technische universität
dortmund

Bachelorarbeit

Statistische Effizienz der Parameterschätzung für Exponentialfamilien

Olga Scheftelowitsch
September 2019

Gutachter:
Dr. Nico Piatkowski
Prof. Dr. Katharina Morik

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für Künstliche Intelligenz (LS-8)
http://www-ai.cs.tu-dortmund.de/intex.html

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation und Hintergrund

Probabilistische graphische Modelle (PGM) basieren auf Abhägigkeiten verschiedener Zufallsvariablen. In dieser Arbeit werden ungerichtete Modelle, auch bekannt als Markov Random Fields (MRF), für diskrete Verteilungen betrachtet. Da sich durch MRFs viele Verteilungen, wie zum Beispiel die Normal- und die Poissonverteilung [5] darstellen lassen, ist das Anwendungsgebiet sehr weit. Einige Beispiele dafür sind Verkehrsvorhersagen [19], Bildanalyse [16] oder Verarbeitung natürlicher Sprachen [21].

Wichtig bei der Modellierung von PGMs ist die Wahl der Struktur des Graphen und die Schätzung der Parameter. In dieser Arbeit gehen wir von einer bekannten Struktur aus und beschäftigen wir uns nur mit der Schätzung der Modellparameter.

Für einen Schätzer ergibt sich die Frage, wie schnell dieser gegen den wahren Parameter konvergiert und wie viele Daten notwendig sind, um ein hinreichend gutes Ergebnis zu erlangen. Diese mindestens notwendige Menge an Daten, die ein Schätzer braucht, damit er mit einer minimalen Wahrscheinlichkeit von $1 - \delta$ einen Fehler von höchstens ϵ hat, wird durch die Stichprobenkomplexität repräsentiert [4]. Eine nicht genau bekannte Stichprobenkomplexität eines Schätzers kann zu ungenügenden Gütegarantien führen.

Ein möglicher Schätzer ist der Maximum-Likelihood-Schätzer (MLE). Die Berechnung des MLE für große MRFs ist sehr rechenintensiv, jedoch gibt es einen weiteren möglichen Schätzer, welcher sich asymptotisch gleich verhält, den Maximum-Pseudo-Likelihood-Schätzer (MPLE). Dieser ist einfacher zu berechnen, jedoch hat laut Duijn et al. der MPLE eine deutlich größere Varianz und damit eine niedrigere statistische Effizienz als der MLE, daher werden mehr Daten benötigt.

Es stellt sich noch die Frage wie viele Daten benötigt werden um einen hinreichend guten Schätzer zu erzielen. Dies wird durch die Stichprobenkomplexität repräsentiert. In [4] wird eine Schranke für die Stichprobenkomplexität des regularisierten MLE und des MPLE für

Conditional Random Fields (CRF) vorgestellt. In dieser Arbeit wird sich stark an dem Beweis aus [4] orientiert, um eine Stichprobenkomplexität für MRFs herzuleiten.

1.2 Verwandte Arbeiten

Die Beweistechnik aus [4] basiert auf einem Beweis zum Lernen der Struktur von Ising-Modellen [26]. Beide Resultate nutzen den Wert des minimalen Eigenwertes der Hessematrix an der Stelle des wahren Parametervektors $\boldsymbol{\theta}^*$, welcher die Stichproben erzeugt hat. Es existieren mehrere weitere Publikationen, welche sich mit der Stichprobenkomplexität verschiedener Lernmethoden für MRFs auseinandersetzten. In [17] wird ein Algorithmus für MRFs mit einer Verteilung über $\{-1, 1\}^n$ vorgestellt, deren maximale Cliquengröße t ist. Die Stichprobenkomplexität dieses Algorithmus hängt logarithmisch von der Knotenanzahl n ab, dafür aber exponentiell von t und λ, wobei λ für Ising-Modelle über alle Knoten $i \in \{1, \ldots, n\}$ die maximale Summe der Einträge im Parametervektor $\boldsymbol{\theta}^*$ ist, die mit dem Knoten i zusammenhängen. Weiterhin muss jeder Parameter, der nicht null ist, eine Mindestgröße η haben. In [10] wird weiterhin ein Algorithmus analysiert, dessen Stichprobenkomplexität auch logarithmisch in n ist, dafür aber von den minimalen und maximalen Werten, die im wahren Parametervektor stehen abhängt. Alle diese Stichprobenkomplexitäten enthalten Annahmen über Parameter, die nicht bekannt sind. In dieser Arbeit konzentrieren wir uns auf die Stichprobenkomplexität des MLE und des MPLE für MRFs. Diese wird anhand des Beweises von Bradley und Guestrin [4] für CRFs bestimmt, sodass auch die hier vorgestellte Stichprobenkomplexität von solchen Parametern abhängt. In [4] wird anstelle des MPLE, der verallgemeinerte Maximum-Composite-Likelihood-Schätzer (MCLE) [20] analysiert. Während die Pseudolikelihood die bedingten Wahrscheinlichkeiten der einzelnen Knoten nutzt, nutzt die Composite-Likelihood die bedingten Wahrscheinlichkeiten von ausgewählen Knotenmengen, sogenannten Komponenten. In [23] werden Rahmenbedingungen für die Wahl solcher Komponenten aufgestellt, sodass die Berechnung der Parameter verteilt ablaufen kann und der resultierende Schätzer konsistent bleibt, währen in [22] das asymptotische Verhalten des MPLE analysiert wird.

1.3 Aufbau der Arbeit

In dieser Arbeit werden als erstes in Kapitel 2 die Grundlagen der Exponentialfamilie, des MLE und des MPLE aufgearbeitet. In Kapitel 3 wird die Stichprobenkomplexität der beiden Schätzer bezüglich generischer MRFs auf Basis von dem Beweis von Bradley und Guestrin bestimmt. In Kapitel 4 wird die praktische Relevanz der Ergebnisse mithilfe von synthetischen Experimenten ermittelt und in Kapitel 5 ein Fazit gezogen.

Kapitel 2

Grundlagen

2.1 Notation

Die Notation in dieser Arbeit lehnt sich hauptsächlich an [25] und [4] an. In dieser Tabelle ist die, in dieser Arbeit genutzte, Notation kurz beschrieben.

$\boldsymbol{X}$	ein Zufallsvektor
$\mathcal{X}$	der Zustandsraum des Zufallsvektors $\boldsymbol{X}$
$\boldsymbol{x}$	eine Realisierung des Zufallsvektors $\boldsymbol{X}$
$\sum_{\boldsymbol{x}} = \sum_{\boldsymbol{x} \in \mathcal{X}}$	eine Summe über alle Realisierungen des Zufallsvektors $\boldsymbol{X}$
$\boldsymbol{X}_{A_j}$	ein Teilzufallsvektor des Zufallsvektors $\boldsymbol{X}$ über die Indexmenge A_j
$\mathcal{X}_{A_j}$	der Zustandsraum des Teilzufallsvektors $\boldsymbol{X}_{A_j}$
$\boldsymbol{x}_{A_j}$	eine Realisierung des Teilzufallsvektors $\boldsymbol{X}_{A_j}$
$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x})$	die Wahrscheinlichkeit $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$, wenn der Parametervektor der Verteilung $\boldsymbol{\theta}$ ist.
$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}\|\boldsymbol{Y})$	die bedingte Wahrscheinlichkeit von $\boldsymbol{X}$ gegeben $\boldsymbol{Y}$, wenn der Parametervektor der Verteilung $\boldsymbol{\theta}$ ist.
$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}\|\boldsymbol{Y})}[\boldsymbol{X}]$	Der Erwartungswert von $\boldsymbol{X}$ gegeben $\boldsymbol{Y}$, wenn der Parametervektor der Verteilung $\boldsymbol{\theta}$ ist.

Weitere Notation wird im Verlauf der Arbeit vorgestellt.

2.2 Exponentialfamilie

2.2.1 Markov Random Fields

Markov Random Fields sind ungerichtete probabilistische graphische Modelle. Sie beschreiben die Abhängigkeiten von Zufallsvariablen innerhalb eines Zufallsvektors $\boldsymbol{X}$ mit

Hilfe eines Graphen $G = (V, E)$. Die Knoten des Graphen repräsentieren die einzelnen Zufallsvariablen im Zufallsvektor. Wir definieren, für $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subset V$, gilt

$$\boldsymbol{A} \perp\!\!\!\perp_G \boldsymbol{B} \mid \boldsymbol{C}$$

genau dann wenn die Knotenmenge $\boldsymbol{C}$ die Knotenmengen $\boldsymbol{A}$ und $\boldsymbol{B}$ trennt. Dies ist der Fall, wenn auf jedem Weg zwischen einem Knoten aus $\boldsymbol{A}$ und einem Knoten $\boldsymbol{B}$ ein Knoten aus $\boldsymbol{C}$ liegt [1]. Seien weiterhin die Bezeichnungen für die Zufallsvariablen und Knoten gleich. Wenn der Graph ein MRF ist, dann gilt

$$\boldsymbol{A} \perp\!\!\!\perp_G \boldsymbol{B} \mid \boldsymbol{C} \text{ impliziert } \mathbb{P}(\boldsymbol{A}|\boldsymbol{B}, \boldsymbol{C}) = \mathbb{P}(\boldsymbol{A}|\boldsymbol{C}) \, .$$

Beispiel

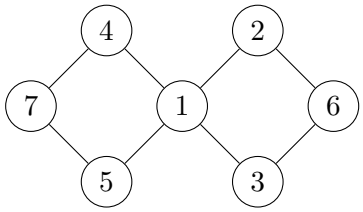


Abbildung 2.1: Beispielgraph

Sei der oben abgebildete Graph ein Beispiel-MRF, dann lässt sich erkennen, dass der Knoten $\{1\}$ die Knoten $\{4, 7, 5\}$ und $\{2, 3, 6\}$ von einander trennt, somit gilt $\mathbb{P}(\{4, 7, 5\}|\{2, 3, 6, 1\}) = \mathbb{P}(\{4, 7, 5\}|\{1\})$.

Wahrscheinlichkeitsverteilung von diskreten MRFs Für die Wahrscheinlichkeitsverteilung des MRF wird der Graph in eine Menge von Cliquen $\mathcal{C}$ eingeteilt, und die Wahrscheinlichkeitsfunktion wird über die Potentialfunktionen ψ_C dieser Cliquen gebildet:

$$\mathbb{P}(\boldsymbol{x}) = \frac{1}{Z}\psi(\boldsymbol{x}) = \frac{1}{Z}\prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C) \tag{2.1}$$

Wobei Z die Normalisierungskonstante $Z = \sum_{\boldsymbol{x}} \mathbb{P}(\boldsymbol{x})$ ist [28]. Solche Wahrscheinlichkeitsverteilung lassen sich als *Exponentialfamilie* darstellen [25].

2.2.1 Definition (Exponentialfamilie [28]). Sei $\boldsymbol{X}$ ein n-dimensionaler Zufallsvektor und sei $\boldsymbol{x} \in \mathcal{X}$ und $\boldsymbol{\theta} \in \mathbb{R}^d$ ein Parametervektor. $\boldsymbol{X}$ gehört genau dann zu einer Exponentialfamilie wenn sich ihre Dichtefunktion schreiben lässt als

$$\mathbb{P}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})) \, . \tag{2.2}$$

Die *suffiziente Statistik* $\phi(\boldsymbol{x})$ wird durch die graphische Struktur $G = (V, E)$ des MRFs bestimmt.

2.2.2 Definition (Suffiziente Statistik [25]). Eine Funktion $\phi : \mathcal{X} \to \mathbb{R}^d$ wird für einen Zufallsvektor $\boldsymbol{X}$ mit $\boldsymbol{x} \in \mathcal{X}$ suffizient genannt, wenn $\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{x}, \phi(\boldsymbol{x})) = \mathbb{P}(\boldsymbol{\theta} \mid \phi(\boldsymbol{x}))$ gilt.

Die log-Partitionsfunktion $A(\boldsymbol{\theta}) = \log(Z(\boldsymbol{\theta})) = \log\left(\sum_{\boldsymbol{x}} \exp\left(\langle \phi(\boldsymbol{x}), \boldsymbol{\theta} \rangle\right)\right)$ sorgt für die Normierung der Wahrscheinlichkeitsverteilung des Modells.

2.2.2 Herleitung der Exponentialfamilie

Ein zentraler Begriff in der Informationstheorie ist die Entropie einer Verteilung. Um die Entropie einer Verteilung zu verstehen, muss zuerst der Informationsgehalt eines einzelnen Ereignisses betrachtet werden. Die Frage, die sich dabei stellt, ist der Zusammenhang zwischen dem Informationsgehalt und der Wahrscheinlichkeit eines Ereignisses, denn umso wahrscheinlicher ein Ereignis ist, umso weniger informativ ist sein Eintreffen [1]. Dies bedeutet, dass jeder Wahrscheinlichkeit ein Informationswert zugeordnet werden kann. Dafür wird die Funktion Inf definiert.

$$\mathrm{Inf}(\boldsymbol{x}) = -\log(\mathbb{P}(\boldsymbol{x})) \tag{2.3}$$

Zu bemerken ist hier, dass

$$\mathrm{Inf}(\boldsymbol{x}) = -\log(\mathbb{P}(\boldsymbol{x})) = -\log(1) = 0$$

und

$$\mathrm{Inf}(\boldsymbol{x}) = -\log(\mathbb{P}(\boldsymbol{x})) = -\log(0) = \infty$$

gilt. Intuitiv formuliert ist ein unmögliches Ereignis unmessbar informativ und ein Ereignis, dessen eintreffen mit Sicherheit bekannt ist, enthält keine neue Information.
Auf Basis des Informationsgehalts, kann die Entropie H einer Verteilung als der Erwartungswert des Informationsgehaltes eines Ereignisses mit dieser Verteilung definiert werden.

2.2.3 Definition (Entropie). Sei $\mathbb{P}$ eine Wahrscheinlichkeitsverteilung und $\boldsymbol{X}$ eine n-dimensionale Zufallsvariable mit dieser Verteilung, dann ist die Entropie von $\mathbb{P}$ definiert als

$$H(\mathbb{P}) = \mathbb{E}(\mathrm{Inf}(\boldsymbol{X})) = \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}(\boldsymbol{x}) \,\mathrm{Inf}(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} -\mathbb{P}(\boldsymbol{x}) \log(\mathbb{P}(\boldsymbol{x})) \tag{2.4}$$

Ohne weitere Bedingungen hätte die Gleichverteilung maximale Entropie, jedoch wäre es wünschenswert, weitere Bedingungen an die Verteilung zu stellen. Angenommen, es existieren schon Beobachtungen der Verteilung und wir wollen, diese Beobachtungen einbinden, aber die Entropie -und somit die Unsicherheit der Verteilung- maximal halten. Um diese Bedingungen in das Problem einzubinden benutzen wir den Erwartungswert einer Funktion ϕ mit $\phi : \mathcal{X} \to \mathbb{R}^d$, und als Bedingung fügen wir $\mathbb{E}_{\mathbb{P}}[\phi(\boldsymbol{X})] = \boldsymbol{a}$ ein. Dabei ist $\boldsymbol{a}$ der

empirische Erwartungswert von $\phi(\boldsymbol{X})$. Bei dieser Problemstellung handelt es sich um ein Optimierungsproblem mit Nebenbedingung. Wir definieren das Problem folgendermaßen:

$$\begin{aligned} &\max_{\mathbb{P}} H(\mathbb{P}) \\ \text{u.d.B. } &\forall i \in \{1, \dots, d\}\, \mathbb{E}[\phi(\boldsymbol{X})]_i = \boldsymbol{a}_i \\ &\sum_{x \in \mathcal{X}} \mathbb{P}(\boldsymbol{x}) = 1 \\ &\forall \boldsymbol{x} \in \mathcal{X}\, \mathbb{P}(\boldsymbol{x}) \geq 0 \end{aligned}$$

Dabei steht $i \in (1, \dots, d)$ für jede Dimension von ϕ. Dieses Problem lässt sich über Lagrange-Multiplikatoren[3] lösen.

2.2.4 Definition (Lagrange-Multiplikator). Für jede Zielfunktion f über $\boldsymbol{x}$ mit den Nebenbedingungen $g_i(\boldsymbol{x}) = 0$ existieren Lagrange-Multiplikatoren $\boldsymbol{\theta}_i$, so dass die Extremstellen von

$$L(x, \boldsymbol{\theta}) = f(\boldsymbol{x}) + \sum_i \boldsymbol{\theta}_i g_i(\boldsymbol{x}) \tag{2.5}$$

die gleichen sind, wie die von f unter den Nebenbedingungen.

Dies nutzen wir um für unser Problem folgende Funktion mithilfe von Lagrange-Multiplikatoren zu definieren:

$$L(\mathbb{P}, \boldsymbol{\theta}, A) = \sum_{x \in \mathcal{X}} -\mathbb{P}(\boldsymbol{x}) \log(\mathbb{P}(\boldsymbol{x})) + \sum_{i=1}^{d} \boldsymbol{\theta}_i (\mathbb{E}[\phi(\boldsymbol{X})]_i - \boldsymbol{a}_i) + A \left(\sum_{x \in \mathcal{X}} \mathbb{P}(\boldsymbol{x}) - 1 \right)$$

Um die Extrema dieser Funktion zu finden, müssen wir zuerst ihren Gradienten bestimmen und ihn gleich null setzen. Der Gradient $\frac{\partial}{\partial \mathbb{P}} L(\mathbb{P}, \boldsymbol{\theta}, A)$ ist ein Vektor, in dem ein Eintag für jedes $\boldsymbol{x} \in \mathcal{X}$ existiert mit der Ableitungen nach $\mathbb{P}(\boldsymbol{x})$:

$$\frac{\partial}{\partial\, \mathbb{P}(\boldsymbol{x})} L(\mathbb{P}, \boldsymbol{\theta}, A) = -\log(\mathbb{P}(\boldsymbol{x})) - 1 + \sum_{i=1}^{d} \boldsymbol{\theta}_i (\phi(\boldsymbol{x})_i) + A$$

Als nächstes setzen wir diese Einträge auf null:

$$\begin{aligned} 0 &= -\log\left(\mathbb{P}(\boldsymbol{x})\right) - 1 + \sum_{i=1}^{d} \boldsymbol{\theta}_i \phi(\boldsymbol{x})_i + A \\ \Leftrightarrow \mathbb{P}(\boldsymbol{x}) &= \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - 1 + A) \end{aligned} \tag{2.6}$$

Jetzt müssen wir prüfen, ob es sich hier tatsächlich um ein lokales Maximum handelt, wofür wir nun die zweite Ableitung berechnen:

$$\frac{\partial}{\partial\, \mathbb{P}(\boldsymbol{x})^2} L(\mathbb{P}, \boldsymbol{\theta}, A) = -\frac{1}{\mathbb{P}(\boldsymbol{x})}$$

Da $\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - 1 + A)$ immer positiv ist, ist $-1/\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - 1 + A)$ negativ und es handelt sich tatsächlich um ein lokales Maximum. Es ist sogar ein globales Maximum, da die Funktion stetig ist, und dies das einzige Extremum ist.
Wenn wir für (2.6) noch die Bedingung $\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}(\boldsymbol{x}) = 1$ beachten, dann lässt sich erkennen, dass

$$-1 + A = \log\left(\sum_{x\in\mathcal{X}} \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle)\right) \tag{2.7}$$

gelten muss, damit die Wahrscheinlichkeitsfunktion normiert ist. Dabei handelt es sich um den Logarithmus der Partitionsfunktion der Exponentialfamilie $A(\boldsymbol{\theta})$ [25]. Insgesamt hat dann $\mathbb{P}(\boldsymbol{x})$ die Form:

$$\mathbb{P}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\theta})). \tag{2.8}$$

Somit folgt also aus der maximalen Entropie mit der Nebenbedingung $\mathbb{E}[\phi(\boldsymbol{X})] = \boldsymbol{a}$ die Exponentialfamilie [25].

2.2.3 Eigenschaften der Exponentialfamilie

Es gibt mehrere suffiziente Statistiken ϕ, die für eine Verteilung aus der Exponentialfamilie gewählt werden können.

2.2.5 Definition (Minimale suffiziente Statistik). Eine Darstellung von $\phi : \mathcal{X} \to \mathbb{R}^d$ heißt *minimal*, wenn kein Vektor $\boldsymbol{a} \in \mathbb{R}^d \setminus \{0\}$ existiert, sodass $\langle \phi(\boldsymbol{X}), \boldsymbol{a}\rangle = b$ gilt, wobei b eine Konstante ist. Eine Darstellung von ϕ heißt *overcomplete* (Übervollständig), wenn sie nicht minimal ist [28].

Die *overcomplete* Darstellung wird häufiger angewendet, da sie eine intuitive Vorstellung ermöglicht.

2.2.6 Lemma. *Jeder nicht leere Teilvektor einer minimalen Statistik ϕ ist auch minimal, auch wenn die dadurch Implizierte Verteilung eine andere ist. Als Teilvektor von ϕ bezeichnen wir einen Vektor ϕ_{teil}, welcher eine Subsequenz von ϕ ist.*

2.2.7 Beweis. Angenommen ϕ ist minimal und ϕ_{teil} ist ein c-dimensionaler Teilvektor von ϕ, welcher nicht minimal ist, dann existiert ein Vektor $\boldsymbol{t} \in \mathbb{R}^c$, so dass $\langle \phi_{teil}(\boldsymbol{X}), t\rangle = b$ gilt. Erstellen wir nun die Funktion $f : \mathbb{N} \to \mathbb{N}$, welche den Index eines Elements im Vektor ϕ auf den dazugehörigen Index im Vektor ϕ_{teil} abbildet, wenn es so eine Abbildung gibt und auf null sonst. Somit könnten wir folgenden Vektor $\boldsymbol{a}$ bilden:

$$\boldsymbol{a}_i = \begin{cases} 0 & \text{wenn } f(i) = 0 \\ \boldsymbol{t}_{f(i)} & \text{sonst} \end{cases}.$$

Dies würde bedeuten, dass $\langle \phi(\boldsymbol{X}), \boldsymbol{a}\rangle = \langle \phi_{teil}(\boldsymbol{X}), t\rangle = b$ gelten würde und ϕ nicht minimal wäre, was ein Widerspruch zu der Ausgangsbedingung ist. Daraus folgt Lemma 2.2.6.

2.3 Maximum Likelihood

Der erste Schätzer mit dem wir uns befassen werden ist der Maximum-Likelihood-Schätzer (MLE), welcher die Likelihood-Funktion maximiert.

