leider uneinig sind.

Unabhängig von dieser Uneinigkeit der Stochastiker ist die Problemstellung die folgende:

Sei $\mathcal{Y}$ eine Grundmenge von linguistischen Datentypen, und $f(.)$ ein Korpus. Sei $\{p_\Theta(.) \mid \Theta \in \Omega\}$ eine Menge von *parametrisierten Wahrscheinlichkeitsverteilungen* über der gegebenen linguistischen Grundmenge $\mathcal{Y}$, wobei in diesem Zusammenhang eine Menge Ω *Parameterraum* und die $\Theta \in \Omega$ *Parameter* genannt werden. Im allgemeinen kann man sich die Parameter als hochdimensionale reellwertige Vektoren vorstellen, die in der Regel einer Reihe von simplen Nebenbedingungen unterworfen sind.

Die entscheidende Frage ist nun: unter der Vermutung, dass der Korpus mit irgendeiner der vorgegebenen Wahrscheinlichkeitsverteilungen $p_\Theta(.)$ generiert wurde, welcher der in Frage kommenden Parameter $\Theta \in \Omega$ ist der *wahre Parameter* $\hat{\Theta}$, oder anders gefragt: mit welchem expliziten Parameter $\hat{\Theta}$ wurde der Korpus vermutlich erzeugt?

Zur Beantwortung dieser Frage, d.h. zur Estimierung des fraglichen Parameters, stehen offensichtlich nur die gegebene Häufigkeitsverteilung $f(.)$ aller Datentypen im Korpus und die möglichen Wahrscheinlichkeitsverteilungen $p_\Theta(.), \Theta \in \Omega$, zur Verfügung. Das am häufigsten angewandte Estimierungsverfahren der Statistik ist die sogenannte *Maximum-Likelihood-Estimierung*. Der Name deutet an, dass die Maximum-Likelihood-Estimierung unter den in Frage kommenden Parametern $\Theta \in \Omega$ gerade den Parameter $\hat{\Theta}$ auswählt, der die Korpuswahrscheinlichkeit $p_\Theta(f)$ maximiert:

$$\hat{\Theta} = \operatorname{argmax}_{\Theta \in \Omega} \quad p_\Theta(f) \ .$$

Obwohl die Maximum-Likelihood-Estimierung unter den gegebenen Bedingungen intuitiv einleuchtend ist, und sie einige hervorragende mathematische Eigenschaften besitzt (sie ist eine konsistente, für grosse Korpora annähernd normalverteilte Estimierung, mit dem wahren Parameter als Erwartungswert und der inversen Fisher-Information als Varianz), ist sie doch nicht konkurrenzlos (Existenz und Eindeutigkeit der Maximum-Likelihood-Estimierung sind problematisch).

In DeGroot (1989) werden neben der Maximum-Likelihood-Estimierung wenigstens vier weitere Estimierungsverfahren angegeben: die Bayes-Estimierung, die Maximum-A-Posteriori-Estimierung, die Mean-Squared-Error-Estimierung, sowie eine auf Konfidenzintervallen beruhende Estimierung. Hierbei wird jede Abbildung von der Menge $\mathcal{F}_n(\mathcal{Y})$ aller Korpora einer festen Länge n auf den Parameterraum Ω eine *Estimierung* genannt:

$$\delta : \ \mathcal{F}_n(\mathcal{Y}) \longrightarrow \Omega, \quad f \mapsto \delta(f) \ .$$

Paradoxerweise teilt sich die Gemeinschaft der Stochastiker an der Frage der Maximum-A-Posteriori-Estimierung. Paradox ist dies aus mehreren Gründen. Zunächst einmal deswegen, weil die Maximum-A-Posteriori-Estimierung ohne ein zusätzliches Inventar überhaupt nicht auf das obige Estimierungsproblem angewandt werden kann. Bei diesem zusätzlichen Inventar handelt es sich um die sogenannte *A-Priori-Wahrscheinlichkeitsverteilung*, eine (in der Regel nicht-diskrete) Wahrscheinlichkeitsverteilung auf dem Parameterraum:

$$p : \ \Omega \to [0,1], \quad \Theta \mapsto p(\Theta) \ .$$

Anstatt wie bei der Maximum-Likelihood-Estimierung die Korpuswahrscheinlichkeit $p_\Theta(f)$ zu maximieren, wird bei der *Maximum-A-Posteriori-Estimierung* das Produkt aus A-Priori-Wahrscheinlichkeit und Korpuswahrscheinlichkeit maximiert:

$$\hat{\Theta} = \operatorname{argmax}_{\Theta \in \Omega} \quad p(\Theta) \cdot p_\Theta(f) \ .$$