2.3.1 Definition (ML-Schätzer). Sei $D = (\boldsymbol{x}^1, ..., \boldsymbol{x}^N)$ ein Datensatz mit den beobachteten Werten des Zufallsvektors $\boldsymbol{X}$. Sei weiterhin die Wahrscheinlichkeitsfunktion $\mathbb{P}_{\boldsymbol{\theta}}$ von $\boldsymbol{X}$ vom Parametervektor $\boldsymbol{\theta}$ abhängig, dann ist die empirische Likelihood-Funktion definiert als

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}^i) \ . \tag{2.9}$$

Und somit ist der ML-Schätzer definiert als

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) \ . \tag{2.10}$$

Im maschinellen Lernen wird generell Loss-Funktion ℓ verwendet, um zu bestimmen, wie genau ein Modell die Verteilung von einer gegebenen Menge an Daten beschreibt. Der Wert der Loss-Funktion muss entweder als eine „Belohnung“ oder als eine „Bestrafung“ interpretiert werden. Bei einer Belohnung soll die Loss-Funktion maximiert und bei einer Bestrafung minimiert werden. Die Likelihood-Funktion selbst ist eine mögliche Loss-Funktion. Da Loss-Funktionen jedoch üblicherweise minimiert werden, nutzen wir hier die negative empirische Log-Likelihood. Dies können wir tun, weil die Logarithmus-Funktion monoton ist, und damit $\operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \log(L(\boldsymbol{\theta})) = \operatorname{argmin}_{\boldsymbol{\theta}} (-\log(L(\boldsymbol{\theta})))$ gilt. Weiterhin gilt $\operatorname{argmin}_{\boldsymbol{\theta}} (-\log(L(\boldsymbol{\theta}))) = \operatorname{argmin}_{\boldsymbol{\theta}} \left(-\frac{1}{N} \log(L(\boldsymbol{\theta}))\right)$. Damit definieren wir nun die von uns im weiteren Laufe der Arbeit als empirische Log-Likelihood bezeichnete Funktion.

2.3.2 Definition (Empirische Log-Likelihood). Sei $L(\boldsymbol{\theta})$ die Likelihood-Funktion, wie sie in (2.9) definiert ist, dann ist die empirische Loss-Funktion definiert als

$$\hat{\ell}_L(\boldsymbol{\theta}) = -\frac{1}{N} \log(L(\boldsymbol{\theta})) = \frac{1}{N} \sum_{i=1}^{N} -\log(\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}^i)) \ . \tag{2.11}$$

Durch diese neue Definition können wir die wahre Log-Likelihood-Funktion definieren, gegen die die empirische Log-Likelihood konvergiert [26].

2.3.3 Definition (Log-Likelihood). Sei $\boldsymbol{\theta}^*$ der Parametervektor der wahren Verteilung, $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x})$ die Wahrscheinlichkeit von $\boldsymbol{X} = \boldsymbol{x}$ und $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}]$ der Erwartungswert von der Zufallsvariable $\boldsymbol{X}$, wenn der Parameter der Verteilung $\boldsymbol{\theta}$ ist. Dann ist die wahre Log-Likelihood für den MLE definiert als

$$\ell_L(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}^*}[-\log(\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}))] \ . \tag{2.12}$$

Hierbei ist es wichtig anzumerken, dass $\operatorname{argmax}_{\boldsymbol{\theta}} \ell_L(\boldsymbol{\theta})$ der wahre Parameter $\boldsymbol{\theta}^*$ ist. Dies gilt weil der MLE für $N \to \infty$ gegen $\boldsymbol{\theta}^*$ konvergiert und somit konsistent ist [9]. Da $\hat{\ell}_L(\boldsymbol{\theta})$ bei einer steigenden Anzahl an Beobachtungen N gegen $\ell_L(\boldsymbol{\theta})$ konvergiert, gilt die Behauptung. Der MLE weist sehr hohe asymptotische Effizienz auf, da er unter allen Schätzern für $N \to \infty$ die kleinstmögliche Varianz hat [22]. Jedoch ist allein die Berechnung der Partitionsfunktion $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle)$, welche für die Berechnung des MLE nötig wär, allgemein #P-schwer [6].

2.3.4 Definition (Komplexitätsklasse #P [29]). Die Komplexitätsklasse #P enthält alle Zählprobleme, für deren dazugehörige Entscheidungsprobleme ein polynomiell nichtdeterministischer Algorithmus existiert.

Eine weitere wichtige Eigenschaft der wahren, sowie der empirischen Log-Likelihood des MLE ist, dass diese für Exponentialfamilien in minimaler Darstellung streng konvex ist. Damit ist die Hessematrix der Funktion positiv definit und ihr minimaler Eigenwert größer null.

2.3.5 Beweis (Strenge Konvexität der Log-Likelihood). Eine multivariate Funktion ist streng konvex, wenn ihre Hessematrix positiv definit ist. Eine Matrix $A \in \mathbb{R}^{d \times d}$ ist positiv semidefinit, wenn für alle $\boldsymbol{v} \in \mathbb{R}^d$ mit $\boldsymbol{v} \neq 0$ die Bedingung

$$\boldsymbol{v}^T A \boldsymbol{v} \geq 0 \tag{2.13}$$

gilt, und positiv definit wenn

$$\boldsymbol{v}^T A \boldsymbol{v} > 0 \tag{2.14}$$

gilt. Die Hessematrix der Log-Likelihood ist die Kovarianzmatrix von $\phi(\boldsymbol{X})$.

2.3.6 Definition (Kovarianzmatrix[24]). Die Kovarianzmatrix eines n-dimensionalen Zufallsvektors $\boldsymbol{X}$ ist definiert als

$$\operatorname{Cov}(\boldsymbol{X}) = \begin{pmatrix} \operatorname{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_1) & \cdots & \operatorname{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_n) \\ \ddots & \vdots & \ddots \\ \operatorname{Cov}(\boldsymbol{X}_n, \boldsymbol{X}_1) & \cdots & \operatorname{Cov}(\boldsymbol{X}_n, \boldsymbol{X}_n) \end{pmatrix} \tag{2.15}$$

Es muss also gezeigt werden, dass die Kovarianzmatrix positiv definit ist, wenn $\phi(\boldsymbol{X})$ minimal ist:

$$\boldsymbol{v}^T \operatorname{Cov}(\phi(\boldsymbol{X}))\boldsymbol{v} = \sum_{i=1}^{d} \sum_{j=1}^{d} \boldsymbol{v}_i \boldsymbol{v}_j \operatorname{Cov}(\phi(\boldsymbol{X})_i, \phi(\boldsymbol{X})_j)$$

Aufgrung von der bilinearität des Kovarianzoperators gilt [12]:

$$\begin{aligned} &= \sum_{i=1}^{d} \sum_{j=1}^{d} \operatorname{Cov}(\phi(\boldsymbol{X})_i \boldsymbol{v}_i, \phi(\boldsymbol{X})_j \boldsymbol{v}_j) \\ &= \operatorname{Var}\left(\sum_{i=1}^{d} \phi(\boldsymbol{X})_i \boldsymbol{v}_i \right) \geq 0 \end{aligned}$$

Die Varianz einer Zufallsvariable ist nur null, wenn diese Zufallsvariable eine Konstante ist. Da ϕ jedoch minimal ist kann es keinen Vektor $\boldsymbol{v}$ geben, so dass $\langle \phi(\boldsymbol{X}), \boldsymbol{v} \rangle = \sum_{i=1}^{d} \phi(\boldsymbol{X})_i \boldsymbol{v}_i$ eine Konstante ist.

2.4 Maximum Pseudo Likelihood

Obwohl die statistische Effizienz des MLEs hoch ist, gibt es, wie bereits erwähnt, für die Berechnung des MLE für MRFs im Allgemeinen keinen effizienten Algorithmus. Aus diesem Grund wurde die Pseudo-Likelihood von Besag [2] entwickelt.

2.4.1 Definition (MPL-Schätzer [2]). Sei $D = (\boldsymbol{x}^1, ..., \boldsymbol{x}^N)$ ein Datensatz mit den beobachteten Werten des n-dimensionalen Zufallsvektors $\boldsymbol{X}$ mit der vom Parametervektor $\boldsymbol{\theta}$ abhängenden Verteilung $\mathbb{P}_{\boldsymbol{\theta}}$. Seien die Zufallvariablen $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ des Zufallsvektors $\boldsymbol{X}$ durch ein MRF mit dem Graphen $G = (V, E)$ representiert. Sei nun $\mathcal{N}(v)$ für $v \in V$ die Nachbarschaft des Knotens v in G, dann ist die empirische Pseudo-Likelihood definiert als

$$L_{PL}(\boldsymbol{\theta})_x = \prod_{i=1}^{N} \prod_{j=1}^{n} \mathbb{P}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_j^i \mid \boldsymbol{x}_{\mathcal{N}(j)}^i \right) . \tag{2.16}$$

Wobei hier $\boldsymbol{x}_j^i$ die Realisierung von $\boldsymbol{x}_j$ in der i-ten Beobachtung ist. Somit ist der MPL-Schätzer definiert als

$$\hat{\boldsymbol{\theta}}_{PL} = \operatorname*{argmax}_{\boldsymbol{\theta}} L_{PL}(\boldsymbol{\theta}) . \tag{2.17}$$

Die Pseudo-Likelihood wurde 1988 von Lindsay [20] durch die Composite-Likelihood verallgemeinert.

2.4.2 Definition (MCLE). Sei $D = (\boldsymbol{x}^1, ..., \boldsymbol{x}^N)$ ein Datensatz mit den beobachteten Werten des Zufallsvektors $\boldsymbol{X}$ und der vom Parametervektor $\boldsymbol{\theta}$ abhängenden Verteilung $\mathbb{P}_{\boldsymbol{\theta}}$. Seien weiterhin die Zufallvariablen $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ des Zufallsvektors $\boldsymbol{X}$ durch ein MRF mit dem Graphen $G = (V, E)$ repräsentiert. $\boldsymbol{X}$ wird in m Teilkomponenten $\boldsymbol{X}_{A_1}, \ldots, \boldsymbol{X}_{A_m}$ aufgeteilt. Hierbei ist $A_j \in \{A_1, \ldots, A_m\}$ die Indexmenge einer der m Komponenten. Sei ferner die Nachbarschaft einer Menge $\mathcal{N}(A) = \bigcup_{v \in A} N(v) \setminus A$, dann ist die empirische Composite-Likelihood definiert als

$$L_{CL}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{m} \mathbb{P}(\boldsymbol{x}_{A_j}^i \mid \boldsymbol{x}_{\mathcal{N}(A_j)}) , \tag{2.18}$$

wobei hier $\boldsymbol{x}_{A_j}^i$ die Realisierung von $\boldsymbol{x}_{A_j}$ in der i-ten Beobachtung ist. Somit ist der MCLE definiert als

$$\hat{\boldsymbol{\theta}}_{CL} = \operatorname*{argmax}_{\boldsymbol{\theta}} L_{CL}(\boldsymbol{\theta}) . \tag{2.19}$$

In der Literatur kann die Composite-Likelihood über die bedingten Wahrscheinlichkeiten, aber auch über die Randwahrscheinlichkeiten der Komponente definiert werden [20], wie in der Definition erkennbar ist, nutzen wir hier die bedingten Wahrscheinlichkeiten.
Wenn $\forall j \in \{1, \dots, m\}$ $\boldsymbol{x}_{A_j} = \boldsymbol{x}_j$ und $m = n$ gilt, daher jede Zufallsvariable eine eigene Komponente ist, dann handelt es sich beim MCLE um den MPLE und wenn $\boldsymbol{x}_{A_1} = \boldsymbol{x}$ und $m = 1$ gilt, wodurch es nur eine Komponente mit allen Zufallsvariablen gibt, dann handelt es sich um den MLE.
Da die Pseudo-Likelihood nur ein Spezialfall der Composite-Likelihood ist, betrachten wir hier die empirische Log-Composite-Likelihood:

$$\ell_{CL}(\boldsymbol{\theta}) = -\frac{1}{N} \log \prod_{i=1}^{N} \prod_{j=1}^{m} \mathbb{P}(\boldsymbol{x}^i_{A_j} \mid \boldsymbol{x}^i_{\mathcal{N}(A_j)})$$

Da es sich hier um ein MRF handelt, gilt $\mathbb{P}(\boldsymbol{x}_{A_j} | \boldsymbol{x}_{\mathcal{N}(A_j)}) = \mathbb{P}(\boldsymbol{x}_{A_j} | \boldsymbol{x}_{\setminus A_j})$ und somit können wir $\mathcal{N}(A_j)$ mit $\setminus A_j$ von hier aus ersetzen.

$$\begin{aligned}
&= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log \mathbb{P}(\boldsymbol{x}^i_{A_j} \mid \boldsymbol{x}^i_{\setminus A_j}) \\
&= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log \frac{\mathbb{P}(\boldsymbol{x}^i_{A_j}, \boldsymbol{x}^i_{\setminus A_j})}{\mathbb{P}(\boldsymbol{x}^i_{\setminus A_j})} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log \frac{\mathbb{P}(\boldsymbol{x}^i_{A_j}, \boldsymbol{x}^i_{\setminus A_j})}{\sum_{\boldsymbol{x}'_{A_i}} \mathbb{P}(\boldsymbol{x}'_{A_i}, \boldsymbol{x}^i_{\setminus A_j})} \\
&= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log \frac{\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}^i) \rangle - A(\boldsymbol{\theta}))}{\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle - A(\boldsymbol{\theta}))} \\
&= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log \frac{\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}^i) \rangle) \exp(-A(\boldsymbol{\theta}))}{\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \exp(-A(\boldsymbol{\theta}))} \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} -\log \frac{\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}^i) \rangle)}{\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle)}
\end{aligned} \tag{2.20}$$

Wir können uns überlegen, dass $\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}^i) \rangle$ und $\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle$ gemeinsame Summanden haben, nämlich die Einträge von $\phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})$, die nur von $\boldsymbol{x}^i_{\setminus A_j}$ abhängen. Dafür definieren wir die Indexmenge B_j.

2.4.3 Definition. Sei A_j mit $j \in \{1, \dots, m\}$ eine Indexmenge von Zufallsvariablen aus dem Zufallsvektor $\boldsymbol{X}$, dann ist B_j definiert als die Indexmenge der Einträge in $\phi(\boldsymbol{X})$, die von $\boldsymbol{X}_{A_j}$ abhängen.

Diese Summanden können wir im Nenner und im Zähler auf die gleiche Weise wie $A(\boldsymbol{\theta})$ kürzen.

$$\begin{aligned}
\ell_{CL}(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} -\log \frac{\exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}^i) \rangle)}{\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle)} \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \Bigg(-\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}^i) \rangle + \log \Bigg(\underbrace{\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{A_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle)}_{Z_{A_j}(\boldsymbol{\theta}, \boldsymbol{x}^i)} \Bigg) \Bigg)
\end{aligned} \tag{2.21}$$

Der Vektor ϕ_{B_j} ist eine Abbildung des Vektors ϕ bezüglich der Indexmenge B_j 2.4.3. Es gibt zwei Möglichkeiten, $\phi_{B_j}(\boldsymbol{X})$ darzustellen. Die erste Möglichkeit wäre, wenn $\phi_{B_j}(\boldsymbol{X})$ die gleiche Dimension hat wie $\phi(\boldsymbol{X})$ und

$$\phi_{B_j}(\boldsymbol{X})_i = \begin{cases} \phi(\boldsymbol{X})_i & \text{wenn } i \in B_j \\ 0 & \text{sonst} \end{cases} .$$

für alle $i \in \{1, \dots, d\}$ gilt. Und die zweite Möglichkeit ist, wenn ϕ_{B_j} nur auf die Indizes in B_j reduziert wird und deswegen eine kleinere Dimension als ϕ hat, diese Interpretation bezeichnen wir als „reduziert“.

2.4.4 Definition (Reduzierter Vektor). Ein Vektor $\boldsymbol{v}_{B_j}$ wird bezüglich einer Indexmenge B_j als reduziert bezeichnet, wenn $\boldsymbol{v}_{B_j}$ eine Subsequenz des Vektors $\boldsymbol{v}$ ist, und die Einträge von $\boldsymbol{v}$, deren Indizes sich in der Indexmenge B_j befinden, enthält.

In dieser Arbeit gehen wir von nicht reduzierten Vektoren $\phi_{B_j}(\boldsymbol{X})$ und $\boldsymbol{\theta}_{B_j}$ aus, außer es wird explizit erwähnt, dass es sich um die reduzierte Version der Vektoren handelt.

Beispiel Sei $\boldsymbol{X}$ ein 3-dimensionaler Zufallsvektor mit einer Verteilung aus der Exponentialfamilie und sei die Funktion $\phi : \mathcal{X} \to \mathbb{R}^d$ wie unten links dargestellt, dann wäre für die Menge $A_1 = \{1\}$ die Funktion $\phi_{B_1} : \mathcal{X} \to \mathbb{R}^d$ in der Mitte und die reduzierte Funktion ϕ_{B_1} rechts dargestellt.

$$\phi(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_1\boldsymbol{x}_2 \\ \boldsymbol{x}_2\boldsymbol{x}_3 \end{pmatrix} \qquad \phi_A(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x}_1 \\ 0 \\ 0 \\ \boldsymbol{x}_1\boldsymbol{x}_2 \\ 0 \end{pmatrix} \qquad \phi_A(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_1\boldsymbol{x}_2 \end{pmatrix}$$

Wie wir sehen können, ist für $A_j = \{1\}$ die Indexmenge $B_j = \{1, 4\}$, da nur die Einträge 1 und 4 von $\boldsymbol{x}_1$ abhängen.

Insgesamt erhalten wir als Definition für die empirische Log-Composite-Likelihood:

2.4.5 Definition (Empirische Log-Composite-Likelihood). Sei $D = (\boldsymbol{x}^1, ..., \boldsymbol{x}^N)$ ein Datensatz mit beobachteten Werten des n-dimensionalen Zufallsvektors $\boldsymbol{X}$. Sei $\boldsymbol{X}$ in m Teilkomponenten $\boldsymbol{X}_{A_1}, \dots, \boldsymbol{X}_{A_m}$ aufgeteilt. Hierbei ist $A_j \in \{A_1, \dots, A_m\}$ die Menge der Indizes einer der m Komponenten. Sei ferner $\forall j \in \{1, \dots, m\}$ die Indexmenge B_j so definiert wie in 2.4.3, dann ist die empirische Log-Composite-Likelihood definiert als

$$\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \left(-\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}^i) \rangle + \log \left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right) \right) . \quad (2.22)$$

Nun können wir, mit der gleichen Begründung wie beim MLE, die wahre Log-Composite-Likelihood-Funktion definieren.

2.4.6 Definition (Log-Composite-Likelihood). Sei $\boldsymbol{\theta}^*$ der Parametervektor der wahren Verteilung, $\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ die Wahrscheinlichkeit von $\boldsymbol{X} = \boldsymbol{x}$ und $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}]$ der Erwartungswert von des n-dimensionalen Zufallsvektors $\boldsymbol{X}$ ist, wenn der Parameter der Verteilung $\boldsymbol{\theta}$ ist. Sei $\boldsymbol{X}$ in m Teilkomponenten $\boldsymbol{X}_{A_1}, \ldots, \boldsymbol{X}_{A_m}$ aufgeteilt. Hierbei ist $A_j \in \{A_1, \ldots, A_m\}$ die Menge der Indizes einer der m Komponenten. Sei ferner $\forall j \in \{1, \ldots, m\}$ die Indexmenge B_j so definiert wie in 2.4.3. Dann ist die wahre Log-Composite-Likelihood für den MCLE definiert als

$$\ell_{CL}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{j=1}^{m}\left(-\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{X})\rangle + \log\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{X}_{\setminus A_j})\rangle)\right)\right)\right] \tag{2.23}$$

Der MPLE konvergiert bei steigender Stichprobengröße gegen den MLE und ist somit auch ein konsistenter Schätzer. Damit gilt $\operatorname{argmax}_{\boldsymbol{\theta}} \ell_{PL}(\boldsymbol{\theta}) = \boldsymbol{\theta}^*$ [2]. Zur besseren Lesbarkeit, werden im Verlauf der Arbeit einige Abkürzungen verwendet. Links in der Tabelle ist die ausgeschriebene Notation und rechts die abgekürzte.

Formel	**Abkürzung**
$\phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{X}_{\setminus A_j})$	$\phi(\boldsymbol{X}_{A_j})$ oder ϕ^j
$\phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})$	$\phi^i(\boldsymbol{X}_{A_j})$ oder $\phi^{j,i}$
$\phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})$	$\phi^i(\boldsymbol{x}'_{A_j})$
$\phi(\boldsymbol{X})$	ϕ

Für den MPLE muss, im Gegensatz zum MLE, nicht die ganze Partitionsfunktion berechnet werden, da sie in (2.21) gekürzt wird. Dadurch wird die Berechnung im Bezug auf den Zustandsraum einfacher, da die Partitionsfunktion der bedingten Wahrscheinlichkeiten $Z_{A_j}(\boldsymbol{\theta}, \boldsymbol{x}^i)$ in (2.21) nur eine Summe über den Zustandsraum von $\boldsymbol{X}_{A_j}$ ist und nicht über den gesamten Zustandsraum von $\boldsymbol{X}$. Da $\boldsymbol{X}_{A_j}$ im Falle des MPLE eine deutlich kleinere Dimension hat, ist die Berechnung des MPLE bezüglich des Zustandsraums einfacher. Im Bezug auf die Stichprobengröße wird sie jedoch schwieriger, da die bedingten Wahrscheinlichkeiten nicht nur vom Parameter, sondern auch von der Beobachtung abhängen und deswegen für jede Stichprobe neu berechnet werden müssen.

Es ist an der Definition der Log-Composite-Likelihood erkennbar, dass diese aus Likelihoods verschiedener bedingter Exponentialverteilungen besteht, welche von den Teilkomponenten $\boldsymbol{X}_{A_j}$ abhängen.

2.4.7 Definition (Likelihood-Komponente der Composite-Likelihood). Wir bezeichnen folgende Funktion als die Likelihood-Komponente von $\ell_{CL}(\boldsymbol{\theta})$ im Bezug auf die Teilkomponente der Zufallsvariable A_j mit $j \in \{1, \dots, m\}$:

$$\ell_{CL}(\boldsymbol{\theta})_{A_j} = \mathbb{E}_{\boldsymbol{\theta}^*}\left[-\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{X})\rangle + \log\left(\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{X}_{\setminus A_j})\rangle)\right)\right] , \quad (2.24)$$

Dafür bezeichnen wir die empirischen Likelihood-Komponente in Bezug auf die Teilkomponente A_j als

$$\hat{\ell}_{CL}(\boldsymbol{\theta})_{A_j} = \frac{1}{N}\sum_{i=1}^{N}\left(-\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}^i)\rangle + \log\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})\rangle)\right)\right) . \quad (2.25)$$

2.4.8 Definition (Reduzierte Likelihood-Komponente). Wir bezeichnen eine empirische und wahre Likelihood-Komponente als reduziert, wenn $\phi_{B_j}(\boldsymbol{X})$ reduziert ist.

Wenn ϕ minimal ist, dann ist ϕ_{B_j} in reduzierter Form für beliebiges $j \in \{1, \dots, m\}$ auch minimal, da ϕ_{B_j} ein Teilvektor von ϕ ist 2.2.7. Daraus können wir den Schluss ziehen, dass eine reduzierte wahre Likelihood-Komponente streng konvex ist.

2.4.9 Beweis. Eine multivariate Funktion ist streng konvex, wenn ihre Hessematrix positiv definit ist. Die Hessematrix einer reduzierten Likelihood-Komponente $\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ sieht folgendermaßen aus:

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_{A_j})}\left[\mathrm{Cov}_{\mathbb{P}(\boldsymbol{X}_{A_j}|\setminus \boldsymbol{X}_{A_j})}(\phi_{B_j}(\boldsymbol{X}))\right] . \quad (2.26)$$

Dies ist an den Ableitungen der Log-Composite-Likelihood (3.48) erkennbar. Damit die Hessematrix der reduzierte Likelihood-Komponente von A_j nicht positiv definit ist, muss ein Vektor $\boldsymbol{v} \in \mathbb{R}^{d_{A_j}}$ existieren, wobei d_{A_j} die Dimension der reduzierten Funktion ϕ_{B_j} ist, sodass $\boldsymbol{v}^T \nabla^2 \ell_{CL}(\boldsymbol{\theta})_{A_j} \boldsymbol{v} = 0$, und damit

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_{A_j})}\left[\boldsymbol{v}^T \mathrm{Cov}_{\mathbb{P}(\boldsymbol{X}_{A_j}|\boldsymbol{X}_{\setminus A_j})}(\phi_{B_j}(\boldsymbol{X}))\boldsymbol{v}\right] = 0$$

gilt. Damit diese Gleichung null ist, muss, wie in Beweis 2.3.5 erkennbar ist,

$$\exists \boldsymbol{v} \in \mathbb{R}^{d_{A_j}} : \forall \boldsymbol{x}_{\setminus A_j} \in \mathcal{X}_{A_j} : \mathrm{Var}_{\mathbb{P}(\boldsymbol{X}_{A_j}|\boldsymbol{x}_{\setminus A_j})}(\langle \boldsymbol{v}, \phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{x}_{\setminus A_j})\rangle) = 0 \quad (2.27)$$

gelten. Da ϕ_{B_j} in reduzierter Form jedoch minimal ist, kann dies nicht für alle $\boldsymbol{x}_{A_j}$ der Fall sein. Somit ist eine wahre reduzierte Log-Composite-Likelihood-Komponente positiv definit.

2.5 Regularisierung

Regularisierung ist eine Methode, bei der zusätzliche Bedingungen zu der Loss-Funktion addiert werden, um weitere Informationen oder Ziele hinzuzufügen. Bei der L1- und der L2-Regularisierung handelt es sich um Methoden, welche Parametervektoren mit möglichst kleinen Parametern bevorzugen, um einfachere Modelle zu erzielen und *Overfitting*, eine zu starke Anpassung an die Daten, zu verhindern [13].
L1-Regularisierung nutzt die 1-Norm und L2-Regularisierung die quadrierte 2-Norm des Parametervektors.

2.5.1 Definition (p-Norm). Sei $\boldsymbol{z} \in \mathbb{R}^d$, dann ist die p-Norm von $\boldsymbol{z}$ definiert als

$$\|\boldsymbol{z}\|_p = \left(\sum_{i=1}^{d} |\boldsymbol{z}_i|^p \right)^{\frac{1}{p}} . \tag{2.28}$$

Sei nun $\ell(\boldsymbol{\theta})$ eine beliebige zu minimierende Loss-Funktion, dann wäre

$$\ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_p^p , \tag{2.29}$$

für $p = 1$ eine L1-regularisierte, und für $p = 2$, eine L2-regularisierte, Loss-Funktion. Hierbei ist λ der Regularisierungsparameter. Umso größer λ ist, umso mehr werden große Parameter „bestraft“, da die Zielfunktion bei komplizierten Parametervektoren größer wird.

2.6 Stichprobenkomplexität und PAC-Schranke

PAC-Learning seht für *Probably Approximatly Correct Learning*. Dabei geht es darum, eine Schranke ε für den Fehler im Lernen, mit einer Wahrscheinlichkeit von $1 - \delta$ zu finden. Für eine PAC-Schranke soll der Algorithmus polynomielle Laufzeit bezüglich des Zustansraumes haben, dies ist für den MLE nicht der Fall, wir werden die von uns berechnete Schranke jedoch trotzdem, wie in [4] als PAC-Schranke bezeichnen. Die Anzahl an Stichproben, die benötigt werden, um einen beliebigen Fehler mit beliebiger Wahrscheinlichkeit zu erreichen, wird als *Stichprobenkomplexität* bezeichnet. In der Literatur werden diese Begriff oft im Zusammenhang mit Klassifikationsverfahren in Verbindung gesetzt [11], hier nutzen wir sie für Parameterschätzung. Die Fehlerschranke kann beliebig gemessen werden, für einen Parametervektor kann es zum Beispiel der Abstand zwischen dem wahren Parameter $\boldsymbol{\theta}^*$ und dem geschätzten Parameter $\hat{\boldsymbol{\theta}}$ sein. In unserem Fall ist dieser Abstand die 1-Norm, von $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$.

Kapitel 3

Stichprobenkomplexität von MLE und MPLE

In diesem Kapitel werden zuerst die PAC-Schranken für den MLE und den MPLE bestimmt und beweisen. Zu jedem Zwischenschritt der Beweise gibt es eine zusammenfassende Anmerkung, in der auf die Schwierigkeiten und besondere Erkenntnisse bezüglich dieses Schrittes eingegangen wird. Beide Beweise haben am Ende einen Diskussionsabschnitt in dem ein Überblick über den gesamten Beweis geschaffen wird. Der letzte Abschnitt dieses Kapitels diskutiert Gemeinsamkeiten und Unterschiede der beiden Beweise, sowie weitere Erkenntnisse.

3.1 MLE

In diesem Abschnitt werden wir die Beweisstrategie von Bradley und Guestrin für die Stichprobenkomplexität von CRFs auf MRFs anwenden, um das folgende Ergebnis herzuleiten:

3.1.1 Theorem (PAC-Schranke des MLE). *Seien* $0 < \lambda < (1-l)\frac{C_{min}^2}{4d^2\phi_{max}^3}N^{-\gamma}$ *mit* $l \in (0,1)$, $\gamma \in (0,1/2)$ *und* C_{min} *der minimale Eigenwert von* $\nabla^2 \ell_L(\boldsymbol{\theta}^*)$ *mit* $0 < C_{min}$. *Dann gilt für* $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \hat{\ell}_L(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p$, *mit einer Wahrscheinlichkeit von mindestens*

$$1 - 2d\exp\left(-\frac{C_{min}^4}{2^9 d^4 \phi_{max}^8} l^2 N^{1-2\gamma}\right) \tag{3.1}$$

die Schranke:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{C_{min}}{4d\phi_{max}^3}\left(-\sqrt{(-N^{-\gamma}+1)}+1\right) \tag{3.2}$$

Im Vergleich dazu sieht die Schranke von Bradley und Guestrin [4] für CRFs folgendermaßen aus:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{C_{min}}{4d\phi_{max}} N^{-\gamma} \tag{3.3}$$

mit einer Wahrscheinlichkeit von mindestens

$$1 - 2d(d+1)\exp\left(-\frac{C_{min}^4}{2^{13}d^4\phi_{max}^8}N^{1-2\gamma}\right) , \tag{3.4}$$

wenn $\lambda = ((C_{min}^2)/(2^6 d^2 \phi_{max}^3))\, N^{-\gamma}$ gewählt wird.
Für $l = \frac{1}{2}$ sind die Bedingungen sehr ähnlich, unsere Schranke ist jedoch mit dem Faktor $(-\sqrt{(-N^{-\gamma}+1)}+1)$ kleiner als die von Bradley und Guestrin mit $N^{-\gamma}$. Dies liegt hauptsächlich an der kleineren Wahl von λ.

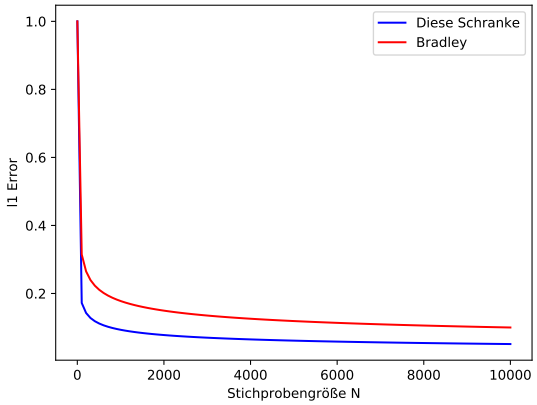


Abbildung 3.1: Vergleich zwischen $N^{-\gamma}$ (Bradley) und $\left(-\sqrt{(-N^{-\gamma}+1)}+1\right)$ (unsere Schranke) für $\gamma = 0.25$

Ein weiterer Unterschied ist der Parameter l. Je kleiner l gewählt wird, desto größer kann λ gewählt, und umso kleiner ist die Wahrscheinlichkeit der Schranke.
Die Stichprobenkomplexität wird durch eine passende Wahl für γ für die Wahrscheinlichkeit und eine spätere Umstellung der Fehler-Schranke nach N berechnet. Um die Stichprobenkomplexität zu berechnen, ist diese Schranke ungeeignet, da sich die Umstellung nach N als schwierig herausstellt. Deswegen benutzen wir für die Berechnung der Stichprobenkomplexität den Faktor $N^{-\gamma}$.

3.1.2 Theorem (Stichprobenkomplexität des MLE). *Seien die Voraussetzungen aus der PAC-Schranke gegeben. Dann gilt* $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1 \leq \epsilon$ *für* $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \hat{\ell}_L(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p$*, mit einer Wahrscheinlichkeit von* $1-\delta$ *bei einer Stichprobengröße von*

$$N \geq \frac{2^5 d^2 \phi_{max}^2}{C_{min}^2 l^2 \epsilon^2} \log\left(\frac{2d}{\delta}\right) . \tag{3.5}$$

3.1.1 Beweis

Betrachten wir zuerst die regularisierte Log-Likelihood-Funktion, für die das Theorem 3.1.2 gilt:

$$f_L(\boldsymbol{\theta}) = \hat{\ell}_L(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p , \tag{3.6}$$

Hierbei handelt es sich bei $p = 1$ um die normale L1-Regularisierung. Bei $p = 2$ gilt $\|\boldsymbol{\theta}\|_2 \neq \|\boldsymbol{\theta}\|_2^2$, weswegen es sich hier nicht um die L2-Regulariesierung handelt.
Der zentrale Punkt des Beweises ist die hier definierte Funktion G.

3.1.3 Definition. Sei $\hat{\ell}_L(\boldsymbol{\theta})$ die empirische Log-Likelihood-Funktion wie in (2.11) beschrieben, $\boldsymbol{\theta}^*$ der wahre Parameter der Verteilung und $\boldsymbol{u} \in \mathbb{R}^d \setminus \{0\}$, dann sei die Funktion $G(\boldsymbol{u})$ definiert als

$$\begin{aligned} G(\boldsymbol{u}) &= \hat{\ell}_L(\boldsymbol{\theta}^* + \boldsymbol{u}) - \hat{\ell}_L(\boldsymbol{\theta}^*) + \lambda(\|\boldsymbol{\theta}^* + \boldsymbol{u}\|_p - \|\boldsymbol{\theta}^*\|_p) \\ &= f_L(\boldsymbol{u} + \boldsymbol{\theta}^*) - f_L(\boldsymbol{\theta}^*) \end{aligned} \tag{3.7}$$

Die empirische Log-Likelihood-Funktion $\hat{\ell}_L(\boldsymbol{\theta})$ ist für Exponentialfamilien konvex 2.3.5 und $\lambda\|\boldsymbol{\theta}^* + \boldsymbol{u}\|_p$ ist auch konvex, da Normen per Definition konvex sind und $\lambda > 0$ gelten soll. Somit ist $G(\boldsymbol{u})$ die Summe zweier konvexer Funktionen und einer Konstanten, $f_L(\boldsymbol{\theta}^*)$, weswegen $G(\boldsymbol{u})$ selber auch konvex ist.
Es gilt $G(0) = f_L(0 + \boldsymbol{\theta}^*) - f_L(\boldsymbol{\theta}^*) = 0$. Wenn $\hat{\boldsymbol{\theta}}$ die Funktion $f_L(\boldsymbol{\theta})$ minimiert, dann befindet sich das Minimum von $G(\boldsymbol{u})$ bei $\hat{\boldsymbol{u}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$.
Dies lässt sich dadurch erkennen, dass der variable Teil der Funktion $G(\boldsymbol{u})$,

$$f_L(\boldsymbol{\theta}^* + \boldsymbol{u}) = \hat{\ell}_L(\underbrace{\boldsymbol{\theta}^* + \boldsymbol{u}}_{a}) + \lambda(\|\underbrace{\boldsymbol{\theta}^* + \boldsymbol{u}}_{a}\|_p) \ ,$$

minimal ist, wenn

$$\begin{aligned} & a = \hat{\boldsymbol{\theta}} \\ \Leftrightarrow\ & \boldsymbol{\theta}^* + \boldsymbol{u} = \hat{\boldsymbol{\theta}} \\ \Leftrightarrow\ & \boldsymbol{u} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \hat{\boldsymbol{u}} \end{aligned}$$

gilt. Daraus folgt, dass $\forall \boldsymbol{u} \in \mathbb{R}^d :\ G(\boldsymbol{u}) \geq G(\hat{\boldsymbol{u}})$ und somit $G(\hat{\boldsymbol{u}}) \leq 0$ gelten muss.

3.1.4 Theorem. *Angenommen es existiert ein $B > 0$, sodass für alle $\boldsymbol{u} \in \mathbb{R}^d$ mit $\|\boldsymbol{u}\|_1 = B$ die Bedingung $G(\boldsymbol{u}) > 0$ gilt, dann muss $\|\hat{\boldsymbol{u}}\|_1 \leq B$ gelten.*

3.1.5 Beweis (Kontraposition). Es gelte die Annahme aus Theorem 3.1.4 und $\|\hat{\boldsymbol{u}}\|_1 > B$. Da es sich bei G um eine konvexe Funktion handelt, müsste es für beliebige zwei Punkte $\boldsymbol{v} \in R^d$ und $\boldsymbol{w} \in \mathbb{R}^d$ und beliebiges $t \in (0, 1)$ die Konvexkombination

$$G(t\boldsymbol{v} + (1 - t)\boldsymbol{w}) \leq tG(\boldsymbol{v}) + (1 - t)G(\boldsymbol{w}) \tag{3.8}$$

geben. Insbesondere müsste dies auch für $\boldsymbol{v} = \hat{\boldsymbol{u}}$ und $\boldsymbol{w} = 0$ gelten. Wenn wir diese Werte in (3.8) einsetzen, dann erhalten wir:

$$G(t\hat{\boldsymbol{u}}) \leq tG(\hat{\boldsymbol{u}}) \leq 0 \tag{3.9}$$

Da $\|\hat{\boldsymbol{u}}\|_1 > B$ gilt, existiert ein t, sodass $\|t\hat{\boldsymbol{u}}\|_1 = t\|\hat{\boldsymbol{u}}\|_1 = B$ gilt, und nach der Konvexkombination wäre dann $G(t\hat{\boldsymbol{u}}) \leq 0$, was ein Widerspruch zu der Annahme ist, dass für alle $\boldsymbol{u} \in \mathbb{R}^d$ mit $\|\boldsymbol{u}\|_1 = B$ die Bedingung $G(\boldsymbol{u}) > 0$ gilt. Daraus folgt das Theorem 3.1.4.

Wenn wir $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ festlegen, dann gilt:

$$G(\boldsymbol{u}) = \hat{\ell}_L(\boldsymbol{\theta}) - \hat{\ell}_L(\boldsymbol{\theta}^*) + \lambda(\|\boldsymbol{\theta}\|_p - \|\boldsymbol{\theta}^*\|_p) \tag{3.10}$$

Nutzen wir nun die Taylor-Formel mit dem Taylorpolynom [8] zweiter Ordnung von $\hat{\ell}_L(\boldsymbol{\theta})$ um den Entwicklungspunkt $\boldsymbol{\theta}^*$ und erhalten:

$$\hat{\ell}_L(\boldsymbol{\theta}) = \hat{\ell}_L(\boldsymbol{\theta}^*) + (\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T(\nabla^2\hat{\ell}_L(\boldsymbol{\theta}^*))\boldsymbol{u} + \frac{1}{6}\sum_{i=1}^{d}\boldsymbol{u}_i\boldsymbol{u}^T\left(\frac{\partial}{\partial\overline{\boldsymbol{\theta}}_i}\nabla\hat{\ell}_L(\overline{\boldsymbol{\theta}})\right)\boldsymbol{u} \tag{3.11}$$

für ein $\overline{\boldsymbol{\theta}} = (\boldsymbol{u} + \alpha\boldsymbol{\theta}^*)$ mit $\alpha \in (0,1)$. Wobei der vierte Term die Lagrangedarstellung des Restgliedes ist. Die Taylor-Formel könnte zu einem beliebigen Grad fortgeführt werden, wodurch sie nur komplizierter wird, deswegen orientieren wir uns an Bradley und Guestrin und wählen das Taylorpolynom dritten Grades mit dem Restglied. Sei außerdem $\|\boldsymbol{u}\|_1 = B$. Damit erhalten wir

$$G(\boldsymbol{u}) = \overbrace{\boxed{(\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T\boldsymbol{u}}}^{\text{Erster Term}} + \overbrace{\boxed{\frac{1}{2}\boldsymbol{u}^T(\nabla^2\hat{\ell}_L(\boldsymbol{\theta}^*))\boldsymbol{u}}}^{\text{Zweiter Term}} + \overbrace{\boxed{\frac{1}{6}\sum_{i=1}^{d}\boldsymbol{u}_i\boldsymbol{u}^T\left(\frac{\partial}{\partial\overline{\boldsymbol{\theta}}_i}\nabla\hat{\ell}_L(\overline{\boldsymbol{\theta}})\right)\boldsymbol{u}}}^{\text{Dritter Term}} + \underbrace{\boxed{\lambda(\|\boldsymbol{\theta}^* + \boldsymbol{u}\|_p - \|\boldsymbol{\theta}^*\|_p}}_{\text{Vierter Term}} . \tag{3.12}$$

Nun wollen wir ein B finden, sodass $\forall\boldsymbol{u} \in \mathbb{R}^d$ mit $\|\boldsymbol{u}\|_1 = B$ die Ungleichung $G(\boldsymbol{u}) > 0$ gilt, denn dann erhalten wir eine obere Schranke für $\|\hat{\boldsymbol{u}}\|_1 = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$.
Wir werden nun eine untere Schranke für $G(\boldsymbol{u})$ in Abhängigkeit von $\|\boldsymbol{u}\|_1 = B$ suchen, indem wir für jeden der oben farbig markierten Terme von $G(\boldsymbol{u})$ eine untere Schranken in Abhängigkeit von $\|\boldsymbol{u}\|_1 = B$ suchen. Danach können wir B so wählen, dass $G(\boldsymbol{u}) > 0$, laut der unteren Schranke, gilt. Für jeden Term wird es im Folgenden einen Abschnitt geben, wo so eine Schranke gesucht wird. Die Schritte für die Berechnung der dafür nötigen Ableitungen befinden sich im Anhang.

3.1.6 Anmerkung (Vorbedingung). Zuerst definieren wir die Funktion f_L (3.6), welche die regularisierte empirische Log-Likelihood darstellt. Danach definieren wir die Funktion $G(\boldsymbol{u})$ (3.10), welche wir mithilfe der Taylorformel als (3.12) definieren. Die Funktion $G(\boldsymbol{u})$ hat die Eigenschaft, dass wenn wir einen Wert $B > 0$ finden, sodass $\forall\boldsymbol{u} \in \mathbb{R}^d$ mit $\|\boldsymbol{u}\|_1 = B$ die Funktion $G(\boldsymbol{u})$ positiv ist, dann gilt für den Schätzer f_L optimierenden Schätzer $\hat{\boldsymbol{\theta}}$ die Fehlerschranke $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq B$. Dies bedeutet, dass wenn es möglich ist, für $G(\boldsymbol{u})$ eine untere Schranke bezüglich $\|u\|_1$ zu finden, dann kann $B = \|\boldsymbol{u}\|_1$ so gewählt werden, dass $G(\boldsymbol{u}) > 0$ gilt. Dafür müssen für alle drei Terme untere Schranken in Abhängigkeit von $B = \|\boldsymbol{u}\|_1$ gefunden werden.

Erster Term

Zuerst wird eine Schranke für $(\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T(\boldsymbol{u})$ gesucht, mit $\boldsymbol{u} = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Jeder Eintrag im Gradient an der Stelle $t \in 1, ..., d$ ist $\nabla\hat{\ell}_L(\boldsymbol{\theta})_t = \frac{\partial}{\partial\boldsymbol{\theta}_t}\hat{\ell}_L(\boldsymbol{\theta})$:

$$\frac{\partial}{\partial\boldsymbol{\theta}_t}\hat{\ell}_L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} -\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\boldsymbol{\theta}}\left[\phi(\boldsymbol{X})_t\right]$$

Somit ist der Gradient insgesamt:

$$\nabla\hat{\ell}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} -\phi(\boldsymbol{x}^i) + \mathbb{E}_{\boldsymbol{\theta}}\left[\phi(\boldsymbol{X})\right] \tag{3.13}$$

Wir versuchen nun eine untere Schranke für den ersten Term bezüglich B zu finden:

$$\begin{aligned}(\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T\boldsymbol{u} &\geq -|(\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T\boldsymbol{u}| \\ &\geq -\|\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)\|_\infty\|\boldsymbol{u}\|_1 \\ &= -\|\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)\|_\infty B \end{aligned} \tag{3.14}$$

Die erste Ungleichung ist eine grobe Abschätzung, da ein Wert entweder positiv ist, dann ist er größer als sein negativer Absolutbetrag, oder der Wert ist negativ und er ist gleich dem negativen Absolutbetrag. Für die zweite Ungleichung haben wir die Hölder-Ungleichung angewendet, um B von dem gesamten Ausdruck zu trennen. Es fehlt nur noch eine Schranke für die Max-Norm von $\nabla\hat{\ell}(\boldsymbol{\theta}^*)$, um eine gesamte untere Schranke für den ersten Term zu erhalten. Dafür betrachten wir nochmal die einzelnen Einträge des Vektors $\nabla\hat{\ell}(\boldsymbol{\theta}^*)$:

$$\nabla\hat{\ell}(\boldsymbol{\theta}^*)_t = \frac{1}{N}\sum_{i=1}^{N} -\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\boldsymbol{\theta}^*}\left[\phi(\boldsymbol{X})_t\right] \tag{3.15}$$

Wir definieren ϕ_{max} so, dass $\phi(\boldsymbol{x})_t$ für ein beliebiges $t \in \{1, \ldots, d\}$ und beliebiges $\boldsymbol{x} \in \mathcal{X}$ nur in $[-\phi_{max}, \phi_{max}]$ liegen kann. Da $\phi(\boldsymbol{X})_t$ auch eine Zufallsvariable ist und $(1/N)\sum_{i=1}^{N}\phi(\boldsymbol{x}^i) = \overline{\phi(\boldsymbol{X})}$ gilt, lässt sich nun die Hoeffding-Ungleichung [14] (siehe A.1.2 im Anhang) anwenden.