Da die Produkt-Wahrscheinlichkeit

$$p(\Theta, f) := p(\Theta) \cdot p_\Theta(f)$$

eine (in der Regel nicht-diskrete) parametrisierte Wahrscheinlichkeitsverteilung auf dem Produkt-Raum $\Omega \times \mathcal{F}_n(\mathcal{Y})$ ist:

$$\sum_{\langle\Theta, f\rangle \in \Omega \times \mathcal{F}_n(\mathcal{Y})} p(\Theta, f) = \sum_{\Theta \in \Omega} p(\Theta) \sum_{f \in \mathcal{F}_n(\mathcal{Y})} p_\Theta(f) = \sum_{\Theta \in \Omega} p(\Theta) \cdot 1 = 1 \ ,$$

kann die Maximum-A-Posteriori-Estimierung aus formaler Sicht als eine Maximum-Likelihood-Estimierung interpretiert werden (was tatsächlich nur bedeutet, dass ein Parameter mittels des Kriteriums der Maximierung einer parametrisierten Wahrscheinlichkeit ausgewählt wird). Umgekehrt kann eine Maximum-Likelihood-Estimierung offensichtlich als eine Maximum-A-Posteriori-Estimierung mit gleichverteilter/uniformer A-Priori-Wahrscheinlichkeitsverteilung angesehen werden.

Trotz der formalen Gleichrangigkeit von Maximum-A-Posteriori-Estimierung und Maximum-Likelihood-Estimierung entbrannte unter den Stochastikern ein Streit um die Akzeptanz der Maximum-A-Posteriori-Estimierung mit mehr oder weniger trivialem Hintergrund: Es scheint klar zu sein, dass eine "A-Priori-Bevorzugung" eines Parameters eine A-Posteriori-Estimierung genau dieses Parameters provoziert. Die Frage ist nun, ob deshalb eine A-Priori-Wahrscheinlichkeitsverteilung in eine Parameter-Estimierung einfliessen darf? Wenn ja, gibt es eine Methode mit der die A-Priori-Wahrscheinlichkeitsverteilung aus objektiven Wissensquellen gewonnen werden kann? Obwohl die Diskussion um das beste Estimierungsverfahren sicher fruchtbar ist, geht sie doch über den Rahmen dieser Arbeit weit hinaus. So soll an dieser Stelle nur darauf hingewiesen werden, dass in dieser Arbeit der EM-Algorithmus auf Grundlage der Maximum-Likelihood-Estimierung vorgestellt wird, dass aber einige Varianten des EM-Algorithmus, zum Beispiel sogenanntes "Penalized EM", auf der Maximum-A-Posteriori-Estimierung beruhen.

Im folgenden soll ein Beispiel vorgestellt werden, mit dem zur einleitenden Bemerkung zurückgekehrt wird: Der wesentliche Aspekt einer Wahrscheinlichkeitsverteilung besteht darin, dass sie für jeden Datentyp die Information bereithält, mit welcher Wahrscheinlichkeit dieser Datentyp aus einer Grundmenge von Datentypen heraus erzeugt wird. Es soll nun gezeigt werden, dass hiermit die Frage, wie oft ein linguistischer Datentyp in einem gegebenen Korpus vorkommt, sehr eng verbunden ist. Das Problem wird angegangen, indem von einem gegebenen Korpus $f(.)$ von Datentypen aus $\mathcal{Y}$ ausgegangen wird. Die zu beantwortende Frage ist, welche Wahrscheinlichkeitsverteilung $p(.)$ auf $\mathcal{Y}$ diesen Korpus angemessen repräsentiert. Der Abschnitt 2.1 legt nahe, dass dies die empirische Wahrscheinlichkeitsverteilung $\tilde{p}(.)$ sein wird. Mit der Maximum-Likelihood-Estimierung soll dies nun bewiesen werden. Die Maximum-Likelihood-Estimierung hat auf die Frage der "angemessenen Repräsentation" die Antwort, dass es diejenige Wahrscheinlichkeitsverteilung ist, welche die Korpuswahrscheinlichkeit maximiert:

$$\hat{p} = \mathrm{argmax}_p \quad p(f)\ .$$

Der Abschnitt 2.3 gibt weitere Antworten: Die Maximierung der Korpuswahrscheinlichkeit ist äquivalent zur Maximierung der Log-Likelihood $L(p;\tilde{p})$ oder zur Minimierung der Cross-Entropie $H_{\text{cross}}(p;\tilde{p})$, oder der Kullback-Leibler-Distanz $D(\tilde{p}\|p)$ oder der Korpusperplexität perplexity$(p,\tilde{p})$. In der Regel wird für das weitere Vorgehen die Maximierung der Log-Likelihood ausgewählt:

$$\begin{aligned}
\hat{p} &= \mathrm{argmax}_p \quad L(p;\tilde{p}) \\
&= \mathrm{argmax}_p \quad \tilde{p}[\log\ p] \\
&= \mathrm{argmax}_p \quad \sum_{y\in\mathcal{Y}} \tilde{p}(y)\cdot\log\ p(y)
\end{aligned}$$

Zur Lösung dieses *Optimierungsproblems* ist es zweckmässig, eine *Parametrisierung* der Wahrscheinlichkeitsverteilung $p(.)$ vorzunehmen:

$$p(y) := \Theta_y \qquad \forall y \in \mathcal{Y}\ .$$

Weil $p(.)$ eine beliebige Wahrscheinlichkeitsverteilung ist, kann über die *Parameter*

$$\Theta_y \in [0,1]$$

nichts weiter ausgesagt werden, als dass der *Parametervektor*

$$\Theta := < \Theta_y \mid y \in \mathcal{Y} >$$

ein Element des folgenden *Parameterraums* ist:

$$\Omega := \left\{ \Theta \in [0,1]^{|\mathcal{Y}|} \;\middle|\; g(\Theta) = 0 \right\}\ .$$

Die *Nebenbedingung* in der Definition des Parameterraums, mit

$$g(\Theta) := 1 - \sum_{y\in\mathcal{Y}} \Theta_y\ ,$$

ist offenbar nötig, damit die Parametrisierung für jede Wahl von $\Theta \in \Omega$ eine Wahrscheinlichkeitsverteilung $p(.)$ auf $\mathcal{Y}$ liefert:

$$\sum_{y\in\mathcal{Y}} p(y) = \sum_{y\in\mathcal{Y}} \Theta_y = 1 - g(\Theta) = 1 - 0 = 1\ .$$

Weitere Nebenbedingungen würden den Parameterraum und damit die Menge der parametrisierten Wahrscheinlichkeitsverteilungen weiter einschränken; was für andere, als die hier zu lösende Fragestellung, durchaus sinnvoll sein kann.

Offensichtlich ist die Wahrscheinlichkeitsverteilung $p(.)$ durch die vorgenommene Parametrisierung von dem Parametervektor Θ abhängig geworden, was wie folgt notiert werden soll:

$$p(.) = p_\Theta(.)\ .$$

Dasselbe gilt dann aber auch für die Log-Likelihood, was ebenfalls explizit notiert wird:

$$L(\Theta) := L(p_\Theta;\tilde{p})\ .$$

Leider hat das Auftreten der Nebenbedingung $g(\Theta) = 0$ zur Folge, dass das zu lösende Optimierungsproblem, welches sich mittlerweise in der Form:

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

präsentiert, nur mit einem tiefen Satz aus der *Differentialrechnung mehrerer Veränderlichen* angegangen werden kann. Im folgenden Satz von J.L. Lagrange, der als sogenannte *Multiplikatorregel* bekannt ist (Erwe 1962), wird eine **notwendige** Bedingung für das Vorliegen eines sogenannten *Extremum mit Nebenbedingungen* angegeben:

Theorem (Multiplikatorregel von J.L. Lagrange): Seien $L(\Theta)$ und $g_1(\Theta)\ldots g_N(\Theta)$ in einer Umgebung eines Punktes $\hat{\Theta} \in [0,1]^{|\mathcal{Y}|}$ stetig differenzierbare reellwertige Abbildungen ($N < |\mathcal{Y}|$). Die *Funktionalmatrix*:

$$\begin{pmatrix} \nabla_\Theta\ g_1 \\ \vdots \\ \nabla_\Theta\ g_N \end{pmatrix}$$

besitze am Punkt $\hat{\Theta}$ maximalen Rang, während

$$g_1(\hat{\Theta}) = \ldots = g_N(\hat{\Theta}) = 0$$

ist. Wenn dann $L(\hat{\Theta})$ Extremum der Werte von $L(\Theta)$ bei Beschränkung auf solche Θ ist, für die:

$$g_1(\Theta) = \ldots = g_N(\Theta) = 0$$

ist, dann gibt es Zahlen $\lambda_1 \ldots \lambda_N$, sodass in $\hat{\Theta}$ gilt:

$$\frac{\partial\ L}{\partial\ \Theta_y} + \lambda_1 \cdot \frac{\partial\ g_1}{\partial\ \Theta_y} + \ldots \lambda_N \cdot \frac{\partial\ g_N}{\partial\ \Theta_y} = 0 \qquad \forall y \in \mathcal{Y}\ .$$

Bemerkung: Der *Gradient* ∇_Θ ist der Vektor der *ersten partiellen Ableitungen* einer Funktion nach allen ihren Parameter-Argumenten:

$$\nabla_\Theta\ L := \left\langle \frac{\partial\ L}{\partial\ \Theta_y} \middle|\ y \in \mathcal{Y} \right\rangle\ .$$

Um die Extrema einer Funktion $L(\Theta)$ unter den Nebenbedingungen $g_1(\Theta) = \ldots = g_N(\Theta) = 0$ zu ermitteln, hat man wie folgt vorzugehen: Von der Funktion $L(\Theta)$ geht man mittels noch unbekannter reeller Zahlen $\lambda_1 \ldots \lambda_N$ (sogenannten *Multiplikatoren*) zu der sogenannten *Lagrangefunktion*:

$$\Lambda(\Theta) := L(\Theta) + \lambda_1 \cdot g_1(\Theta) + \ldots \lambda_N \cdot g_N(\Theta)$$

über und tut so, als müsste man deren Extrema bestimmen, dann führt dies zu den in der Multiplikatorregel angegebenen $|\mathcal{Y}|$ Gleichungen. Diese ergeben zusammen mit den N Nebenbedingungen ein System von $|\mathcal{Y}| + N$ Gleichungen für die $|\mathcal{Y}| + N$ Unbekannten $\Theta_y, (y \in \mathcal{Y})$ und $\lambda_1 \ldots \lambda_N$. Höchstens die als Lösungen dieses Gleichungssystems auftretenden Θ, kommen als Stellen in Frage, in denen $L(.)$ ein Extremum mit jenen Nebenbedingungen besitzt. Eine solche Stelle braucht allerdings nicht stets wirklich ein derartiges

Extremum zu liefern, nicht einmal ein "relatives". Die Methode wird besonders in denjenigen Fällen nützlich sein, in denen aus anderen Gründen die Existenz eines Extremums sichergestellt ist und nur noch die Berechnung aussteht.

Die Anwendung der Multiplikatorregel auf die Log-Likelihood ergibt offenbar die folgende Lagrangefunktion ($N = 1$):

$$\begin{aligned} \Lambda(\Theta) &:= L(\Theta) + \lambda \cdot g(\Theta) \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log\ \Theta_y + \lambda \cdot \left(1 - \sum_{y \in \mathcal{Y}} \Theta_y\right)\ . \end{aligned}$$

Die ersten partiellen Ableitungen sind ($y \in \mathcal{Y}$):

$$\frac{\partial\ \Lambda}{\partial\ \Theta_y} := \frac{\tilde{p}(y)}{\Theta_y} - \lambda\ .$$