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}|\sum_{i=1}^{N}\left(-\phi(\boldsymbol{X}^i)_t + \mathbb{E}_{\boldsymbol{\theta}^*}[\phi(\boldsymbol{X})_t]\right)| \geq \beta\right) &= \mathbb{P}\left(|-\overline{\phi(\boldsymbol{X})_t} + \mathbb{E}_{\boldsymbol{\theta}^*}[\overline{\phi(\boldsymbol{X})_t}]| \geq \beta\right) \\ &\leq 2\exp\left(-\frac{2N^2\beta^2}{\sum_{i=1}^{N}(2\phi_{max})^2}\right) \\ &= 2\exp\left(-\frac{N\beta^2}{2\phi_{max}^2}\right)\end{aligned}$$

Dabei beschreibt der Strich über einer Variable den empirischen Erwartungswert über N Beobachtungen. Wenn wir nun die Bonferroni-Ungleichung [12] (siehe A.1.1) über alle Ereignisse $[\nabla\hat{\ell}(\boldsymbol{\theta}^*)]_t \geq \beta$ anwenden, dann ist dies die Wahrscheinlichkeit, dass ein $t \in$

$\{1, \ldots, d\}$ existiert, sodass $\nabla\hat{\ell}(\boldsymbol{\theta}^*)_t \geq \beta$ gilt. Das ist äquivalent zu der Wahrscheinlichkeit, dass dies für das maximale $\nabla\hat{\ell}(\boldsymbol{\theta}^*)_t$ gilt:

$$\mathbb{P}\left(\|\frac{1}{N}\sum_{i=1}^{N}\left(-\phi(\boldsymbol{X}^i) + \mathbb{E}_{\boldsymbol{\theta}^*}\left[\phi(\boldsymbol{X})\right]\right)\|_\infty \geq \beta\right) = \mathbb{P}\left(\|\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)\|_\infty \geq \beta\right) \leq 2d\exp\left(-\frac{N\beta^2}{2\phi_{max}^2}\right) \tag{3.16}$$

Durch Einsetzen dieses Ergebnisses in (3.14) erhalten wir mit einer Wahrscheinlichkeit von mindestens $1 - 2d\exp\left(-\frac{N\beta^2}{2\phi_{max}^2}\right)$ die untere Schranke:

$$\begin{aligned}(\nabla\hat{\ell}_L(\boldsymbol{\theta}^*))^T(\boldsymbol{u}) &\geq -\|\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)\|_\infty B \\ &\geq -\beta B\end{aligned} \tag{3.17}$$

3.1.7 Anmerkung (Erster Term). Zum Beweis einer unteren Schranke des ersten Terms wird zuerst der Gradient der empirischen Log-Likelihood bestimmt. Als lose untere Schranke für den Term nutzen wir zunächst $y \geq -|y|$. Danach wird der Term mithilfe der Hölder-Ungleich aufgeteilt in $\|\boldsymbol{u}\|_1 = B$, und die Max-Norm des Gradienten von $\nabla\hat{\ell}(\boldsymbol{\theta}^*)$, sodass nur noch eine untere Schranke für die negative Max-Norm des Gradienten, also eine obere Schranke für die Max-Norm, formuliert werden muss.
Wenn jeder Eintrag im Gradienten an der Stelle $\boldsymbol{\theta}^*$ betrachtet wird, dann lässt sich erkennen, dass es sich um die Differenz des empirischen und des echten Erwartungswertes von $\phi(\boldsymbol{X})$ handelt. In diesem Fall lässt sich die Hoeffding-Ungleichung anwenden. Durch diese Ungleichung kann jedem Eintrag eine untere Schranke mit einer gewissen Wahrscheinlichkeit zugeordnet werden. Durch die Bonferroni-Ungleichung und die Berechnung der Gegenwahrscheinlichkeit wird eine obere Schranke für die Max-Norm des Gradienten bestimmt, sodass sie in die Ungleichung eingefügt werden kann.
Diese obere Schranke, sowie ihre Wahrscheinlichkeit, hängen von einem Parameter β ab. Eine wichtige Beobachtung an dieser Stelle ist, dass β sich so wählen lässt, dass die Wahrscheinlichkeit bei einer steigenden Anzahl an Beobachtungen N immer größer wird und die untere Schranke gegen null konvergiert.

Zweiter Term

Der zweite Term von $G(\boldsymbol{u})$ ist $\boldsymbol{u}^T(\nabla^2\hat{\ell}_L(\boldsymbol{\theta}^*))\boldsymbol{u}$. Jeder Eintrag in der Hessematrix, $\nabla^2\hat{\ell}_L(\boldsymbol{\theta})_{t,k} = \frac{\partial}{\partial\boldsymbol{\theta}_t\boldsymbol{\theta}_k}\hat{\ell}_L(\boldsymbol{\theta})$, sieht wie folgt aus:

$$\frac{\partial}{\partial\boldsymbol{\theta}_t\boldsymbol{\theta}_k}\hat{\ell}_L(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_k] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k] \tag{3.18}$$

An den Ableitungen ist erkennbar, dass $\nabla^2\hat{\ell}(\boldsymbol{\theta}) = \nabla^2\left(\frac{1}{N}\sum_{i=1}^{N}A(\boldsymbol{\theta})\right) = \nabla^2 A(\boldsymbol{\theta})$ gilt. Dies liegt daran, dass $A(\boldsymbol{\theta})$ vom Datensatz unabhängig ist. Aus diesem Grund gilt für die Hessematrix von $\ell_L(\boldsymbol{\theta})$:

$$\nabla^2\ell_L(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}*}[\nabla^2 - \langle\phi(\boldsymbol{X}), \boldsymbol{\theta}\rangle + A(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}*}[\nabla^2 A(\boldsymbol{\theta})] = \nabla^2 A(\boldsymbol{\theta}) \tag{3.19}$$

Somit sind die Hessematrizen von $\hat{\ell}_L(\boldsymbol{\theta})$ und $\ell_L(\boldsymbol{\theta})$ äquivalent. Dies ist ein Unterschied zu den von Bradley und Guestrin untersuchten CRFs.

3.1.8 Definition (Minimaler Eigenwert). Sei $\Lambda_{\min}(V)$ definiert als der minimale Eigenwert einer Matrix V. Dieser ist außerdem nach Courant-Fischer definiert als

$$\min_{\|v\|_2=1} v^T V v \,,$$

vergleiche Definition A.2.1.

Definieren wir C_{min} mit $0 < C_{min} \leq \Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*))$ als eine untere Schranke vom minimalen Eigenwert von $\nabla^2 \ell_L(\boldsymbol{\theta}^*)$. So eine Schranke kann es ohne Annahmen über den wahren Parameter $\boldsymbol{\theta}^*$ nicht geben. Da sich unser Beweis stark an dem Beweis in der Publikation von Bradley und Guestrin [4] orientiert und dieser C_{min} nutzt, werden wir dies hier auch tun.
Nun können wir die untere Schranke vom dritten Term bestimmen:

$$\begin{aligned}
\frac{1}{2}\boldsymbol{u}^T(\nabla^2 \hat{\ell}_L(\boldsymbol{\theta}^*))\boldsymbol{u} &= \frac{1}{2}\left(\frac{\boldsymbol{u}^T}{\|\boldsymbol{u}\|_2}(\nabla^2 \hat{\ell}_L(\boldsymbol{\theta}^*))\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}\right)\|\boldsymbol{u}\|_2^2 \\
&\geq \frac{1}{2}\Lambda_{\min}(\nabla^2 \hat{\ell}_L(\boldsymbol{\theta}^*))\|\boldsymbol{u}\|_2^2 \\
&\geq \frac{1}{2d}\Lambda_{\min}(\nabla^2 \hat{\ell}_L(\boldsymbol{\theta}^*))\|\boldsymbol{u}\|_1^2 \\
&= \frac{1}{2d}\Lambda_{\min}(\nabla^2 \hat{\ell}_L(\boldsymbol{\theta}^*))B^2
\end{aligned}$$

Da, wie schon angemerkt, die Hessematrizen der empirischen und der wahren Log-Likelihood für MRFs gleich sind und somit auch die Eigenwerte gleich sind, können wir den nächsten Schritt machen.

$$\begin{aligned}
&= \frac{1}{2d}\Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*))B^2 \\
&\geq \frac{1}{2d}C_{min}B^2 \qquad (3.20)
\end{aligned}$$

Somit erhalten wir eine untere Schranke für den zweiten Term.

3.1.9 Anmerkung (Zweiter Term). Für den zweiten Term wird die Hessematrix der empirischen Log-Likelihood an der Stelle $\boldsymbol{\theta}^*$ betrachet, um festzustellen, dass sie der Hessematrix der wahren Log-Likelihood entspricht. Diese Hessematrix ist für Exponentialfamilien mit minimaler suffizienter Statistik ϕ positiv definit (siehe Beweis 2.3.5) und ihr minimaler Eigenwert größer null. Dies spielt eine Rolle in Ungleichung (3.20). Um die Courant-Fischer Definition des minimalen Eigenwertes zu nutzen, teilen wir $\boldsymbol{u}$ auf beiden Seiten durch $\|\boldsymbol{u}\|_2$, sodass $\|(\boldsymbol{u})/(\|\boldsymbol{u}\|_2)\|_2 = 1$ gilt. Damit der gesamte Term sich nicht ändert, multiplizieren wir ihn mit $\|\boldsymbol{u}\|_2^2$. Nun können wir die Courant-Fischer Definition des minimalen Eigenwertes anwenden, da nach diesem $\boldsymbol{v}^T\nabla^2\hat{\ell}_L(\boldsymbol{\theta}^*)\boldsymbol{v}$ für einen beliebigen Vektor $\boldsymbol{v}$ mit $\|\boldsymbol{v}\|_2 = 1$ nicht kleiner sein kann, als der minimale Eigenwert von $\nabla^2\hat{\ell}_L(\boldsymbol{\theta}^*)$.

Nach diesem Schritt ist die Ungleichung abhängig von der 2-Norm von $\boldsymbol{u}$, dabei wollten wir eine untere Schranke, die abhängig ist von der 1-Norm. Um dies zu gewährleisten nutzen wir die Eigenschaft der 1-Norm und 2-Norm, dass für einen Vektor $\boldsymbol{u} \in \mathbb{R}^d$ die Ungleichung $\|\boldsymbol{u}\|1 \leq \sqrt{d}\|\boldsymbol{u}\|_2$ gilt [15]. Da beide Normen positiv sind gilt auch $\|\boldsymbol{u}\|1^2 \leq d\|\boldsymbol{u}\|_2^2$ und damit $\frac{1}{d}\|\boldsymbol{u}\|1^2 \leq \|\boldsymbol{u}\|_2^2$, was uns den nächsten Schritt erlaubt. Als letztes nutzen wir aus, dass die Hessematrizen der empirischen und der wahren Log-Likelihood äquivalent sind.

In Bradley und Guestrins Beweis für CRFs ist dieser letzte Schritt direkt nicht möglich, da für CRFs die Hessematrizen unterschiedlich sind. Die Hessematrix der wahren Log-Likelihood ist positiv definit, aber nicht unbedingt die der empirischen, weswegen dort zuerst eine Schranke für den minimalen Eigenwert von $\hat{\ell}_L(\boldsymbol{\theta}^*)$ bezüglich des minimalen Eigenwertes von $\ell_L(\boldsymbol{\theta}^*)$ bestimmt und angewendet wird. Dieser Ansatz soll später bei der Betrachtung des MPLE verwendet werden.

Dritter Term

Als nächstes betrachten wir den dritten Term von (3.12): $\frac{1}{6}\sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{u}^T \left(\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_i} \nabla^2 \hat{\ell}_L(\overline{\boldsymbol{\theta}})\right) \boldsymbol{u}$. Dafür betrachten wir zuerst $\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_i} \nabla^2 \hat{\ell}_L(\boldsymbol{\theta})$ an der Stelle $\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_s}[\nabla^2 \hat{\ell}_L(\boldsymbol{\theta})]_{t,k}$.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_L(\boldsymbol{\theta})_{t,k} = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s \phi(\boldsymbol{X})_k \phi(\boldsymbol{X})_t] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k \phi(\boldsymbol{X})_t]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s] \\ + 2\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t] \\ - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k \phi(\boldsymbol{X})_s] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t \phi(\boldsymbol{X})_s] \qquad (3.21)\end{aligned}$$

Für $\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_L(\boldsymbol{\theta})$ gilt nun:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_L(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s \phi(\boldsymbol{X})^{\otimes}] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})^{\otimes}]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s] + 2\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})]^{\otimes} \\ - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]^T - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]\, \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{x})]^T \qquad (3.22)\end{aligned}$$

Hierbei ist $\boldsymbol{v}^{\otimes}$ für einen Vektors $\boldsymbol{v}$ als $\boldsymbol{v}^{\otimes} = \boldsymbol{v}\boldsymbol{v}^T$ definiert.

Wenn wir jetzt für $\boldsymbol{\theta}$ den Parameter $\overline{\boldsymbol{\theta}}$ aus dem dritten Term einfügen erhalten wir:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_L(\overline{\boldsymbol{\theta}}) = \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})_s \phi(\boldsymbol{X})^{\otimes}] - \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})^{\otimes}]\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})_s] + 2\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})_s]\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})]^{\otimes} \\ - \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})]\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]^T - \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}[\phi(\boldsymbol{x})]^T\ .\end{aligned}$$

Dies bedeutet, dass wir nun dem dritten Term eine Schranke zuordnen können. Übersicht halber gilt hier $\phi(\boldsymbol{X}) = \phi$:

$$
\begin{aligned}
&\frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \boldsymbol{u}^T \left(\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_L(\overline{\boldsymbol{\theta}})\right) \boldsymbol{u} \\
=&\frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \boldsymbol{u}^T (\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s \phi^{\otimes}\right] - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi^{\otimes}\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s\right] + 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right]^{\otimes} \\
&- \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\phi_s\right]^T - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\phi_s\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right]^T) \boldsymbol{u} \\
=&\frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s (\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s (\boldsymbol{u}^T \phi^{\otimes} \boldsymbol{u})\right] - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\boldsymbol{u}^T \phi^{\otimes} \boldsymbol{u}\right] \mathbb{E}\left[\phi_s\right] + 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s\right] (\boldsymbol{u}^T\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right]^{\otimes}\, \boldsymbol{u}) \\
&- \boldsymbol{u}^T\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\phi_s\right]^T \boldsymbol{u} - \boldsymbol{u}^T\, \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\phi_s\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi\right]^T \boldsymbol{u}) \\
=&\frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_i (\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s (\boldsymbol{u}^T\phi)^2\right] - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\boldsymbol{u}^T\phi)^2\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s\right] + 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi_s\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\boldsymbol{u}^T\phi)\right]^2 - 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\boldsymbol{u}^T\phi\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\phi \boldsymbol{u}^T \phi_s\right]) \\
=&\frac{1}{6}(\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\boldsymbol{u}^T\phi)^3\right] + 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\boldsymbol{u}^T\phi)\right]^3 - 3\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\boldsymbol{u}^T\phi\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\boldsymbol{u}^T\phi)^2\right]) \qquad (3.23) \\
\geq&\frac{1}{6}(-|\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}[(\boldsymbol{u}^T\phi)^3]| - |2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}[(\boldsymbol{u}^T\phi)]^3| - |3\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}[\boldsymbol{u}^T\phi]\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}[(\boldsymbol{u}^T\phi)^2]|) \qquad (3.24) \\
\geq&\frac{1}{6}(-\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[|(\boldsymbol{u}^T\phi)|^3\right] - 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(|\boldsymbol{u}^T\phi)|\right]^3 - 3\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[|\boldsymbol{u}^T\phi|\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[|(\boldsymbol{u}^T\phi)|^2\right]) \\
\geq&\frac{1}{6}\left(-\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\|\boldsymbol{u}^T\|_1 \|\phi\|_\infty)^3\right] | - 2\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\|\boldsymbol{u}^T\|_1 \|\phi)\|_\infty\right]^3 - 3\,\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\|\boldsymbol{u}^T\|_1 \|\phi\|_\infty\right] \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[(\|\|\boldsymbol{u}^T\|_1 \|\phi\|_\infty)^2\right]\right)
\end{aligned}
$$

Nun können wir ϕ_{max} nutzen, um weiter unzuformen:

$$
\begin{aligned}
\geq&\frac{1}{6}(-\|\boldsymbol{u}\|_1^3 \phi_{max}^3 - 2\|\boldsymbol{u}\|_1^3 \phi_{max}^3 - 3\|\boldsymbol{u}\|_1^3 \phi_{max}^3) \\
=&-\|\boldsymbol{u}\|_1^3 \phi_{max}^3 = -B^3 \phi_{max}^3 \qquad (3.25)
\end{aligned}
$$

Bis (3.24) sind die Rechenschritte Identisch zu denen aus [4]. Wie auch im zweiten Term nutzen wir hier in (3.24) die grobe untere Schranke durch den negativen Absolutbetrag, um in den letzten zwei Ungleichungen die Hölder-Ungleichung zu verwenden und haben nun eine untere Schranke für den dritten Term gefunden.

3.1.10 Anmerkung (Dritter Term). Für den dritten Term haben wird zuerst die dritte Ableitung von $\hat{\ell}(\boldsymbol{\theta}^*)$ bestimmt um damit eine Schranke für den dritten Term zu erhalten. Die Umformungsschritte bis (3.24) gleich denen von Bradley und Guestrin. Danach nutzen wir größtenteils die Hölder-Ungleichung, sowie, dass $\mathbb{E}[|\boldsymbol{X}|] \geq |\,\mathbb{E}[\boldsymbol{X}]| \geq \mathbb{E}[\boldsymbol{X}]$ gilt. In dem Beweis von Bradley und Guestrin wird an der Stelle (3.23) die Jensen-Ungleichung verwendet. Jensens Ungleichung sagt aus, dass $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ gilt, wenn f eine konvexe Funktion ist [12]. Die Funktion $f(x) = x^3$ ist im Negativen nicht konvex und wir können nicht ausschließen, dass $\boldsymbol{u}^T\phi(\boldsymbol{X})$ negativ sein kann. Deswegen wenden wir diese Ungleichung nicht an.
In [4] wird davon ausgegangen, dass $\phi(\boldsymbol{x})$ immer positiv ist und $\min_{\boldsymbol{x}, i \in \{1,\ldots,d\}} \phi(\boldsymbol{x})_i = 0$ gilt, jedoch kann $\boldsymbol{u}$ trotzdem negative Einträge enthalten, sodass $\boldsymbol{u}^T\phi(\boldsymbol{X})$ auch in diesem Fall nicht nur positiv sein kann. Somit darf diese Ungleichung auch in dem Beweis von Bradley und Guestrin nicht anwendet werden.

Vierter Term

Um eine untere Schranke für den letzten Term zu finden, wenden wir die Dreiecksungleichung an:

$$\begin{aligned}
\|\boldsymbol{\theta}^*\|_p &= \|(\boldsymbol{\theta}^* - (\boldsymbol{u}+\boldsymbol{\theta}^*)) + (\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p && \Leftrightarrow \|\boldsymbol{\theta}^*\|_p - \|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p \leq \|\boldsymbol{u}\|_p \\
&\leq \|(\boldsymbol{\theta}^* - (\boldsymbol{u}+\boldsymbol{\theta}^*))\|_p + \|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p && \Leftrightarrow \|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p - \|\boldsymbol{\theta}^*\|_p \geq -\|\boldsymbol{u}\|_p \\
&= \|-\boldsymbol{u}\|_p + \|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p && \Leftrightarrow \lambda(\|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_p - \|\boldsymbol{\theta}^*\|_p) \geq -\lambda\|\boldsymbol{u}\|_p
\end{aligned}$$

An dieser Stelle gibt es eine Fallunterscheidung zwischen $p = 1$ und $p = 2$. Für $p = 1$ gilt

$$\lambda(\|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_1 - \|\boldsymbol{\theta}^*\|_1) \geq -\lambda\|\boldsymbol{u}\|_1 \,,$$

während für $p = 2$

$$\lambda(\|(\boldsymbol{u}+\boldsymbol{\theta}^*)\|_2 - \|\boldsymbol{\theta}^*\|_2) \geq -\lambda\|\boldsymbol{u}\|_2 \geq -\lambda\|\boldsymbol{u}\|_1 \tag{3.26}$$

gilt, so haben wir für beide Fälle die gleiche Schranke.

3.1.11 Anmerkung (Vierter Term). Für den vierten Term wird durch eine geschickte Addition von null und die Anwendung der Dreiecksungleichung eine untere Schranke erzielt.

Schlussfolgerung

Wenn wir nun die Resultate aus (3.17), (3.20), (3.25) und (3.26) zusammenfügen, erhalten wir

$$\begin{aligned}
G(\boldsymbol{u}) &\geq -\beta B + \frac{C_{min}}{2d}B^2 - \phi_{max}^3 B^3 - \lambda B \\
&= B\left(-\beta + \frac{C_{min}}{2d}B - \phi_{max}^3 B^2 - \lambda\right)
\end{aligned} \tag{3.27}$$

mit Wahrscheinlichkeit von mindestens $1 - 2d\exp\left(-\left((\beta^2 N)/(2\phi_{max}^2)\right)\right)$, denn alle Schranken, außer der für den ersten Term, gelten mit Wahrscheinlichkeit 1.

Weiterhin muss nach Voraussetzung $G(\boldsymbol{u}) > 0$ und $B > 0$ ($B = \|\boldsymbol{u}\|_1$) gelten. Dies bedeutet, dass

$$-\beta + \frac{C_{min}}{2d}B - \phi_{max}^3 B^2 - \lambda > 0 \tag{3.28}$$

erfüllt sein muss. Nun können wir, mithilfe der quadratischen Ergänzung, den Bereich berechnen in dem sich B befinden muss, damit die Bedingung in (3.28) gilt.

$$
\begin{aligned}
&-(\beta+\lambda)+\frac{C_{min}}{2d}B-\phi_{max}^3B^2>0\\
\Leftrightarrow&\frac{(\beta+\lambda)}{\phi_{max}^3}-\frac{C_{min}}{2d\phi_{max}^3}B+B^2<0\\
\Leftrightarrow&\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2-\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2-\frac{C_{min}}{2d\phi_{max}^3}B+B^2<0\\
\Leftrightarrow&\frac{(\beta+\lambda)}{\phi_{max}^3}-\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2+\left(B-\frac{C_{min}}{4d\phi_{max}^3}\right)^2<0\\
\Leftrightarrow&\left(B-\frac{C_{min}}{4d\phi_{max}^3}\right)^2<-\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2\\
\Leftrightarrow&-\sqrt{-\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2}<B-\frac{C_{min}}{4d\phi_{max}^3}<\sqrt{-\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2}\\
\Leftrightarrow&-\sqrt{-\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2}+\frac{C_{min}}{4d\phi_{max}^3}<B<\sqrt{-\frac{(\beta+\lambda)}{\phi_{max}^3}+\left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2}+\frac{C_{min}}{4d\phi_{max}^3}
\end{aligned}
\tag{3.29}
$$

Es wäre nun naheliegend, B möglichst nahe an der unteren Schranke zu wählen. Weiterhin sollte B bei einer steigender Stichprobengröße gegen null konvergieren, und damit auch die untere Schranke von B.

Damit diese Bedingung gegeben ist, muss der Term $(\beta+\lambda)$ gegen null konvergieren, dafür müssen sowohl β, als auch λ dieses Kriterium erfüllen, weil beide Variablen nicht negativ sind. Da wir bei der Formulierung der Optimierungs-Funktion die Variable β nicht wählen können, definieren wir sie über eine Funktion von λ. Um die Funktion möglichst einfach zu halten, wählen wir ein Polynom ersten Grades.

Eine einfache Wahl für λ ist z.B. $wN^{-\gamma}$ mit $w,\gamma\in\mathbb{R}^+$, dann ist β der Form $rwN^{-\gamma}=vN^{-\gamma}$ mit $v\in\mathbb{R}^+$.

Wenn wir uns nun die Wahrscheinlichkeit $1-2d\exp\left(-(\beta^2N)/(2\phi_{max}^2)\right)$ anschauen, dann können wir erkennen, dass wenn $(\beta^2N)/(2\phi_{max}^2)$ gegen unendlich konvergiert, der gesamte Ausdruck gegen eins konvergiert, was von uns gewünscht ist, damit die Schranke für den Fehler bei einer steigenden Stichprobengröße wahrscheinlicher wird. Deswegen wollen wir β so wählen, dass $\beta^2N=v^2N^{-2\gamma}N=v^2N^{-2\gamma+1}$ gegen unendlich konvergiert. Dies ist gegeben, wenn $-2\gamma+1>0$ gilt, das bedeutet für γ, dass $\gamma<1/2$ gelten muss. Mit dieser Einschränkung wissen wir nun, dass unter unseren vorherigen Annahmen $\gamma\in(0,1/2)$ gelten muss.

Betrachten wir nun die untere Schranke von B in (3.29) etwas genauer. Damit der Term unter der Wurzel $\sqrt{-(\beta+\lambda)/(\phi_{max}^3)+\left(C_{min}/(4d\phi_{max}^3)\right)^2}$ nicht negativ ist und somit einen gültigen Wert hat, muss $(\beta+\lambda)/\phi_{max}^3<\left(C_{min}/(4d\phi_{max}^3)\right)^2$ gelten. Da $(\beta+\lambda)$ stetig fallend ist, genügt es, wenn die Ungleichung für das kleinste N gilt. Dies muss nicht notwendig $N=1$ sein, weil dies eine unrealistische Stichprobengröße ist. Die Schranke dürfte auch ab

einem größeren N gelten, jedoch wählen wir hier trotzdem einfachheitshalber als kleinste Stichprobengröße $N = 1$. Es muss also nur $(v + w)/\phi_{max}^3 < \left(C_{min}/(4^2 d\phi_{max}^3)\right)^2$ gelten.

$$\begin{aligned} 0 < (v + w) &< \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \\ \Rightarrow 0 < v &< \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \end{aligned}$$

Dies bedeutet, dass wir v als

$$v = l \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \tag{3.30}$$

mit $l \in (0, 1)$ beschreiben können. Da aber noch

$$\begin{aligned} 0 < (v + w) &< \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \\ 0 < l \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} + w &< \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \\ \Leftrightarrow 0 < w &< (1 - l) \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \end{aligned}$$

gelten muss, wählen wir für w

$$w = a(1 - l) \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} \tag{3.31}$$

mit $a \in (0, 1)$. Setzen wir nun die resultierenden λ und β:

$$\lambda = a(1 - l) \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} N^{-\gamma} \qquad \beta = l \frac{C_{min}^2}{4^2 d^2 \phi_{max}^3} N^{-\gamma} \tag{3.32}$$

in die untere Schranke von B ein.

$$\begin{aligned} B &> -\sqrt{-(a(1 - l) + l) \frac{C_{min}^2}{4^2 d^2 \phi_{max}^6} N^{-\gamma} + \left(\frac{C_{min}}{4d\phi_{max}^3}\right)^2} + \frac{C_{min}}{4d\phi_{max}^3} \\ &= -\sqrt{(-(a(1 - l) + l) N^{-\gamma} + 1)} \frac{C_{min}}{4d\phi_{max}^3} + \frac{C_{min}}{4d\phi_{max}^3} \\ &< -\sqrt{(-((1 - l) + l) N^{-\gamma} + 1)} \frac{C_{min}}{4d\phi_{max}^3} + \frac{C_{min}}{4d\phi_{max}^3} \\ &= \frac{C_{min}}{4d\phi_{max}^3} \left(-\sqrt{(-N^{-\gamma} + 1)} + 1\right) \end{aligned} \tag{3.33}$$

Wir können also $B = \frac{C_{min}}{4d\phi_{max}^3} \left(-\sqrt{(-N^{-\gamma} + 1)} + 1\right)$ wählen. Damit gilt:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{C_{min}}{4d\phi_{max}^3} \left(-\sqrt{(-N^{-\gamma} + 1)} + 1\right) \tag{3.34}$$

mit einer Wahrscheinlichkeit von mindestens $1 - 2d \exp\left(-\left((C_{min}^4)/(2^9 d^4 \phi_{max}^8)\right) l^2 N^{1-2\gamma}\right)$. Da die Variable a aus (3.31) beliebig aus dem Intervall $(0, 1)$ gewählt werden kann, reicht es, wenn $\lambda < (1 - l)\left((C_{min}^2)/(4d^2\phi_{max}^3)\right) N^{-\gamma}$ gewählt wird. Damit ist die anfänglich vorgestellte PAC-Schranke bewiesen.

3.1.12 Anmerkung (PAC-Schranke). Dieser Ansatz unterscheidet sich von dem von Bradley und Guestrin dadurch, dass diese zuerst B so gewählt haben, dass λ maximal sein kann und danach B mit der asymptotischen Konvergenzrate von konsistenten MCLEs multiplizieren, wobei die Konvergenzrate selber $N^{-1/2}$ ist [18], in [4] wird jedoch mit $N^{-\gamma}$ mit $0 < \gamma < 1/2$ multipliziert. Dies wird dort getan, um durch γ eine Verbindung zu der Wahrscheinlichkeit herzustellen. Danach wählen Bradley und Guestrin $\lambda = \beta$ so, dass $G(\boldsymbol{u}) > 0$ gilt.
Wir berechnen im Gegensatz zuerst die untere und obere Schranke von B, sodass $G(\boldsymbol{u})$ positiv ist und B unsere Fehlerschranke sein kann. Danach überlegen wir uns eine Belegung für β und λ, sodass die Wahrscheinlichkeit steigen und die untere Schranke von B fallen kann. Wir orientieren uns an dieser unteren Schranke, um B zu wählen
Der Grund wieso wir durch diesen Ansatz eine kleinere Schranke erhalten als Bradley und Guestrin ist, dass wir $\lambda < \beta$ wählen. Für $l = 1/2$ wäre unser β gleich zu dem von Bradley und Guestrin, dadurch aber, dass wir λ kleiner als β wählen, ist unsere untere Schranke kleiner als deren, sodass wir effektiv für B die untere Schranke von Bradley und Guestrin wählen könnten.

Stichprobenkomplexität

Um die Stichprobenkomplexität zu berechnen, wollen wir die Stichprobengröße berechnen, die nötig ist, damit der l1-Fehler $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1$, kleiner ϵ ist mit einer Wahrscheinlichkeit von $1 - \delta$. Zunächst berechnen wir dafür γ in Abhängigkeit von δ und N:

$$
\begin{aligned}
& 2d \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^9 d^4 \phi_{max}^8}\right) = \delta \\
\Leftrightarrow & \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^9 d^4 \phi_{max}^8}\right) = \frac{\delta}{2d} \\
\Leftrightarrow & -\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^9 d^4 \phi_{max}^8} = \log(\delta) - \log(2d) \\
\Leftrightarrow & N^{1-2\gamma} = -\frac{2^9 d^4 \phi_{max}^8}{C_{min}^4 l^2} (\log(\delta) - \log(2d)) \\
\Leftrightarrow & (1-2\gamma) \log N = \log\left(-\frac{2^9 d^4 \phi_{max}^8}{C_{min}^4 l^2} (\log(\delta) - \log(2d))\right) \\
\Leftrightarrow & \gamma = -\frac{1}{2}\left(\frac{\log\left(-\frac{2^9 d^4 \phi_{max}^8}{C_{min}^4 l^2} (\log(\delta) - \log(2d))\right)}{\log N} - 1\right)
\end{aligned} \tag{3.35}
$$

Hier gilt $\gamma \in (0, 1/2)$ nicht mehr für alle δ und N, da für kleinere Stichprobengrößen γ negativ sein muss, damit die Wahrscheinlichkeit $1 - \delta$ gilt. Daraufhin berechnen wir N, sodass $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1 \leq B \leq \epsilon$ gilt. Dafür muss N in Abhängigkeit von ϵ und γ gewählt werden. Nur mit unserer Schranke ist dies schwierig:

$$
\begin{aligned}
& \frac{C_{min}}{4d\phi_{max}^3}\left(-\sqrt{(-N^{-\gamma}+1)}+1\right) \leq \epsilon \\
& \sqrt{(-N^{-\gamma}+1)} \geq 1 - \epsilon\frac{4d\phi_{max}^3}{C_{min}}
\end{aligned} \tag{3.36}
$$

Wenn wir hier quadrieren würden, um nach N umzustellen, dann wäre die Richtung der Ungleichung abhängig von ϵ. Deswegen nutzen wir zur Berechnung der Stichprobenkomplexität eine gröbere Abschätzung:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{C_{min}}{4d\phi_{max}^3}\left(-\sqrt{(-N^{(-\gamma)}+1)}+1\right) \leq \frac{C_{min}}{4d\phi_{max}^3}N^{-\gamma} \leq \epsilon \tag{3.37}$$

Die Abschätzung gilt, da

$$\begin{aligned} -\sqrt{(-N^{(-\gamma)}+1)}+1 &\leq N^{-\gamma} \\ -N^{(-\gamma)}+1 &\leq (N^{-\gamma}-1)^2 \\ -N^{(-\gamma)}+1 &\leq N^{-2\gamma}-2N^{-\gamma}+1 \\ 0 &\leq N^{-2\gamma}-N^{-\gamma} \end{aligned} \tag{3.38}$$

eine wahre Aussage ist. Diese Schranke ist die gleiche wie die von Bradley und Guestrin, weswegen wir nun eine ähnliche Stichprobenkomplexität berechnen werden:

$$\begin{aligned} \frac{C_{min}}{4d\phi_{max}^3}N^{-\gamma} &\leq \epsilon \\ N^{-\gamma} &\leq \frac{4d\phi_{max}^3}{C_{min}}\epsilon \\ -\gamma\log(N) &\leq \log\frac{4d\phi_{max}^3\epsilon}{C_{min}} \\ \gamma\log(N) &\geq \log\frac{C_{min}}{4d\phi_{max}^3\epsilon} \end{aligned}$$

Nun können wir das γ aus (3.35) einfügen:

$$\begin{aligned} &\gamma\log(N) \geq \log\frac{C_{min}}{4d\phi_{max}^3\epsilon} \\ &-\frac{1}{2}\left(\frac{\log\left(-\frac{2^9d^4\phi_{max}^8}{C_{min}^4l^2}\left(\log(\delta)-\log(2d)\right)\right)}{\log N}-1\right)\log(N) \geq \log\frac{C_{min}}{4d\phi_{max}^3\epsilon} \\ &-\frac{1}{2}\log\left(-\frac{2^9d^4\phi_{max}^8}{C_{min}^4l^2}\log\left(\frac{\delta}{2d}\right)\right)+\frac{1}{2}\log(N) \geq \log\frac{C_{min}}{4d\phi_{max}^3\epsilon} \\ &-\log\left(-\frac{2^9d^4\phi_{max}^8}{C_{min}^4l^2}\log\left(\frac{\delta}{2d}\right)\right)+\log(N) \geq 2\log\frac{C_{min}}{4d\phi_{max}^3\epsilon} \\ &\log(N) \geq 2\log\frac{C_{min}}{4d\phi_{max}^3\epsilon}+\log\left(-\frac{2^9d^4\phi_{max}^8}{C_{min}^4l^2}\log\left(\frac{\delta}{2d}\right)\right) \\ &N \geq \exp\left(2\log\frac{C_{min}}{4d\phi_{max}^3\epsilon}+\log\left(-\frac{2^9d^4\phi_{max}^8}{C_{min}^4l^2}\log\left(\frac{\delta}{2d}\right)\right)\right) \\ &N \geq \frac{2^5d^2\phi_{max}^2}{C_{min}^2l^2\epsilon^2}\log\left(\frac{2d}{\delta}\right) \end{aligned}$$

Womit Theorem 3.1.2 bewiesen ist.

3.1.13 Anmerkung (Stichprobenkomplexität). Um die Stichprobenkomplexität zu berechnen, muss zuerst γ so gewählt werden, dass die Wahrscheinlichkeit mindestens so groß ist wie $1 - \delta$, wobei δ beliebig gewählt werden kann. Dafür setzen wir δ gleich der oberen Wahrscheinlichkeitsschranke für einen Fehler und stellen nach γ um. Danach kann γ in die Fehlerschranke eingefügt und nach der Stichprobengröße N umgestellt werden. Mit unserer Schranke ist dies nicht gelungen (siehe (3.36)). Deswegen haben wir eine obere Schranke der Fehlerschranke benutzt (3.38). Diese Schranke ist der von Bradley und Guestrin sehr ähnlich.

3.1.2 Diskussion MLE

Das Ziel des Beweises ist es, eine PAC-Schranke und eine Stichprobenkomplexität für die regularisierte Maximum-Likelihood-Schätzung zu finden. Zuerst wird eine PAC-Schranke ermittelt. Dafür wird die Funktion $G(\boldsymbol{u})$ (3.12) definiert. Diese Funktion hat die Eigenschaft, dass sie an der Stelle $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ einen negativen Wert annimmt. Durch die Konvexität der Funktion können wir feststellen, dass wenn wir ein B finden, sodass $\forall \boldsymbol{u} \in \mathbb{R}^d \wedge \|\boldsymbol{u}\|_1 = B$: $G(\boldsymbol{u}) > 0$ gilt, dann gilt $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq B$. Dies bedeutet, dass B unter diesen Bedingungen eine obere Schranke für den Fehler darstellt. Wenn es nun eine untere Schranke für $G(\boldsymbol{u})$ in Abhängigkeit von $\|\boldsymbol{u}\|_1$ geben würde, dann könnten wir $B = \|\boldsymbol{u}\|_1$ so wählen, dass diese Funktion mit einer bestimmten Wahrscheinlichkeit definitiv positiv ist. Um eine solche Schranke zu finden, wird die Taylorformel verwendet, wodurch es vier Terme gibt, für die eine untere Schranke in Abhängigkeit von B ermittelt werden muss. Wir bestimmen für alle vier Terme eine untere Schranke und wählen am Ende die übrig gebliebenen Parameter und $\|u\|_1 = B$ so, dass die oben beschriebene Bedingung gilt und B möglichst klein ist. Damit können wir eine PAC-Schranke und eine Stichprobenkomplexität aufstellen.

Obwohl der Beweis sich an dem von Bradley und Guestrin orientiert und somit diesem sehr ähnlich ist, gibt es trotzdem einige Unterschiede zwischen MRFs und CRFs, sowie einige andere Ideen. Die Hauptunterschiede zum Beweis von Bradley und Guestrin sind, dass wir nicht davon ausgehen, dass $\phi(\boldsymbol{x})$ für alle $\boldsymbol{x} \in \mathcal{X}$ nur positive Einträgen haben kann. Deswegen ist ϕ_{max} als ϕ_{max} als $\phi_{max} = \max_{\boldsymbol{x}, i \in \{1,\dots,d\}} |\phi(\boldsymbol{x})_i|$ definiert, weiterhin wurde durch die Äquivalenz von $\nabla^2 \ell_L(\boldsymbol{\theta})$ und $\nabla^2 \hat{\ell}_L(\boldsymbol{\theta})$ die Schranke für den zweiten Term einfacher als die von Bradley und Guestrin und gilt mit Wahrscheinlichkeit eins.

Für den dritten Term verwenden wir im Gegensatz zu Bradley und Guestrin die Jensen-Ungleichung nicht, da die Bedingung für diese Ungleichung nicht erfüllt sind. Zum Schluss orientieren wir uns, anstelle B mithilfe der asymptotischen Konvergenzrate zu wählen, an der unteren Schranke für B, wodurch wir eine bessere PAC-Schranke erzielen. Für das Berechnen der Stichprobenkomplexität stoßen wir mit unserer Schranke auf Schwierigkeiten (3.36), weswegen wir stattdessen eine obere Schranke von ihr benutzen (3.38).

3.2 MPLE

Die gleiche Beweisstrategie wie schon beim MLE, wenden wir hier für den konsistenten MCLE an, um eine PAC-Schranke und eine Stichprobenkomplexität herzuleiten. Seinen für die beiden folgenden Theoreme C_{min} als $\min_{j\in\{1,\dots,m\}} \Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*)_{A_j})$ und ρ_{min} als $\min_{t\in\{1,\dots,t\}}(\sum_{j:t\in B_j} \Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*)_{A_j}))$ definiert.

3.2.1 Theorem (PAC-Schranke des MCLE). *Seien* $0 < \lambda < (1-l)\frac{C_{min}^2}{4d^2\phi_{max}^3} N^{-\gamma}$ *mit* $l \in (0,1)$ *und* $\gamma \in (0, 1/2)$. *Dann gilt für* $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \hat{\ell}_{CL}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p$, *mit einer Wahrscheinlichkeit von mindestens*

$$1 - 2d\exp\left(-\frac{\rho_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right) - 2md^2 \exp\left(-\frac{NC_{min}^2}{36 d^2 \phi_{max}^4}\right) \tag{3.39}$$

die Schranke:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \left(1 - \sqrt{-N^{-\gamma}+1}\right) \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} \tag{3.40}$$

3.2.2 Theorem (Stichprobenkomplexität des MCLE). *Angenommen* $\operatorname{argmin}_{\boldsymbol{\theta}} \hat{\ell}_{CL}(\boldsymbol{\theta})$ *sei ein konsistenter Schätzer. Sei* $l \in (0,1)$ *und* λ *entsprechend der PAC-Schranke. Dann gilt* $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1 \leq \epsilon$ *für* $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \hat{\ell}_{CL}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p$, *mit einer Wahrscheinlichkeit von* $1-\delta$ *bei einer Stichprobengröße von*

$$N \geq \log\left(\frac{2d + 2md^2}{\delta}\right) \frac{2^9 d^2 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4 l^2 \epsilon^2} \tag{3.41}$$

3.2.1 Beweis

Die Stichprobenkomplexität des MCLE lässt sich auf eine ähnliche Weise berechnen, wie die des MLE. Auch hier minimieren wir

$$f_{CL}(\boldsymbol{\theta}) = \hat{\ell}_{CL}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_p \tag{3.42}$$

für $p \in \{1,2\}$ und definieren die Funktion $G(\boldsymbol{u})$ mit $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$:

$$G(\boldsymbol{u}) = \hat{\ell}_{CL}(\boldsymbol{\theta}^* + \boldsymbol{u}) - \hat{\ell}_{CL}(\boldsymbol{\theta}^*) + \lambda(\|\boldsymbol{\theta}^* + u\|_p - \|\boldsymbol{\theta}^*\|_p) \tag{3.43}$$

Die Hessematrix der empirischen Log-Composite-Likelihood $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})$ ist eine Summe aus Kovarianzmatrizen, welche immer positiv semidefinit sind. Deswegen ist $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})$ selber auch positiv semidefinit und damit die Funktion $\hat{\ell}_{CL}(\boldsymbol{\theta})$ konvex. Aus diesem Grund können wir auch hier die gleichen Eigenschaften wie bei Theorem 3.1.4 erkennen.
Wir wollen wieder, aus dem gleichen Grund wie beim MLE, eine untere Schranke für $G(\boldsymbol{u})$ in Abhängigkeit von B finden. Wenn wir $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ definieren, dann gilt $\hat{\ell}_{CL}(\boldsymbol{\theta}^* + \boldsymbol{u}) = \hat{\ell}_{CL}(\boldsymbol{\theta})$. Um diese Funktion und $\hat{\ell}_{CL}(\boldsymbol{\theta}^*)$ in Verbindung zu bringen, wenden wir hier die Taylor-Formel für $\hat{\ell}_{CL}(\boldsymbol{\theta})$ um den Punkt $\boldsymbol{\theta}^*$ an. Daraus resultiert

$$\hat{\ell}_{CL}(\boldsymbol{\theta}) = \hat{\ell}_{CL}(\boldsymbol{\theta}^*) + (\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*))^T\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*))\boldsymbol{u} + \frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s\boldsymbol{u}^T\left(\frac{\partial}{\partial\overline{\boldsymbol{\theta}_s}}\nabla\hat{\ell}_{CL}(\overline{\boldsymbol{\theta}})\right)\boldsymbol{u}\,, \tag{3.44}$$

mit $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ und $\overline{\boldsymbol{\theta}} = \boldsymbol{u} + \alpha\boldsymbol{\theta}^*$ für ein $\alpha \in (0,1)$. Jetzt können wir das Resultat aus (3.44) in (3.43) einfügen, um

$$G(\boldsymbol{u}) = \overbrace{(\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*))^T\boldsymbol{u}}^{\text{Erster Term}} + \overbrace{\frac{1}{2}\boldsymbol{u}^T(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*))\boldsymbol{u}}^{\text{Zweiter Term}} + \overbrace{\frac{1}{6}\sum_{i=1}^{d}\boldsymbol{u}_i\boldsymbol{u}^T\left(\frac{\partial}{\partial\overline{\boldsymbol{\theta}}_i}\nabla\hat{\ell}_{CL}(\overline{\boldsymbol{\theta}})\right)\boldsymbol{u}}^{\text{Dritter Term}} \tag{3.45}$$
$$+ \lambda(\|\boldsymbol{\theta}^* + \boldsymbol{u}\|_p - \|\boldsymbol{\theta}^*\|_p$$

zu erhalten. Auch hier werden wir für jeden Term einzeln eine untere Schranke in Abhängigkeit von B suchen. Da der letzte Term, dem aus (3.12) entspricht, werden wir für diesen Beweis die gleiche Schranke aus (3.26) verwenden.