Das Gleichungssystem aus der Multiplikatorregel erhält man, indem die ersten partiellen Ableitungen der Lagrangefunktion gleich Null gesetzt werden ($y \in \mathcal{Y}$):

$$0 = \frac{\tilde{p}(y)}{\hat{\Theta}_y} - \lambda\ .$$

Dies liefert $\hat{\Theta}_y = \frac{\tilde{p}(y)}{\lambda}$. Die Nebenbedingung liefert noch, dass $\lambda = 1$ ist, d.h. es gilt $\hat{\Theta}_y = \tilde{p}(y)$. Zusammenfassend erhält man die folgende mögliche Lösung der Ausgangsfragestellung:

$$\begin{aligned} \hat{p}(y) &= p_{\hat{\Theta}}(y) \\ &= \hat{\Theta}_y \\ &= \tilde{p}(y)\ . \end{aligned}$$

D.h. die Vermutung erhärtet sich, dass die empirische Wahrscheinlichkeitsverteilung diejenige Wahrscheinlichkeitsverteilung ist, die die Korpuswahrscheinlichkeit maximiert und damit den Korpus am angemessensten repräsentiert. Leider ist die Bemerkung, die sich der Multiplikatorregel anschliesst, sehr ernst zu nehmen: Parameter, die mit der Multiplikatorregel errechnet werden, können Extrema sein, müssen es aber nicht sein; hierfür gibt es in der Literatur eine reiche Anzahl von Beispielen. Zudem weiss man (ohne zusätzliche Berechnungen) nicht, welche Art von Extremum, Minimum oder Maximum, berechnet wurde. Deshalb soll im folgenden ein alternatives Verfahren vorgestellt werden, mit dem man in die Lage versetzt wird, die Extrema der Log-Likelihood zu berechnen, zusätzlich aber auch nachprüfen kann, ob ein Minimum oder Maximum vorliegt. Leider ist auch für dieses Verfahren ein tiefer Satz aus der Differentialrechnung mehrerer Veränderlichen nötig. Er gibt hinreichende Bedingungen für das Vorliegen von Extrema an und wurde wieder Erwe (1962) entnommen.

Theorem: Wenn $L(\Theta)$ zweimal stetig differenzierbar nach allen ihren Variablen ist und in $\hat{\Theta}$

$$\nabla_\Theta L = 0$$

gilt, und die Matrix

$$\left(\frac{\partial^2 L}{\partial\Theta_y \partial\Theta_{y'}}\right)_{y,y' \in \mathcal{Y}}$$

in $\hat{\Theta}$ *negativ definit* ist, so hat $L(\Theta)$ in $\hat{\Theta}$ ein ***strenges** relatives Maximum.* Ist die Matrix hingegen *positiv definit*, so hat $L(\Theta)$ in $\hat{\Theta}$ ein ***strenges** relatives Minimum.*

Bemerkung: Man sagt, $L(\Theta)$ besitze in $\hat{\Theta}$ ein ***strenges** relatives Maximum*, wenn es eine Umgebung $\mathcal{U}$ von $\hat{\Theta}$ gibt, sodass:

$$L(\Theta) < L(\hat{\Theta}) \qquad \forall \Theta \in \mathcal{U} \ .$$

Entsprechendes gilt für ein ***strenges** relatives Minimum.* Man sagt ferner, eine *symmetrische Matrix*

$$\mathcal{M} = \left(m_{\ \Theta_y \Theta_{y'}}\right)_{y,y' \in \mathcal{Y}}$$

ist *positiv definit*, falls für beliebige reelwertige Vektoren $\Theta \neq 0$ gilt, dass:

$$\Theta^t \cdot \mathcal{M} \cdot \Theta \ > \ 0 \ .$$

Entsprechendes gilt für *negativ definite* Matrizen.

Im folgenden wird vorgestellt, wie dieses Theorem benutzt werden kann, um zu zeigen, dass die empirische Verteilung $\tilde{p}(.)$ tatsächlich die Korpuswahrscheinlichkeit maximiert. Die geradlinigste Vorgehensweise ist, eine Parametrisierung herzustellen, die ohne Nebenbedingungen auskommt. Mit einer Parametrisierung ohne Nebenbedingungen sollte es möglich sein, mit dem ersten Teil des Theorems, mögliche Extrema zu berechnen. Mit dem zweiten Teil des Theorems sollte verifizierbar sein, welche der möglichen Extrema tatsächlich Maxima sind. Die skizzierte Vorgehensweise wird erfolgreich sein. Obwohl das gesamte Verfahren sehr einfach ist, ist dem Autor nicht bekannt, ob es schon einmal irgendwo in der Literatur angewandt wurde. In Abschnitt 3.8 wird es im Rahmen des EM-Algorithmus erneut für kontextfreie Grammatiken benutzt werden.