3.2.3 Anmerkung (Vorbedingung). Der Ansatz des Beweises ist der gleiche wie bei schon beim MLE. Der einzige Unterschied ist, dass wir hier anstelle von der regularisierten empirischen Log-Likelihood-Funktion f_L (3.6), die regularisierte empirische Log-Composite-Likelihood-Funktion f_{CL} (3.42) verwenden. Auch hier wird die Funktion $G(\boldsymbol{u})$ (3.43) mithilfe der Taylorformel als (3.45) definiert. Da diese Funktion die gleichen Eigenschaften besitzt wie beim MLE, suchen wir auch hier für jeden Term eine untere Schranke in Abhängigkeit von $\|u\| = B$ um dadurch eine obere Schranke für den Fehler $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ zu finden. Wobei hier $\hat{\boldsymbol{\theta}}$ die Funktion f_{CL} optimiert.

Erster Term

Für den ersten Term berechnen wir den Gradienten $\nabla\hat{\ell}_{CL}(\boldsymbol{\theta})$. Weil es sich hierbei um einen Vektor handelt, betrachten wir von ihm zuerst einen einzelnen Eintrag $\hat{\ell}_{CL}(\boldsymbol{\theta})_t = \frac{\partial}{\partial\boldsymbol{\theta}_t}\hat{\ell}_{CL}(\boldsymbol{\theta})$

$$\frac{\partial}{\partial\boldsymbol{\theta}_t}\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\left(-\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t]\right)$$

Hierbei bedeutet die Notation $j : t \in B_j = \{j \in \{1,\dots,m\} | t \in B_j\}$. Daraus folgt für den Gradienten $\nabla\hat{\ell}_{CL}(\boldsymbol{\theta})$:

$$\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\left(-\phi_{B_j}(\boldsymbol{x}^i) + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})]\right) \tag{3.46}$$

Für alle Einträge $\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_t$ mit $t \in \{1, ..., d\}$ gilt, dass sie im Erwartungswert null sind:

$$\begin{aligned}
&\mathbb{E}_{P_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\left(-\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t]\right)\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{P_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\phi(\boldsymbol{X})_t\right] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{P_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}\left[\phi(\boldsymbol{X'}_{A_j}, \boldsymbol{X}_{\setminus A_j})_t\right]\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\phi(\boldsymbol{X})_t\right] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_{\setminus A_j})}\left[\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}\left[\phi(\boldsymbol{X'}_{A_j}, \boldsymbol{X}_{\setminus A_j})_t\right]\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{P_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\phi(\boldsymbol{X})_t\right] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\phi(\boldsymbol{X})_t\right] = 0
\end{aligned}$$

Weiterhin können wir feststellen, dass jedes $\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_t$ durch das Intervall $[-2\phi_{max}M_t, 2\phi_{max}M_t]$ beschränkt ist, dabei ist M_t mit $t \in \{1, \ldots, d\}$ die Anzahl an Indexmengen B_j mit $j \in \{1, \ldots, m\}$ in denen der Index t vorkommt und ϕ_{max} ist, wie schon im Abschnitt zum MLE, definiert als $\max_{\boldsymbol{x}\in\mathcal{X}, i\in\{1,...,d\}} |\phi(\boldsymbol{x})_i|$. Mithilfe dieser Erkenntnisse wenden wir nun die Hoeffding-Ungleichung an und erhalten

$$\begin{aligned}
&\mathbb{P}\left(|\frac{1}{N}\sum_{i=1}^{N}\sum_{j:\boldsymbol{\theta}_t\in\boldsymbol{\theta}_{B_j}}\left(-\phi_{B_j}(\boldsymbol{x}^i)_t + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t]\right)| > \beta\right) \\
&\leq 2\exp\left(-\frac{\beta^2 N}{8M_t^2\phi_{max}^2}\right).
\end{aligned}$$

Wenn wir nun $M_{max} = \max_{t\in\{1,...,d\}} M_t$ definieren und die Bonferroni-Ungleichung anwenden, erhalten wir:

$$\mathbb{P}\left(\|\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*)\|_\infty > \beta\right) \leq 2d\exp\left(-\frac{\beta^2 N}{8M_{max}^2\phi_{max^2}}\right) \tag{3.47}$$

Als nächstes können wir für den ersten Term von (3.45) die gleichen Schritte wie bei (3.17) nutzen, um die selbe Schranke $(\nabla\hat{\ell}_{CL}(\boldsymbol{\theta}^*))^T\boldsymbol{u} \geq -\beta B$ mit einer Wahrscheinlichkeit von mindestens $1 - 2d\exp\left(-\left((\beta^2 N)/(8M_{max}^2\phi_{max}^2)\right)\right)$ zu erhalten.

3.2.4 Anmerkung (Erster Term). Die Berechnung der Schranke für den ersten Term beim MCLE ist ähnlich zu der Berechnung des ersten Terms beim MLE. Auch hier nutzen wir die Schranke $y \geq -|y|$ und teilen den Term mithilfe der Hölder-Ungleichung in $\|\boldsymbol{u}\|_1 = B$ und $\|\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)\|_\infty$ auf. Um eine untere Schranke zu finden, muss noch eine obere Schranke für die Max-Norm von $\nabla\hat{\ell}_L(\boldsymbol{\theta}^*)$ gefunden werden. Dadurch, dass es sich um den Gradienten des empirischen Log-Composite-Likelihood-Schätzers handelt, ist dies nicht die Differenz des empirischen und des richtigen Erwartungswerts von $\phi(\boldsymbol{X})$. Deswegen nutzen wir die Hoeffding-Ungleichung leicht abgewandelt im Vergleich zu Beweis 3.17. Wir definieren jeden Eintag im Gradienten als eine Summe über Variablen, dessen Erwartungswert null ist, sodass die Hoeffding-Ungleichung angewendet werden kann. Der Rest der Berechnung ist jedoch gleich zu Beweis 3.17. Deswegen ist die Wahrscheinlichkeit für die Schranke kleiner

als beim MLE. Sie ist auch kleiner als die von Bradley und Guestrin, da wir nicht davon ausgehen, dass $\phi(\boldsymbol{X})$ positiv ist, weswegen wir $\phi_{max} = \max_{\boldsymbol{x}\in\mathcal{X},i\in\{1,\dots,d\}} |\phi(\boldsymbol{x})|$ definiert haben.

Zweiter Term

Um als nächstes für den zweiten Term von (3.45) eine Schranke zu finden, müssen wir zunächst die empirische Hessematrix $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})$ bestimmen. Dabei sehen die Einträge dieser Hessematrix $\nabla^2\ell_{CL}(\boldsymbol{\theta})_{t,k}$ wie folgt aus.

$$\frac{\partial}{\partial\boldsymbol{\theta}_t\boldsymbol{\theta}_k}\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t,k\in B_j}\Big(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_k\phi(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_t] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_k]\,\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_t]\Big) \tag{3.48}$$

Die Hessematrix der empirischen Log-Composite-Likelihood selber sieht demnach folgendermaßen aus:

$$\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\left(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}\left[\phi_{B_j}(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})^{\otimes}\right] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}\left[\phi_{B_j}(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\right]^{\otimes}\right) \tag{3.49}$$

Dies bedeutet, dass jeder Eintrag der Hessematrix der echten Log-Composite-Likelihood-Funktion wie folgt aussieht:

$$\frac{\partial}{\partial\boldsymbol{\theta}_t\boldsymbol{\theta}_k}\ell_{CL}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\sum_{j:t,k\in B_j}\Big(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})_k\phi(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})_t] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})_k]\,\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})_t]\Big)\right] \tag{3.50}$$

Und die Hessematrix der echten Log-Composite-Likelihood folgendermaßen:

$$\nabla^2\ell_{CL}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\sum_{j=1}^{m}(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})^{\otimes}] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j},\boldsymbol{X}_{\setminus A_j})]^{\otimes})\right] \tag{3.51}$$

Bei der Betrachtung von (3.49) und (3.51) lässt sich erkennen, dass diese Hessematrizen aus den Hessematrizen $\nabla^2\ell_{CL}(\boldsymbol{\theta})_{A_j}$ beziehungsweise $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})_{A_j}$ der Composite-Likelihood-Komponenten bestehen. Wenn wir jetzt für den zweiten Term den gleichen Ansatz nutzen wollten wie beim MLE, dann stellen wir fest, dass die empirische und die erwartete Hessematrix nicht gleich sind, womit auch die minimalen Eigenwerte nicht gleich sind. Der Anfang der Herleitung ist jedoch ähnlich:

$$\frac{1}{2}\boldsymbol{u}^T\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)\right)\boldsymbol{u} = \frac{1}{2}\sum_{j=1}^{m}\boldsymbol{u}_{B_j}^T\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\boldsymbol{u}_{B_j}$$

Ab hier bis zum Ende des Abschnittes zur unteren Schanke des zweiten Terms nutzen wir die reduzierten Schreibweisen von $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$ und von $\boldsymbol{u}_{B_j}$ (vgl. 2.4.4 und 2.4.8). Für den zweiten Term ergibt sich dadurch kein Unterschied, weil die obere Summe sich dadurch nicht ändert.

$$
\begin{aligned}
&= \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\|\boldsymbol{u}_{B_j}\|_2}\boldsymbol{u}_{B_j}^T \left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right) \frac{1}{\|\boldsymbol{u}_{B_j}\|_2}\boldsymbol{u}_{B_j}\|\boldsymbol{u}_{B_j}\|_2^2 \\
&\geq \frac{1}{2}\sum_{j=1}^{m} \Lambda_{\min}\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\|\boldsymbol{u}_{B_j}\|_2^2 \\
&= \frac{1}{2}\left(\sum_{j=1}^{m} \Lambda_{\min}\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\sum_{t:t\in B_j}\boldsymbol{u}_t^2\right) \\
&= \frac{1}{2}\sum_{t=1}^{d}\underbrace{\left(\sum_{j:t\in B_j}\Lambda_{\min}\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\right)}_{\rho_t}\boldsymbol{u}_t^2
\end{aligned}
\tag{3.52}
$$

Wenn anstelle von $\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$ das richtige $\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ stehen würde, und solange $\forall t \in \{1,\dots,d\}\ \exists j \in \{1,\dots,m\} :\ t \in B_j$ gilt, dann ist die Summe ρ_t nie null, da die Hessematrizen der reduzierten wahren Log-Composite-Likelihood-Komponenten $\nabla^2\ell(\boldsymbol{\theta})_{A_j}$ positiv definit sind 2.4.9. Weiterhin gilt $\exists t \in \{1,\dots,d\} :\ \boldsymbol{u}_t \neq 0$, da $\boldsymbol{u} \neq 0$ nach der Voraussetzung aus Theorem 3.1.4 gilt. Somit wäre der Term (3.52) größer null.
Da die Teil-Matrizen von $\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)$ und $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)$ unterschiedlich sind, werden wir die Differenz zwischen den beiden analysieren. Als erstes berechnen wir die Differenz zwischen den einzelnen Einträgen der Teil-Matrizen. Der Übersicht halber, nutzen wir die abgekürzte Notation $\phi(\boldsymbol{X'}_{A_j}) = \phi(\boldsymbol{X'}_{A_j}, \boldsymbol{X}_{\setminus A_j})$.

$$
\begin{aligned}
&[\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} \\
&= \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X})}\left[\underbrace{\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j})_k\phi(\boldsymbol{X'}_{A_j})_t] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j})_k]\,\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{X}_{\setminus A_j})}[\phi(\boldsymbol{X'}_{A_j})_t]}_{Y}\right] \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi^i(\boldsymbol{X'}_{A_j})_k\phi^i(\boldsymbol{X'}_{A_j})_t] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi^i(\boldsymbol{X'}_{A_j})_k]\,\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi^i(\boldsymbol{X'}_{A_j})_t]\right)
\end{aligned}
$$

Da der obere Term der Erwartungswert der Zufallsvariable Y ist und der untere Term der empirische Erwartungswert von Y ist, können wir an dieser Stelle wieder die Hoeffding-Ungleichung anwenden. Dabei ist zu beachten, dass diese Zufallsvariable höchstens $2\phi_{max}^2$ und mindestens $-2\phi_{max}^2$ sein kann, weil $\mathbb{E}[\phi(\boldsymbol{X'}_{A_j})_k\phi(\boldsymbol{X'}_{A_j})_t]$, sowie $\mathbb{E}[\phi(\boldsymbol{X'}_{A_j})_k]\,\mathbb{E}[\phi(\boldsymbol{X'}_{A_j})_t]$ durch $[-\phi_{max}^2, \phi_{max}^2]$ beschränkt sind.
Daraus folgt:

$$
\begin{aligned}
\mathbb{P}\left(|[\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k}| \geq \epsilon\right) &= \mathbb{P}\left(\left([\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k}\right)^2 \geq \epsilon^2\right) \\
&\leq 2\exp\left(-\frac{N\epsilon^2}{8\phi_{max}^4}\right)
\end{aligned}
$$

Sei $d_{A_j} = |B_j|$, die Anzahl an Elementen in $\phi_{B_j}(\boldsymbol{X})$. Wenn wir nun die Bonferroni-Ungleichung über alle Elemente $t, k \in B_j$ anwenden, erhalten wir:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=1}^{d_{A_j}}\sum_{k=1}^{d_{A_j}}\left([\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k}\right)^2 \geq d_{A_j}^2\epsilon^2\right) \leq 2d_{A_j}^2 \exp\left(-\frac{N\epsilon^2}{8\phi_{max}^4}\right)\\
\Leftrightarrow &\mathbb{P}\left(\sqrt{\sum_{t=1}^{d_{A_j}}\sum_{k=1}^{d_{A_j}}\left([\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k}\right)^2} \geq d_{A_j}\epsilon\right) \leq 2d_{A_j}^2 \exp\left(-\frac{N\epsilon^2}{8\phi_{max}^4}\right)\\
\Leftrightarrow &\mathbb{P}\left(\sqrt{\sum_{t=1}^{d_{A_j}}\sum_{k=1}^{d_{A_j}}\left([\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}]_{t,k}\right)^2} \geq \epsilon'\right) \leq 2d_{A_j}^2 \exp\left(-\frac{N\epsilon'^2}{8d_{A_j}^2\phi_{max}^4}\right) \qquad (3.53)
\end{aligned}$$

Im letzten Schritt haben wir links $\epsilon' = d_{A_j}\epsilon$ gesetzt und damit auf der rechten Seite ϵ mit $\frac{\epsilon'}{d_{A_j}}$ ersetzt. Bei dem Term in der Wurzel handelt es sich um die Frobeniusnorm (vgl. A.2.3) der Matrix $[\ell_{CL}(\boldsymbol{\theta}^*)]_{A_j} - [\hat{\ell}_{CL}(\boldsymbol{\theta}^*)]_{A_j}$.
Als nächstes versuchen wir, für die Frobeniusnorm mit dem minimalen Eigenwert von $\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ in Verbindung zu bringen. Zur Abkürzung der Notation definieren wir $Q^{A_j} = \nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ und $\hat{Q}^{A_j} = \nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$. Wir nutzen hier die Notation aus Definition 3.1.8. Weiterhin definieren wir die Funktion v_{min}.

3.2.5 Definition. Sei V eine quadratische Matrix, dann sei die Funktion $v_{min}(V) : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ definiert als $v_{min}(V) = \text{argmin}_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\, \boldsymbol{v}^T V \boldsymbol{v}$.

Damit lässt sich der minimale Eigenwert einer Matrix V als $v_{min}(V)^T V v_{min}(V)$ schreiben.

$$\Lambda_{\min}(Q^{A_j}) = v_{min}(Q^{A_j})^T Q^{A_j} v_{min}(Q^{A_j})$$

Da $v_{min}(Q^{A_j}) = \text{argmin}_{\|\boldsymbol{v}\|=1}\, \boldsymbol{v}^T Q^{A_j}\boldsymbol{v}$ gilt, kann für jeden anderen Vektor $\boldsymbol{v}$ mit $\|\boldsymbol{v}\|_2 = 1$ der Term $\boldsymbol{v}^T Q^{A_j}\boldsymbol{v}$ nur größer werden kann, somit auch für $v_{min}(\hat{Q}^{A_j})$, was zu dem nächsten Schritt führt.

$$\begin{aligned}
&\leq v_{min}(\hat{Q}^{A_j})^T Q^{A_j} v_{min}(\hat{Q}^{A_j})\\
&= v_{min}(\hat{Q}^{A_j})^T \hat{Q}^{A_j} v_{min}(\hat{Q}^{A_j}) + v_{min}(\hat{Q}^{A_j})^T(Q^{A_j} - \hat{Q}^{A_j})v_{min}(\hat{Q}^{A_j}) \qquad (3.54)\\
&= \Lambda_{\min}(\hat{Q}^{A_j}) + v_{min}(\hat{Q}^{A_j})^T(Q^{A_j} - \hat{Q}^{A_j})v_{min}(\hat{Q}^{A_j})\\
\Leftrightarrow \Lambda_{\min}(\hat{Q}^{A_j}) &\geq \Lambda_{\min}(Q^{A_j}) - v_{min}(\hat{Q}^{A_j})^T(Q^{A_j} - \hat{Q}^{A_j})v_{min}(\hat{Q}^{A_j})
\end{aligned}$$

Hier haben wir in (3.54) zuerst null addiert und dann nach $\Lambda_{\min}(\hat{Q}^{A_j})$ umgestellt. Jetzt nutzen wir aus, dass der maximale Eigenwert einer Matrix V: $\Lambda_{max}(V)$ nach Courant-Fischer definiert ist als $\max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T V \boldsymbol{v}$ (vgl. A.2.1), dies bedeutet, dass $\boldsymbol{v}'^T V \boldsymbol{v}'$ für einen beliebigen Vektor $\boldsymbol{v}'$ mit $\|\boldsymbol{v}'\|_2 = 1$ nicht größer ist als $\Lambda_{max}(V)$.

$$\geq \Lambda_{\min}(Q^{A_j}) - \Lambda_{max}(Q^{A_j} - \hat{Q}^{A_j})$$

Der maximale Eigenwert einer quadratischen Matrix ist gleich dem maximalen Singulärwert dieser Matrix. Der maximale Singulärwert wird auch als Spektralnorm bezeichnet, welche kleiner ist als die Frobeniusnorm [24], was uns den Schritt zur nächsten Ungleichung ermöglicht und einen Zusammenhang zu (3.53) bietet.

$$\geq \Lambda_{\min}(Q^{A_j}) - \sqrt{\sum_{t=1}^{d_{A_j}} \sum_{k=1}^{d_{A_j}} \left(Q_{t,k}^{A_j} - \hat{Q}_{t,k}^{A_j}\right)^2}$$

$$\Leftrightarrow \Lambda_{\min}(Q^{A_j}) - \Lambda_{\min}(\hat{Q}^{A_j}) \leq \sqrt{\sum_{t=1}^{d_{A_j}} \sum_{k=1}^{d_{A_j}} \left(Q_{t,k}^{A_j} - \hat{Q}_{t,k}^{A_j}\right)^2} \tag{3.55}$$

Fügen wir nun (3.55) in (3.53) ein:

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\sum_{t=1}^{d_{A_j}} \sum_{k=1}^{d_{A_j}} \left(Q_{t,k}^{A_j} - \hat{Q}_{t,k}^{A_j}\right)^2} \geq \epsilon'\right) \leq 2d_{A_j}^2 \exp\left(-\frac{N\epsilon'^2}{8d_{A_j}^2 \phi_{max}^4}\right) \\
\Rightarrow &\mathbb{P}\left(\Lambda_{\min}(Q^{A_j}) - \Lambda_{\min}(\hat{Q}^{A_j}) \geq \epsilon'\right) \leq 2d^2 \exp\left(-\frac{N\epsilon'^2}{8d_{A_j}^2 \phi_{max}^4}\right) \\
\Leftrightarrow &\mathbb{P}\left(\Lambda_{\min}(\hat{Q}^{A_j}) \leq \Lambda_{\min}(Q^{A_j}) - \epsilon'\right) \leq 2d_{A_j}^2 \exp\left(-\frac{N\epsilon'^2}{8d_{A_j}^2 \phi_{max}^4}\right)
\end{aligned} \tag{3.56}$$

Wenn wir $\epsilon' = \frac{1}{2}\Lambda_{\min}(Q^{A_j})$ wählen und $C_j = \Lambda_{\min}(Q^{A_j})$ definieren, dann erhalten wir

$$\mathbb{P}\left(\Lambda_{\min}(\hat{Q}^{A_j}) \leq \frac{1}{2}C_j\right) \leq 2d^2 \exp\left(-\frac{NC_j^2}{36d^2\phi_{max}^4}\right) . \tag{3.57}$$

Definieren wir

$$C_{min} = \min_{j \in \{1,\ldots,m\}} \Lambda_{\min}\left(\nabla^2 \ell_{CL}(\boldsymbol{\theta}^*)_{A_j}\right) , \tag{3.58}$$

da wir damit für alle $j \in \{1, \ldots, m\}$ die gleiche Schranke

$$\mathbb{P}\left(\Lambda_{\min}(\hat{Q}^{A_j}) \leq \frac{1}{2}C_j\right) \leq 2d^2 \exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right)$$

aufstellen können.
Jetzt können wir die Bonferroni-Ungleichung über alle $\nabla^2 \hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$ mit $j \in \{1, ..., m\}$ anwenden.

$$\mathbb{P}\left(\exists j \in \{1, \ldots m\} : \Lambda_{\min}\left(\nabla^2 \hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right) \leq \frac{C_j}{2}\right) \leq 2md^2 \exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right) \tag{3.59}$$

Sei ρ_t definiert als $\rho_t = \sum_{j:t\in B_j}\left(\Lambda_{\min}\left(\nabla^2 \ell_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\right)$ und $\hat{\rho}_t$ analog für die empirische Log-Conditional-Likelihood-Funktion definiert als $\hat{\rho}_t = \sum_{j:t\in B_j}\left(\Lambda_{\min}\left(\nabla^2 \hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\right)$. Mithilfe der Ungleichung (3.59) schließen wir auf

$$\mathbb{P}\left(\exists t \in \{1, \ldots, d\} : \hat{\rho}_t \leq \frac{\rho_t}{2}\right) \leq 2md^2 \exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right) . \tag{3.60}$$

Da alle $\Lambda_{\min}\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)$ mit $j \in \{1,\dots,m\}$ mit einer gewissen Wahrscheinlichkeit größer als $C_j/2$ sind, gilt die gleiche Wahrscheinlichkeit für alle Summen von diesen minimalen Eigenwerten, denn damit die Summe $\hat{\rho}_t$ kleiner ist als $\rho_t/2$, muss ein Summand $\hat{C}_j$ existieren, der kleiner ist als $C_j/2$. Dies gilt nur, wenn es sich tatsächlich um eine Summe von minimalen Eigenwerten handelt, weswegen $\forall t \in \{1,\dots,d\}\ \exists j \in \{1,\dots,m\} :\ t \in B_j$ gelten muss. Mit diesen Erkenntnissen werden wir (3.52) fortsetzen und die gleichen Umformungsschritte nutzten wie Bradley und Guestrin [4]:

$$\frac{1}{2}\boldsymbol{u}^T\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)\right)u \geq \frac{1}{2}\sum_{t=1}^{d}\left(\sum_{j:u_t\in u_{A_j}}\Lambda_{\min}\left(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\right)\right)u_t^2$$

In den Klammern steht nun die Definition von $\hat{\rho}_t$, sodass wir diese hiermit ersetzen können und darauf (3.60) anwenden können.

$$\begin{aligned}
&= \frac{1}{2}\sum_{t=1}^{d}\hat{\rho}_t u_t^2 \\
&\geq \frac{1}{2}\sum_{t=1}^{d}\frac{1}{2}\rho_t u_t^2
\end{aligned} \tag{3.61}$$

Wenn wir $\rho_{min} = \min_{t\in\{1,\dots,d\}}\rho_t$ definieren, dann können wir weiter abschätzen.

$$\begin{aligned}
&\geq \frac{1}{4}\rho_{min}\sum_{t=1}^{d}u_t^2 \\
&\geq \frac{\rho_{min}}{4d}\|u\|_1^2 = \frac{\rho_{min}}{4d}B^2\ .
\end{aligned} \tag{3.62}$$

Dies gilt dann mit einer Wahrscheinlichkeit von mindestens $1 - 2md^2\exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right)$, wie es in (3.60) erkennbar ist.

3.2.6 Anmerkung (Zweiter Term). Die Herleitung für die Schranke des zweiten Terms für den MCLE ist am kompliziertesten von den drei Termen. Nach ähnlichen Rechenschritten wie beim MLE schon, stoßen wir in (3.53) auf das Problem, dass die Hessematrizen $\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)$ und $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)$ unterschiedlich sind und damit auch ihre Eigenwerte.
Bradley und Guestrin vergleichen an dieser Stelle die beiden Matrizen, um eine Schranke für den minimalen Eigenwert zu finden. Beide Matrizen bestehen aus den Hessematrizen der wahren, beziehungsweise empirischen, Log-Composite-Likelihood Komponenten, weswegen der Unterschied zwischen den einzelnen Komponenten $\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$ und $\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ betrachtet wird.
Der Vergleich zwischen der wahren und der empirischen Hessematrix der Komponenten ist nötig, da wir eine Schranke wollen, die nicht von den spezifischen Daten in der Stichproben abhängig ist, sondern nur von der Größe der Stichprobe. Der minimale Eigenwert der empirischen Hessematrix kann außerdem null sein, dies liegt daran, dass eine minimale suffiziente Statistik für den Zufallsvektor $\boldsymbol{X}$ nicht minimal ist, wenn ein Teil der Zufallsvariablen, $\boldsymbol{X}_{A_j}$ durch Konstanten ersetzt wird. Für die Hessematrix der wahren Komponenten ist

dies jedoch irrelevant, da dort durch den Erwartungswert über alle Realisierungen von $\boldsymbol{X}$ aufsummiert wird, wie es in Beweis 2.4.9 erkennbar ist. Hier ist es wichtig, darauf zu achten, dass für die minimalen Eigenwerte der Komponenten, die reduzierte Form der Hessematrix verwendet werden muss, da sonst ϕ_{B_j} nicht minimal ist, und damit der minimale Eigenwert null.
Mithilfe der Hoeffding- und der Bonferroni-Ungleichung können wir eine Schranke für die Frobeniusnorm der reduzierten Matrix $\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j} - \nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}$ bestimmen (3.53). Die Frobeniusnorm einer symmetrischen und quadratischen Matrix ist größer als der maximale Eigenwert, deswegen haben wir in (3.59) festgestellt, dass $\Lambda_{\min}(\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j}) - \Lambda_{\min}(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}) \leq \Lambda_{\max}(\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j} - \nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}) \leq \|\nabla^2\ell_{CL}(\boldsymbol{\theta}^*)_{A_j} - \nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j}\|_F$ gilt. Dafür haben wir die Courant-Fischer Definition des minimalen und maximalen Eigenwertes benutzt (vgl. A.2.1). Dieser Beweis stammt aus [26]. Diese Ungleichung konnten wir nutzen, um sie in (3.53) einzusetzen und dadurch eine untere Schranke für eine einzelne Komponente $\Lambda_{\min}(\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}^*)_{A_j})$ mit $j \in \{1,\dots,d\}$ zu finden, sodass die Wahrscheinlichkeit in N steigt. Durch die Bonferroni-Ungleichung beschränken wir die minimalen Eigenwerte der Komponenten. Und damit auch Summen über diese minimalen Eigenwerte, wie $\rho_t = \sum_{B_j: t\in B_j} \Lambda_{\min}(\nabla^2\ell_{CL}(\boldsymbol{\theta}^*))_{A_j}$.
Solange ρ_t für alle $t \in \{1,\dots,d\}$ nicht null ist, daher jede Dimension von ϕ durch die Log-Composite-Likelihood abgedeckt ist, können wir mithilfe von (3.60) die untere Schranke für den dritten Term herleiten (3.62). Dieser Teil der Beweises ist für MRFs beinahe identisch zum Beweis von Bradley und Guestrin, ist jedoch durch die Menge an Ungleichungen, die dafür verwendet wurden, am schwierigsten zu überblicken.
Die Bedingung, dass jede Dimension von dem Composite-Likelihood ϕ abgedeckt sein muss, wird so in [4] nicht explizit erwähnt, muss jedoch auch dort gelten, damit ρ_{min} nicht null ist.

Dritter Term

Für den dritten Term von (3.45) müssen wir $\frac{\partial}{\partial\boldsymbol{\theta}_s}\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})$ bestimmen. Dafür bestimmen wir zuerst die Ableitungen $\frac{\partial}{\partial\boldsymbol{\theta}_s}\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})_{k,t}$. Der Übersicht halber gilt in diesem Abschnitt $\mathbb{E}^i_j[\boldsymbol{X}] = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\boldsymbol{X}]$, sowie $\phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i) = \phi^{j,i}$ und $\phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i)_t = \phi^{j,i}_t$.

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}_s}\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})_{k,t} =& \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t,k\in B_j}\Big(\mathbb{E}^i_j[\phi^{j,i}_k\phi^{j,i}_t\phi^{j,i}_s] - \mathbb{E}^i_j[\phi^{j,i}_k\phi^{j,i}_t]\,\mathbb{E}^i_j[\phi^{j,i}_s] \\ &- \mathbb{E}^i_j[\phi^{j,i}_k]\,\mathbb{E}^i_j[\phi^{j,i}_k\phi^{j,i}_s] + 2\,\mathbb{E}^i_j[\phi^{j,i}_k]\,\mathbb{E}^i_j[\phi^{j,i}_k]\,\mathbb{E}^i_j[\phi^{j,i}_s] - \mathbb{E}^i_j[\phi^{j,i}_t]\,\mathbb{E}^i_j[\phi^{j,i}_t\phi^{j,i}_s]\Big)\end{aligned}$$

Wenn wir nun $\frac{\partial}{\partial\boldsymbol{\theta}_s}\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta})$ insgesamt betrachten, erhalten wir:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}_s}\nabla^2\hat{\ell}_{CL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j:s\in B_j}&(\mathbb{E}^i_j[(\phi^{j,i})^{\otimes}\phi^{j,i}_s] + 2\,\mathbb{E}^i_j[\phi^{j,i}_s]\,\mathbb{E}^i_j[\phi^{j,i}]^{\otimes} - 2\,\mathbb{E}^i_j[\phi^{j,i}_s]\,\mathbb{E}^i_j[(\phi^{j,i})^{\otimes}] \\ &- \mathbb{E}^i_j[\phi^{j,i}]\,\mathbb{E}^i_j[\phi^{j,i}_s\phi^{j,i}]^T - \mathbb{E}^i_j[\phi^{j,i}_s\phi^{j,i}]\,\mathbb{E}^i_j[\phi^{j,i}]^T)\end{aligned} \tag{3.63}$$

Da jetzt die dritte Ableitung bekannt ist, suchen wir nun für den dritten Term eine untere Schranke in Abhängigkeit von B.

$$\frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \boldsymbol{u}^T \left(\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_{CL}(\overline{\boldsymbol{\theta}})\right) \boldsymbol{u}$$

$$= \frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \boldsymbol{u}^T \Bigg(\frac{1}{N}\sum_{i=1}^{N} \sum_{j:s\in B_j} (\mathbb{E}_j^i[(\phi^{j,i})^{\otimes}\phi_s^{j,i}] + 2\,\mathbb{E}[\phi_s^{j,i}]\,\mathbb{E}_j^i[\phi^{j,i}]^{\otimes} - \mathbb{E}_j^i[(\phi^{j,i})^{\otimes}]\,\mathbb{E}_j^i[\phi_s^{j,i}]$$

$$- \mathbb{E}_j^i[\phi^{j,i}]\,\mathbb{E}[\phi_s^{j,i}\phi^{j,i}]^T - \mathbb{E}_j^i[\phi_s^{j,i}\phi^{j,i}]\,\mathbb{E}_j^i[\phi^{j,i}]^T)\Bigg)\boldsymbol{u}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \Bigg(\sum_{j:s\in B_j} (\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T(\phi^{j,i})^{\otimes}\boldsymbol{u}_{B_j})\phi_s^{j,i}] + 2\,\mathbb{E}_j^i[\phi_s^{j,i}]\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}\boldsymbol{u}_{B_j}]^{\otimes}$$

$$- \mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T(\phi^{j,i})^{\otimes}\boldsymbol{u}_{B_j}]\,\mathbb{E}_j^i[\phi_s^{j,i}] - 2\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]\,\mathbb{E}_j^i[\phi_s^{j,i}\boldsymbol{u}_{B_j}^T\phi^{j,i}]^T\Bigg)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{s=1}^{d} \boldsymbol{u}_s \Bigg(\sum_{j:s\in B_j} (\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^2\phi_s^{j,i}] + 2\,\mathbb{E}_j^i[\phi_s^{j,i}]\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]^2$$

$$- \mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^2]\,\mathbb{E}_j^i[\phi_s^{j,i}] - 2\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]\,\mathbb{E}_j^i[\phi_s^{j,i}\boldsymbol{u}_{B_j}^T\phi^{j,i}]^T\Bigg)$$

Hier nutzen wir, dass $\sum_{t\in\{1,\dots,d\}}\sum_{j:t\in B_j}\boldsymbol{u}_t = \sum_{j\in\{1,\dots,m\}}\sum_{t:t\in B_j}\boldsymbol{u}_t$ gilt.

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{j=1}^{m} \left(\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^3] + 2\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]^2 - 3\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]\,\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^2]\right)$$

$$\geq \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{j=1}^{m} \left(-|\,\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^3]| - |2\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]^2| - 3|\,\mathbb{E}_j^i[\boldsymbol{u}_{B_j}^T\phi^{j,i}]|\,|\,\mathbb{E}_j^i[(\boldsymbol{u}_{B_j}^T\phi^{j,i})^2]|\right)$$

$$\geq \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{j=1}^{m} \left(-\,\mathbb{E}_j^i\left[|(\boldsymbol{u}_{B_j}^T\phi^{j,i})|^3\right] - 2\,\mathbb{E}_j^i\left[|\boldsymbol{u}_{B_j}^T\phi^{j,i}|\right]\mathbb{E}_j^i\left[|\boldsymbol{u}_{B_j}^T\phi^{j,i}|\right]^2 - 3\,\mathbb{E}_j^i\left[|\boldsymbol{u}_{B_j}^T\phi^{j,i}|\right]\mathbb{E}_j^i\left[|(\boldsymbol{u}_{B_j}^T\phi^{j,i})|^2\right]\right)$$

$$\geq \frac{1}{N}\sum_{i=1}^{N} \frac{1}{6}\sum_{j=1}^{m} -6\|\boldsymbol{u}_{B_j}\|_1^3\phi_{max}^3 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} -\|\boldsymbol{u}_{B_j}\|_1^3\phi_{max}^3 \tag{3.64}$$

Die Summanden von $\|\boldsymbol{u}_{B_j}\|_1^3$ sind eine Teilmenge der Summanden von $\|\boldsymbol{u}\|_1^3$. Wenn ein Index t existiert, der in mehreren Komponenten B_j vorkommt, dann reicht es für eine untere Schranke aus, $\|\boldsymbol{u}\|_1^3$ mit der maximalen Anzahl an Komponenten zu multiplizieren, in denen der selbe Index vorkommt.

$$\geq \frac{1}{N}\sum_{i=1}^{N} -M_{max}\|\boldsymbol{u}\|_1^3\phi_{max}^3 = -M_{max}\|\boldsymbol{u}\|_1^3\phi_{max}^3 = -M_{max}B^3\phi_{max}^3 \tag{3.65}$$

Somit haben wir eine untere Schranke für den dritten Term erhalten.

3.2.7 Anmerkung (Dritter Term). Die Herleitung der unteren Schranke für den dritten Term für den MCLE ist sehr ähnlich zu der Herleitung des dritten Terms für den MLE. Auch hier wird mehrfach die Hölder-Ungleichung angewendet, um eine untere Schranke herzuleiten und auch hier wird, aus dem gleichen Grund wie beim MLE, im Gegensatz zu Bradley und Guestrin die Jensen-Ungleichung nicht angewendet. Da es sich hier um

eine Summe von Komponenten handelt, wäre ein einfacher Gedanke alle $\|\boldsymbol{u}_{B_j}\|_1^3$ mit $\|\boldsymbol{u}\|_1^3$ nach oben abzuschätzen und mit der Anzahl an Komponenten m zu multiplizieren. Da jedoch die Summanden von allen $\|\boldsymbol{u}_{B_j}\|_1^3$ mit $j \in \{1, \ldots, m\}$ Teilmengen der Summanden von $\|\boldsymbol{u}\|_1^3$ sind, reicht es in (3.65) aus $\|\boldsymbol{u}\|_1^3$ mit der Anzahl an Überschneidungen dieser Mengen zu multiplizieren.

Schlussfolgerung

Wenn wir nun die unteren Schranken für die einzelnen Terme (3.47), (3.62), (3.65), (3.26) in $G(\boldsymbol{u})$ einfügen, erhalten wir für $G(\boldsymbol{u})$ mit einer Wahrscheinlichkeit von mindestens

$$1 - 2d \exp\left(-\frac{\beta^2 N}{8M_{max}^2 \phi_{max}^2}\right) - 2md^2 \exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right)$$

eine untere Schranke. Diese Wahrscheinlichkeit entsteht durch die Nutzung der Bonferroni-Ungleichung über die Wahrscheinlichkeiten, dass die Schranken (3.47) und (3.62) nicht gelten. Bis zu dieser Stelle ist der Beweis beinahe identisch zu dem Beweis von Bradley und Guestrin. Die Unterschiede sind nur entstanden durch die unterschiedliche Definition von ϕ_{max} und dadurch, dass wir für den dritten Term nicht die Jensen-Ungleichung anwenden.

$$\begin{aligned} G(\boldsymbol{u}) &\geq -\beta B + \frac{\rho_{min}}{4d} B^2 - M_{max}\phi_{max}^3 B^3 - \lambda B \\ &= B\left(-(\beta + \lambda) + \frac{\rho_{min}}{4d} B - M_{max}\phi_{max}^3 B^2\right) . \end{aligned} \tag{3.66}$$

Wie sich an dieser Schranke erkennen lässt, müssen damit $G(\boldsymbol{u}) > 0$ gilt, $B > 0$ und

$$-(\beta + \lambda) + \frac{\rho_{min}}{4d} B - M_{max}\phi_{max}^3 B^2 > 0 \tag{3.67}$$

gelten. Als nächstes berechnen wir B, sodass (3.67) erfüllt ist:

$$\begin{aligned} & -(\beta + \lambda) + \frac{\rho_{min}}{4d} B - M_{max}\phi_{max}^3 B^2 > 0 \\ \Leftrightarrow & \frac{\beta + \lambda}{M_{max}\phi_{max}^3} - \frac{\rho_{min}}{4dM_{max}\phi_{max}^3} B + B^2 < 0 \\ \Leftrightarrow & \frac{\beta + \lambda}{M_{max}\phi_{max}^3} - \frac{\rho_{min}}{4dM_{max}\phi_{max}^3} B + B^2 + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6} - \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6} < 0 \\ \Leftrightarrow & \frac{\beta + \lambda}{M_{max}\phi_{max}^3} - \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6} + \left(B - \frac{\rho_{min}}{8dM_{max}\phi_{max}^3}\right)^2 < 0 \\ \Leftrightarrow & -\sqrt{-\frac{\beta + \lambda}{M_{max}\phi_{max}^3} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6}} < B - \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} < \sqrt{-\frac{\beta + \lambda}{M_{max}\phi_{max}^3} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6}} \end{aligned}$$

Genauso wie beim MLE möchten wir B möglichst klein wählen, jedoch soll $G(\boldsymbol{u}) > 0$ weiterhin gelten. Aus diesen Grund orientieren wir uns an der unteren Schranke von B:

$$-\sqrt{-\frac{\beta + \lambda}{M_{max}\phi_{max}^3} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6}} + \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} < B \tag{3.68}$$

Wenn wir wollen, dass $\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}$ bei steigender Stichprobengröße gegen null konvergiert, dann müssen nach dieser Schranke β und λ gegen null konvergieren. Eine einfache Wahl für λ ist wie beim MLE schon $wN^{-\gamma}$ mit $w, \gamma \in \mathbb{R}^+$ und wenn wir β wieder als ein einfaches Polynom ersten Grades von λ definieren, dann wäre $\beta = vN^{-\gamma}$ mit $v \in \mathbb{R}^+$. Da es aber auch wünschenswert wäre, wenn die Wahrscheinlichkeit gegen eins konvergieren würde, gibt es Einschränkungen für β, weil es in der Wahrscheinlichkeit vorkommt:

$$1 - \underbrace{2d \exp\left(-\frac{\beta^2 N}{8M_{max}^2 \phi_{max}^2}\right)}_{t1} - \underbrace{2md^2 \exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right)}_{t2} \tag{3.69}$$

Der Term $t1$ konvergiert bei steigendem N gegen null und enthält keine weiteren unbelegten Variablen, es muss also noch $t2$ gegen null konvergieren, damit (3.69) gegen eins konvergiert. Dafür muss $\lim_{N\to\infty} \beta^2 N = \lim_{N\to\infty} v^2 N^{-2\gamma} N = \lim_{N\to\infty} v^2 N^{1-2\gamma} = \infty$ gelten. Diese Bedingung ist erfüllt, wenn $\gamma < 1/2$ gilt.
Auch hier müssen wir darauf achten, dass der Term unter der Wurzel

$$-\frac{\beta + \lambda}{M_{max}\phi_{max}^3} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6}$$

nicht negativ ist. Dies ist bei $((\beta + \lambda)/(M_{max}\phi_{max}^3)) < ((\rho_{min}^2)/(8^2 d^2 M_{max}^2 \phi_{max}^6))$ gegeben. Solange die Ungleichung für eine kleinste Stichprobengröße gilt, gilt sie auch für alle größeren Stichprobengrößen. Genauso wie beim MLE wählen wir als kleinste Stichprobengröße einfachheitshalber $N = 1$, auch wenn die untere Schranke für die Wahrscheinlichkeit dafür negativ ist.

$$\begin{aligned} 0 &< \frac{v + w}{M_{max}\phi_{max}^3} < \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2 \phi_{max}^6} \\ 0 &< v + w < \frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \\ \Rightarrow 0 &< v < \frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \end{aligned}$$

v können wir mit $l \in (0, 1)$ als

$$v = l\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \tag{3.70}$$

definieren. Jetzt können wir w in Abhängigkeit von v wählen:

$$\begin{aligned} v + w &< \frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \\ l\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} + w &< \frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \\ \Leftrightarrow w &< (1 - l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \end{aligned}$$

Also definieren wir w mit $a \in (0,1)$ als

$$w = a(1-l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3} \tag{3.71}$$

Wenn wir nun (3.70) und (3.71) in (3.68) unter der Bedingung, dass $\lambda = wN^{-\gamma}$ und $\beta = vN^{-\gamma}$ gilt

$$\lambda = a(1-l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma} \qquad \beta = l\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma} \tag{3.72}$$

, einfügen, dann erhalten wir:

$$\begin{aligned}
B &> -\sqrt{-\frac{l\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma} + a(1-l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma}}{M_{max}\phi_{max}^3} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2\phi_{max}^6}} + \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} \\
&= -\sqrt{-(l+a(1-l))\frac{\rho_{min}^2}{8^2 d^2 M_{max}^2\phi_{max}^6}N^{-\gamma} + \frac{\rho_{min}^2}{8^2 d^2 M_{max}^2\phi_{max}^6}} + \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} \\
&= -\sqrt{-(l+a(1-l))N^{-\gamma}+1}\frac{\rho_{min}}{8dM_{max}\phi_{max}^3} + \frac{\rho_{min}}{8dM_{max}\phi_{max}^3} \\
&= \left(1-\sqrt{-(l+a(1-l))N^{-\gamma}+1}\right)\frac{\rho_{min}}{8dM_{max}\phi_{max}^3} \\
&< \left(1-\sqrt{-N^{-\gamma}+1}\right)\frac{\rho_{min}}{8dM_{max}\phi_{max}^3}
\end{aligned} \tag{3.73}$$

Als B können wir nun (3.73) wählen. Dann gilt bei $\lambda < (1-l)\frac{\rho_{min}}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma}$

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \left(1-\sqrt{-N^{-\gamma}+1}\right)\frac{\rho_{min}}{8dM_{max}\phi_{max}^3}\ , \tag{3.74}$$

mit einer Wahrscheinlichkeit von mindestens

$$1 - 2d\exp\left(-\frac{\rho_{min}^4 l^2 N^{1-2\gamma}}{2^{15}d^4 M_{max}^4\phi_{max}^8}\right) - 2md^2\exp\left(-\frac{NC_{min}^2}{36d^2\phi_{max}^4}\right)\ .$$

3.2.8 Anmerkung (PAC-Schranke). Genauso wie beim MLE überlegen wir in welchem Bereich $\|\boldsymbol{u}\|_1 = B$ liegen muss, ausgehend von unsrer unteren Schranke für G, damit $G(\boldsymbol{u}) > 0$ gilt, sodass wir B, wie in Theorem 3.1.4 beschrieben, als eine Schranke für den Fehler nutzen können. Um ein solches B zu finden, wird in (3.68) eine untere Schranke für B aufgestellt, sodass diese Bedingung erfüllt ist. Es bleibt also nur, λ und β so zu wählen, dass diese untere Schranke gültig ist, und sie bei steigendem N gegen null konvergiert, während die Wahrscheinlichkeit gegen eins konvergiert. Dazu muss B etwas größer als diese untere Schranke gewählt werden. Nach einigen Überlegungen wählen wir $\beta = l\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma}$ und $\lambda < (1-l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma}$. Hätten wir $\lambda = (1-l)\frac{\rho_{min}^2}{8^2 d^2 M_{max}\phi_{max}^3}N^{-\gamma}$ gewählt, dann wäre die untere Schranke für B größer als die bisherige, weswegen wir diese als unser B wählen können.