Der Schlüsselschritt ist, eine Parametrisierung zu konstruieren, die ohne Nebenbedingungen auskommt. Hierfür wird in der Menge $\mathcal{Y}$ der Datentypen zunächst ein beliebiger Datentyp y_0 markiert, und danach $\mathcal{Y}$ wie folgt aufgespaltet:

$$\mathcal{Y} = \{y_0\} + \mathcal{Y}_0 \ .$$

(Das '+'-Zeichen steht für die disjunkte Vereinigung von Mengen.) Die gewünschte Parametrisierung wird folgendermassen vorgenommen:

$$p_\Theta(y) := \begin{cases} 1 - \sum_{y \in \mathcal{Y}_0} \Theta_y & \text{falls } y = y_0 \\ \Theta_y & \text{falls } y \in \mathcal{Y}_0 \end{cases}$$

Unten wird sich zeigen, dass ausserdem $\tilde{p}(y_0) \neq 0$ gelten muss. Da der Korpus nicht leer ist, muss es natürlich einen solchen Datentyp geben. Die Parametrisierung kommt mit $|\mathcal{Y}_0| = |\mathcal{Y}| - 1$ Parametern Θ_y, $(y \in \mathcal{Y}_0)$ aus. Für die neue Parametrisierung wird im Vergleich zu der alten Parametrisierung also ein Parameter weniger benötigt. Ferner ist für sie keine Nebenbedingung erforderlich, weil:

$$\begin{aligned} \sum_{y \in \mathcal{Y}} p_\Theta(y) &= p_\Theta(y_0) + \sum_{y \in \mathcal{Y}_0} p_\Theta(y) \\ &= \left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right) + \sum_{y \in \mathcal{Y}_0} \Theta_y \\ &= 1 \ . \end{aligned}$$

Der Parameterraum ist daher von der einfachen Gestalt[13]:

$$\Omega_0 := [0,1]^{|\mathcal{Y}_0|} \ .$$

Die Log-Likelihood ergibt sich für diese Parametrisierung, wie folgt:

$$\begin{aligned} L_0(\Theta) &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log\ p_\Theta(y) \\ &= \tilde{p}(y_0) \cdot \log\ p_\Theta(y_0) + \sum_{y \in \mathcal{Y}_0} \tilde{p}(y) \cdot \log\ p_\Theta(y) \\ &= \tilde{p}(y_0) \cdot \log \left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right) + \sum_{y \in \mathcal{Y}_0} \tilde{p}(y) \cdot \log\ \Theta_y \ . \end{aligned}$$

Vergleicht man die neue Log-Likelihood $L_0(\Theta)$ mit der Lagrangefunktion $\Lambda(\Theta)$ für die alte Parametrisierung, so fällt die grosse Ähnlichkeit dieser beiden Funktionen auf: Der wesentliche Unterschied ist, dass $L_0(\Theta)$ einen Term $\log\left(1 - \sum_y \Theta_y\right)$ enthält, und $\Lambda(\Theta)$ den einfacheren Term $\left(1 - \sum_y \Theta_y\right)$. Es wird sich im folgenden zeigen, dass dieser Unterschied nur zu einem unwesentlich grösseren Aufwand zur Berechnung der Extrema führt[14].

[13] Natürlich muss zusätzlich gewährleistet sein, dass

$$0 \leq p_\Theta(y_0) = 1 - \sum_{y \in \mathcal{Y}_0} \Theta_y \leq 1 \ .$$

Diese Nebenbedingung kann jedoch nach der Optimierung der Parameter überprüft werden und wird dann in den meisten Fällen erfüllt sein. Im Gegensatz hierzu hätte für die alte Parametrisierung eine Optimierung ohne Nebenbedingung keinesfalls Parameter geliefert, die im nachhinein $g(\Theta) = 0$ erfüllt hätten.

[14] Der zusätzliche Rechenaufwand, der sich im Vergleich zur alten Parametrisierung ergibt, wird geleistet, um zu zeigen, dass die errechneten Extrema Maxima sind; was mit der alten Parametrisierung prinzipiell unmöglich ist.

Der erste Teil des Theorems verlangt, dass der Gradient der Log-Likelihood $L_0(\Theta)$ am Extremum gleich Null ist. Die ersten partiellen Ableitungen von $L_0(\Theta)$ sind ($y \in \mathcal{Y}_0$):

$$\frac{\partial\ L_0}{\partial\ \Theta_y} = \tilde{p}(y_0) \cdot \frac{-1}{\left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right)} + \tilde{p}(y) \cdot \frac{1}{\Theta_y} \ .$$

Setzt man diese gleich Null, erhält man das folgende lineare Gleichungssystem mit $|\mathcal{Y}_0|$ Gleichungen und $|\mathcal{Y}_0|$ Unbekannten ($y \in \mathcal{Y}_0$):

$$\left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right) \cdot \tilde{p}(y) = \tilde{p}(y_0) \cdot \Theta_y \ .$$

Man entnimmt den Gleichungen, dass sich Θ_y und $\tilde{p}(y)$ proportional zueinander verhalten; die einzige Lösung ist daher ($y \in \mathcal{Y}_0$):

$$\hat{\Theta}_y = \tilde{p}(y) \ .$$

Der zweite Teil des Theorems verlangt die Berechnung der Matrix:

$$\left(\frac{\partial^2 L_0}{\partial\Theta_y \partial\Theta_{y'}}\right)_{y,y' \in \mathcal{Y}_0}$$

an der Stelle $\hat{\Theta}$. Für $y, y' \in \mathcal{Y}_0$ gilt:

$$\begin{aligned}
\frac{\partial^2 L_0}{\partial\Theta_y \partial\Theta_{y'}} &= \frac{\partial}{\partial\ \Theta_y}\left(\tilde{p}(y_0) \cdot \frac{-1}{\left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right)} + \tilde{p}(y') \cdot \frac{1}{\Theta_{y'}}\right) \\
&= \tilde{p}(y_0) \cdot \frac{-1}{\left(1 - \sum_{y \in \mathcal{Y}_0} \Theta_y\right)^2} + \begin{cases} 0 & \text{falls } y \neq y' \\ -\tilde{p}(y) \cdot \frac{1}{\Theta_y^2} & \text{falls } y = y' \end{cases}
\end{aligned}$$

An der Stelle $\hat{\Theta}$ erhält man daher:

$$\begin{aligned}
\left.\frac{\partial^2 L_0}{\partial\Theta_y \partial\Theta_{y'}}\right|_{\hat{\Theta}=\tilde{p}} &= \tilde{p}(y_0) \cdot \frac{-1}{\left(1 - \sum_{y \in \mathcal{Y}_0} \tilde{p}(y)\right)^2} + \begin{cases} 0 & \text{falls } y \neq y' \\ -\tilde{p}(y) \cdot \frac{1}{\tilde{p}(y)^2} & \text{falls } y = y' \end{cases} \\
&= \begin{cases} -\frac{1}{\tilde{p}(y_0)} & \text{falls } y \neq y' \\ -\frac{1}{\tilde{p}(y_0)} - \frac{1}{\tilde{p}(y)} & \text{falls } y = y' \end{cases}
\end{aligned}$$

Die Matrix:

$$\left(\frac{\partial^2 L_0}{\partial\Theta_y \partial\Theta_{y'}}\right)_{y,y' \in \mathcal{Y}_0}$$

lässt sich an der Stelle $\hat{\Theta}$ also als Summe $\mathcal{C} + \mathcal{D}$ zweier Matrizen $\mathcal{C}$ und $\mathcal{D}$ darstellen, wobei die Matrix $\mathcal{C}$ eine *konstante Matrix* mit den Matrizenelementen $-\frac{1}{\tilde{p}(y_0)}$, und die Matrix $\mathcal{D}$

eine *Diagonalmatrix* mit den Diagonalelementen $-\frac{1}{\tilde{p}(y)}$, ($y \in \mathcal{Y}_0$) ist. Im folgenden wird gezeigt, dass beide Matrizen $\mathcal{C}$ und $\mathcal{D}$ negativ definit sind. Für beliebige reellwertige Vektoren $\Theta \neq 0$ gilt:

$$\begin{aligned}
\Theta^t \cdot \mathcal{C} \cdot \Theta &= \left(\sum_{y \in \mathcal{Y}_0} \Theta_y\right) \cdot \frac{-1}{\tilde{p}(y_0)} \cdot \left(\sum_{y \in \mathcal{Y}_0} \Theta_y\right) \\
&= -\frac{1}{\tilde{p}(y_0)} \cdot \left(\sum_{y \in \mathcal{Y}_0} \Theta_y\right)^2 \\
&< 0 \ ,
\end{aligned}$$

vorausgesetzt, dass die empirische Wahrscheinlichkeit des markierten Elements nicht gleich Null ist. Solch ein Element muss aber existieren, denn sonst wäre der Korpus leer. Ferner gilt:

$$\begin{aligned}
\Theta^t \cdot \mathcal{D} \cdot \Theta &= \sum_{y \in \mathcal{Y}_0} \Theta_y \cdot \left(-\frac{1}{\tilde{p}(y_0)} - \frac{1}{\tilde{p}(y)}\right) \cdot \Theta_y \\
&= \sum_{y \in \mathcal{Y}_0} \left(-\frac{1}{\tilde{p}(y_0)} - \frac{1}{\tilde{p}(y)}\right) \cdot \Theta_y^2 \\
&< 0 \ ,
\end{aligned}$$

mit derselben Begründung. Es folgt, dass sowohl die Matrix $\mathcal{C}$ als auch die Matrix $\mathcal{D}$ negativ definit ist. Da die Summe zweier negativ definiter Matrizen ebenfalls negativ definit ist, ist alles gezeigt. Gemäss Theorem folgt, dass:

$$\hat{\Theta}_y = \tilde{p}(y), \qquad (y \in \mathcal{Y}_0),$$

ein strenges relatives Maximum der Log-Likelihood $L_0(\Theta)$ ist.

Da sehr einfach zu zeigen ist, dass $L_0(\Theta)$ auf dem *Rand* des Parameterraums $\Omega_0 = [0,1]^{|\mathcal{Y}_0|}$ nur den Wert $-\infty$ annimmt, welcher nicht grösser als $L_0(\hat{\Theta})$ ist, ist $\hat{\Theta} = \tilde{p}$ sogar ein *globales Maximum* der Log-Likelihood $L_0(\Theta)$.

Damit konnte das angekündigte Ziel bewiesen werden, dass die empirische Wahrscheinlichkeitsverteilung einen Korpus am besten repräsentiert: Unter allen Wahrscheinlichkeitsverteilungen ist es die empirische Wahrscheinlichkeitsverteilung, welche die Log-Likelihood (bzw. die Korpuswahrscheinlichkeit) maximiert:

$$\tilde{p} = \operatorname{argmax}_p \ \ L(p; \tilde{p}) \ .$$

Dies ist äquivalent zu der Aussage, dass für alle Wahrscheinlichkeitsverteilungen $p(.)$ gilt, dass: $L(\tilde{p};\tilde{p}) \geq L(p;\tilde{p})$. Es ist recht interessant, dass dieses befriedigende Ergebnis zeigt (gleichwertig zu der Aussage ist), dass die Cross-Entropie stets grösser als die Entropie ist. Mit $H(\tilde{p}) = -L(\tilde{p};\tilde{p})$ und $H_{\text{cross}}(p;\tilde{p}) = -L(p;\tilde{p})$ folgt nämlich, dass:

$$H_{\text{cross}}(p;\tilde{p}) \geq H(\tilde{p}) \qquad \forall p \ .$$

Im Abschnitt 2.3 wurde darauf hingewiesen, dass diese Ungleichung kein triviales Resultat ist, sondern die fundamentale *Informationsungleichung der Informationstheorie* beweist, welche aussagt, dass die Kullback-Leibler-Distanz $D(\tilde{p}\|p)$ keine negativen Werte annimmt:

$$D(\tilde{p}\|p) = H_{\text{cross}}(p;\tilde{p}) - H(\tilde{p}) \geq 0 \ .$$

Mit diesem schönen Ergebnis soll der Abschnitt beschlossen werden.