Stichprobenkomplexität

Um die Stichprobenkomplexität zu erhalten, müssen wir die Stichprobengröße berechnen, die nötig ist, damit $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1 \leq \epsilon$ mit einer Wahrscheinlichkeit von mindestens $1-\delta$ gilt. Als erstes berechnen wir γ in Abhängigkeit von δ. Da es schwierig ist, mit der Wahrscheinlichkeit zu rechnen, vereinfachen wir sie zuerst:

$$\begin{aligned}
&2d \exp\left(-\frac{\rho_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right) + 2md^2 \exp\left(-\frac{NC_{min}^2}{36 d^2 \phi_{max}^4}\right) \\
\leq &2d \exp\left(-\frac{\rho_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right) + 2md^2 \exp\left(-\frac{N^{1-2\gamma} C_{min}^2}{36 d^2 \phi_{max}^4}\right) \\
\leq &2d \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right) + 2md^2 \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^1 5 d^4 M_{max} \phi_{max}^8}\right) \\
= &(2d + 2md^2) \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right)
\end{aligned}$$

Diese Abschätzung gilt nur, wenn $\phi_{max} \geq 1$ gilt. Wir können nun diese obere Schranke für die Berechnung von γ verwenden.

$$\begin{aligned}
&\delta = (2d + 2md^2) \exp\left(-\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8}\right) \\
\Leftrightarrow &\log(\delta) = \log(2d + 2md^2) - \frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8} \\
\Leftrightarrow &\frac{C_{min}^4 l^2 N^{1-2\gamma}}{2^{15} d^4 M_{max}^4 \phi_{max}^8} = \log\left(\frac{2d + 2md^2}{\delta}\right) \\
\Leftrightarrow &N^{1-2\gamma} = \log\left(\frac{2d + 2md^2}{\delta}\right) \frac{2^{15} d^4 M_{max}^4 \phi_{max}^8}{C_{min}^4 l^2} \\
\Leftrightarrow &(1-2\gamma)\log(N) = \underbrace{\log\left(\log\left(\frac{2d + 2md^2}{\delta}\right) \frac{2^{15} d^4 M_{max}^4 \phi_{max}^8}{C_{min}^4 l^2}\right)}_{K} \\
\Leftrightarrow &\gamma = \frac{1}{2} - \frac{K}{\log(N)}
\end{aligned} \tag{3.75}$$

Als zweites berechnen wir N, sodass $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_1 < B \leq \epsilon$ gilt. Auch hier können wir die von uns aufgestellte Schranke nicht gut verwenden, da wir wieder auf das gleiche Problem des Quadrierens stoßen würden wie beim MLE. Aus diesem Grund nutzen wir die Schranke

$$\left(1 - \sqrt{-N^{-\gamma} + 1}\right) \frac{\rho_{min}}{8 d M_{max} \phi_{max}^3} \leq N^{-\gamma} \frac{\rho_{min}}{8 d M_{max} \phi_{max}^3} \leq \epsilon \,, \tag{3.76}$$

weil wir schon beim MLE gezeigt haben, dass $\left(1 - \sqrt{-N^{-\gamma} + 1}\right) \leq N^{-\gamma}$ gilt (3.38).

$$\begin{aligned}
&N^{-\gamma} \frac{\rho_{min}}{8 d M_{max} \phi_{max}^3} \leq \epsilon \\
\Leftrightarrow &N^{-\gamma} \leq \frac{8 d M_{max} \phi_{max}^3}{\rho_{min}} \epsilon \\
\Leftrightarrow &N^{-\gamma} \leq \frac{8 d M_{max} \phi_{max}^3}{\rho_{min}} \epsilon \\
\Leftrightarrow &-\gamma \log(N) \leq \log\left(\frac{8 d M_{max} \phi_{max}^3}{\rho_{min}} \epsilon\right)
\end{aligned}$$

Nun können wir γ aus (3.75) einfügen:

$$
\begin{aligned}
&-\left(\frac{1}{2}-\frac{K}{2\log(N)}\right)\log(N) \leq \log\left(\frac{8dM_{max}\phi_{max}^3}{\rho_{min}}\epsilon\right)\\
\Leftrightarrow &-\log(N)+K \leq 2\log\left(\frac{8dM_{max}\phi_{max}^3}{\rho_{min}}\epsilon\right)\\
\Leftrightarrow &N \geq \exp\left(K-2\log\left(\frac{8dM_{max}\phi_{max}^3}{\rho_{min}}\epsilon\right)\right)\\
\Leftrightarrow &N \geq \log\left(\frac{2d+2md^2}{\delta}\right)\frac{2^9d^2M_{max}^2\phi_{max}^2\rho_{min}^2}{C_{min}^4l^2\epsilon^2}
\end{aligned}
$$

Womit Theorem 3.2.2 bewiesen ist.

3.2.9 Anmerkung (Stichprobenkomlexität). Die Vorgehensweise zum Berechnen der Stichprobenkomplexität des MCLE ist identisch zu der für den MLE, der einzige Unterschied ist, dass zwei von unseren Schranken der Terme, eine Wahrscheinlichkeit haben zu scheitern. Deswegen muss die gemeinsame Wahrscheinlichkeit in (3.2.1) vereinfacht werden und erst danach die üblichen Schritte zur Berechnung durchgeführt werden.

3.2.2 Diskussion MCLE

Genauso wie beim MLE ist es das Ziel des Beweises, eine PAC-Schranke und eine Stichprobenkomplexität herzuleiten. Die Struktur des Beweises ist ähnlich wie beim MLE, auch hier haben wir eine Funktion $G(\boldsymbol{u})$ definiert, die die gleichen Eigenschaften hat wie beim MLE. Auch hier entstehen durch die Taylorformel drei Terme, für die eine untere Schranke in Abhängigkeit von $\|\boldsymbol{u}\|_1 = B$ zu finden ist, damit wir B so wählen können, dass $G(\boldsymbol{u}) > 0$ gilt. Der größte Unterschied im Vergleich zu der Herleitung des MLE und die größte Schwierigkeit ist, dass die Log-Composite-Likelihood aus einer Summe von Likelihood-Komponenten besteht. Deswegen müssen wir bei jedem Term darauf achten, dass wir über alle diese Komponenten summieren, um eine untere Schranke zu bestimmen. Der zweite Term fordert dafür die meisten Überlegungen, welche in [4] nicht im Detail beschrieben sind. Eine solche Überlegung ist die Nutzung des reduzierten Vektors ϕ_{B_j} (vgl. 2.4.4). Diese ist wichtig, da die reduzierte Funktion ϕ_{B_j} minimal ist, solange ϕ minimal ist (vgl. 2.2.6), sodass die minimalen Eigenwerte der Hessematrizen der wahren reduzierten Komponenten tatsächlich nicht null sind (siehe Beweis 2.4.9). Eine weitere Feststellung ist, dass die Log-Composite-Likelihood alle Dimensionen von ϕ abdecken muss, damit die Composite-Likelihood konsistent ist. Dies ist eine Bedingung damit dieser Beweis gültig ist. Bradley und Guestrin stellen folgende Proposition auf:

3.2.10 Proposition. *Wenn jede Zufallsvariable im Zufallsvektor in genau einer Komponente vorkommt, dann ist der MCLE konsistent.*

Wie beim Berechnen der unteren Schranke für den zweiten Term angemerkt ist (vgl. (3.55)), ist es nur wichtig, dass alle Dimensionen von ϕ von der Log-Composite-Likelihood abgedeckt sind. Dies ist der Fall, wenn jede Zufallsvariable in mindestens einer Komponente vorkommt. Der Schätzer ist demnach auch konsistent, wenn eine Zufallsvariable in mehreren Komponenten vorkommt.
Die Berechnung der PAC-Schranke läuft ähnlich wie beim MLE ab. Im Gegensatz zum MLE haben diesmal jedoch zwei von unseren Schranken eine Wahrscheinlichkeit zu scheitern, weswegen wir die Bonferroni-Ungleichung anwenden, um eine obere Schranke für die Wahrscheinlichkeit aufzustellen, dass eine dieser Schranken scheitert. Basierend darauf berechnen wir, mit den gleichen Überlegungen wie beim MLE, eine PAC-Schranke, die nicht direkt zur Berechnung der Stichprobenkomplexität geeignet ist. Deswegen nutzen wir eine obere Schranke der PAC-Schranke, wodurch ein ähnliches Resultat entsteht wie bei Bradley und Guestrin.

3.3 Diskussion

Betrachten wir nun die Gemeinsamkeiten zwischen den Schranken für den MLE und den MCLE. Beide Schranken hängen von den Werten C_{min} und ϕ_{max} ab. Der Wert ϕ_{max} ist in beiden Fällen der größte absolute Wert, den ein Eintrag im Vektor $\phi(\boldsymbol{X})$ annehmen kann. Es wäre jedoch möglich, die minimale suffiziente Statistik ϕ mit dem Faktor $1/\phi_{max}$ zu skalieren, wodurch der wahre Parameter $\boldsymbol{\theta}^*$ mit ϕ_{max} skaliert wird, sodass ϕ_{max} immer eins ist. Die Wahrscheinlichkeitsverteilung würde sich so auch nicht ändern, da

$$\langle \frac{1}{\phi_{max}} \phi(\boldsymbol{X}), \phi_{max}\boldsymbol{\theta} \rangle = \langle \phi(\boldsymbol{X}), \boldsymbol{\theta} \rangle$$

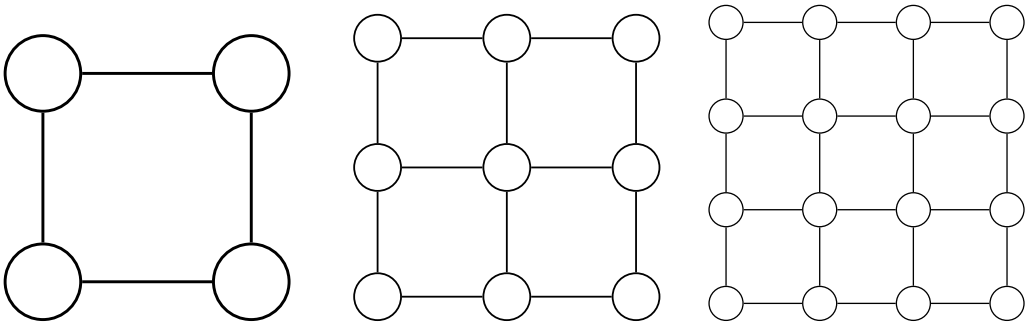
gilt. Der minimale Eigenwert C_{min} ändert sich jedoch, sodass durch diese Umrechnung der ehemalige Einfluss von ϕ_{max} auch durch C_{min} ausgedrückt wird.
Der Wert von C_{min} ist jedoch für MCLE und MLE unterschiedlich definiert. Im Falle des MCLE ist er definiert als $C_{min} = \min_{j\in\{1,\dots,m\}} \Lambda_{\min}(\nabla^2 \ell_{A_j}(\boldsymbol{\theta}^*))$ und im Falle des MLE definiert als $C_{min} = \Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*))$.
An beiden Schranken lässt sich gut erkennen, dass die Stichprobenkomplexität für die jeweilig definierten C_{min} kleiner ist, je größer C_{min} ist. Es stellt sich jedoch das Problem, dass beide Definitionen vom - unter realen Umständen unbekannten - wahren Parameter $\boldsymbol{\theta}^*$ abhängen. Und obwohl $C_{min} > 0$ gilt, kann der Wert beliebig klein sein. Die Schranken in [4] und [26] nutzen jedoch auch diesen Parameter. Ein weiterer Nachteil ist, dass der Regularisierungsparameter für MLE und MCLE so gewählt werden musste, dass der Beweis an den Stellen (3.67) und (3.28) funktioniert. Es wäre wünschenswerter eine Schranke zu erhalten, die von einem frei wählbaren λ abhängt.
Betrachten wir weiter den Wert der berechneten Stichprobenkomplexität des MLE. Es lässt sich feststellen, dass laut ihr die Anzahl an Stichproben, die benötigt werden sehr hoch ist.

Sei zum Beispiel $d = 21, C_{min} = 0,01, l = 1/2$ und $\phi_{max} = 1$. Der Wert von ϕ_{max} und von d entspricht einem binären Zufallsvektor mit einem 3×3-Gittergraphen als MRF. Wenn wir einen Fehler von $\epsilon = 21$ mit einer Wahrscheinlichkeit von $1 - \delta = 0,67$ erhalten wollen, dann bräuchten wir 6203305 Stichproben. Diese Anzahl an Stichproben scheint für ein so kleines Modell und einen relativ großen Fehler unrealistisch. Deswegen werden wir in synthetischen Experimenten diese Vermutung prüfen.

Kapitel 4

Experimente

In diesem Kapitel betrachten wir einige synthetische Experimente, um unsere Schranken zu prüfen. Für die Experimente benutzen wir quadratische Gittergraphen der Größen $n = 4,\ n = 9,\ n = 16$.



(a) 2×2-Gitter **(b)** 3×3-Gitter **(c)** 4×4-Gitter

Abbildung 4.1: Beispiel Gittergraphen

Diese kleinen Größen werden gewählt, da die Größe ihres Zustandsraums mit 2^4, 2^9 und 2^{16} noch klein ist, und die Berechnung des MLE möglich. Weiterhin beschränken wir uns auf Ising-Modelle, binäre MRFs, da für diese die minimale suffiziente Statistik $\phi(\boldsymbol{X})$ bekannt ist [28].

$$\mathbb{P}(\boldsymbol{x}) = \exp\left(\sum_{s\in V} \boldsymbol{\theta}_s \boldsymbol{x}_s + \sum_{(s,t)\in E} \boldsymbol{\theta}_{st} \boldsymbol{x}_s \boldsymbol{x}_t - A(\boldsymbol{\theta})\right) \tag{4.1}$$

Obwohl der Regularisierungsparameter laut PAC-Schranke nicht null sein soll, dafür aber beliebig nah an null, wurde zur Vereinfachung der Experimente keine Regularisierung verwendet. Zur Anlehnung an [4] wurde $l = 1/2$ für die Schranke gewählt.
Wir haben für jede Gittergröße zufällig gleichverteilt Parametervektoren $\boldsymbol{\theta}$ aus $[-5, 5]^d$ gewählt und die Hessematrix, so wie sie in (3.18) beschrieben ist, bestimmt. Daraufhin haben wir den dazugehörigen minimalen Eigenwert berechnet. Dies haben wir solange wiederholt, bis wir 500 Parametervektoren hatten, deren minimale Eigenwerte über das

Intervall $(10^{-8}, 10^{-3})$ spannen. Danach wurden für jedes Modell und jeden Parameter 10000 Strichproben gezogen. Diese nutzen wir um für verschiedene Stichprobengrößen die L1-Abweichung zwischen dem wahren $\boldsymbol{\theta}^*$ und dem von MLE bzw. dem MPLE bestimmten $\hat{\boldsymbol{\theta}}$ zu berechnen. Das ziehen der Stichproben, sowie die Schätzung des MLE und MPLE werden mithilfe von der PX-Software[1] berechnet.

Der MLE und der MPLE von MRFs existieren häufig nicht in geschlossener Form, dies bedeutet, dass ein Optimierungsverfahren notwendig ist, um diese Werte zu berechnen. Die Log-Likelihood ist, wie wir sie definiert haben, konvex und ihr Minimum ist der MLE. Das Minimum einer konvexen Funktion befindet sich an der Stelle, wo der Gradient null ist. Ein mögliches Optimierungverfahren zur bestimmung des Punktes an dem der Gradient null ist, ist der Gradientenabstieg. Dabei wird der Parameter iterativ mit einer bestimmten Schrittweite in Richtung des negativen Gradienten angepasst, sodass nach einer endlichen Anzahl an Iterationen der Parameter, wo der Gradient null ist gefunden wird. Eine weitere Version des Gradientenabstiesverfahrens ist der beschleunigte Gradientenabstieg von Nesterov, bei dem die Schrittweite bei jeder Iteration angepasst wird, sodass weniger Iterationen notwendig sind [27]. Die PX-Software nutzt den beschleunigten Gradientenabstieg. Für diese Experimente wurden tausend Iterationen dieser Methode durchgeführt, da danach kein Unterschied im Gradienten mehr bemerkbar war.

4.1 MLE

Als erstes betrachten wir den l1-Fehler des MLE aus unseren Schätzungen, in Abhängigkeit von der Stichprobengröße N und vergleichen ihn mit dem Fehler, den unsere Schranke für diese Stichprobengröße beschreiben würde. Dafür haben wir die Stichprobengrößen $\{10, 25, 50, 75, 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$ gewählt und für jeden Parametervektor und jede dieser Stichprobengrößen den MLE bestimmt. Da es sich um 500 Parametervektoren mit unterschiedlichen C_{min} handelt die alle einen unterschiedlichen Fehler und unterschiedliche Schranken aufweisen, ist der Fehler in Form eines Boxplots dargestellt.

[1] https://randomfields.org

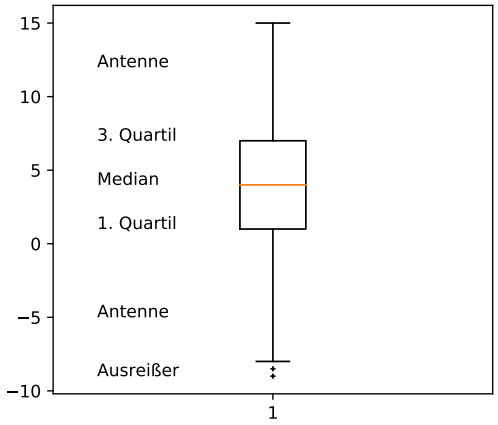
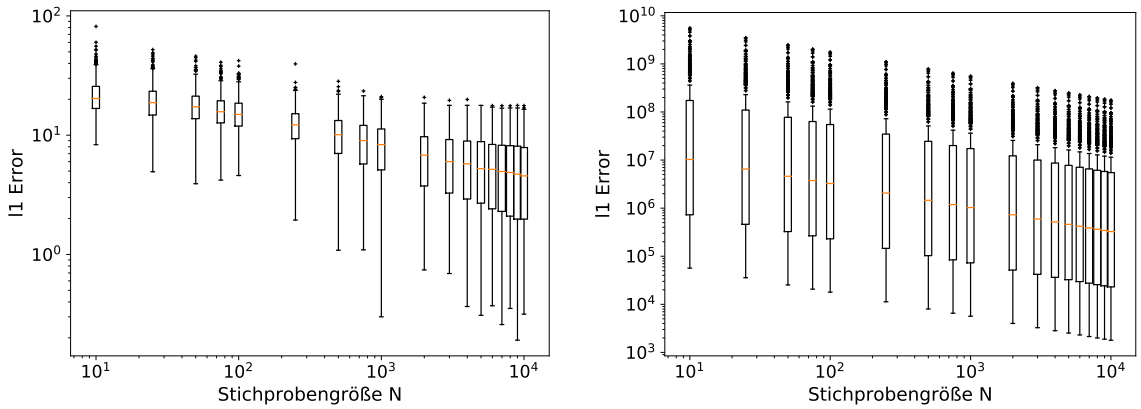


Abbildung 4.2: Beispiel Boxplot

Der Kasten eines Boxplots beinhaltet 50% der Datenpunkte, sodass die obere Grenze des Kastens das obere Quartil Q_3 und die untere Grenze das untere Quartil Q_1 darstellt, während der gelbe Strich im Kasten den Median darstellt. Die Antennen reichen bis zu den minimalen und maximalen Datenpunkt, solange diese innerhalb von $Q_3 + (Q_3 - Q_1)1.5$ beziehungsweise $Q_1 - (Q_3 - Q_1)1.5$ liegen. Wenn der minimale oder maximale Datenpunkt außerhalb dieser Reichweite liegt, dann ist er als Ausreißer mit dem Zeichen + über beziehungsweise unter der Antenne gekennzeichnet.

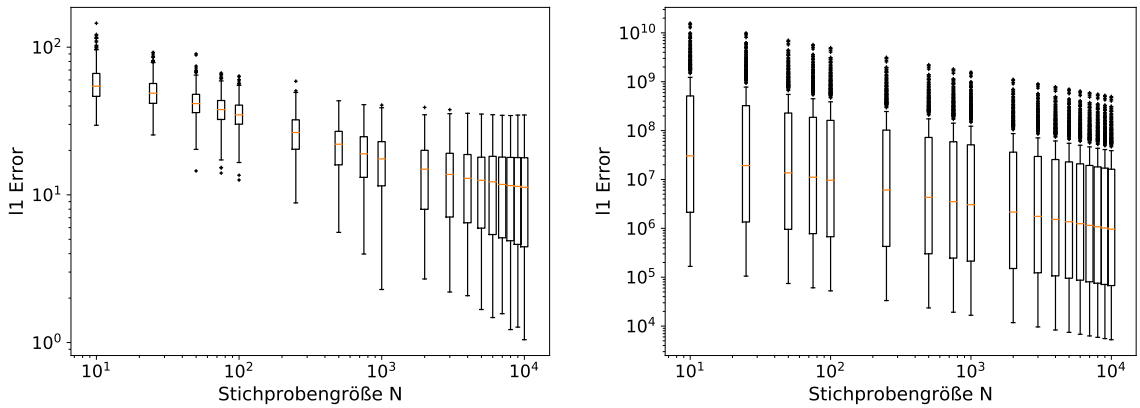
Die x- und y-Achsen aller folgenden Grafiken logarithmische skaliert.



(a) Empirisch aus den Daten

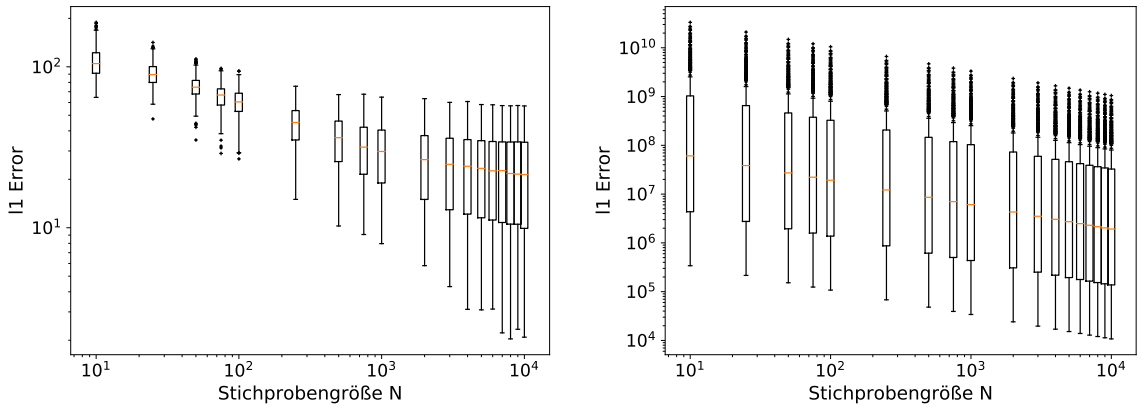
(b) Nach der Schranke

Abbildung 4.3: l1-Fehler des MLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 2×2-Gitter in log-log-scale.



(a) Empirisch aus den Daten **(b)** Nach der Schranke

Abbildung 4.4: l1-Fehler des MLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 3×3-Gitter in log-log-scale.



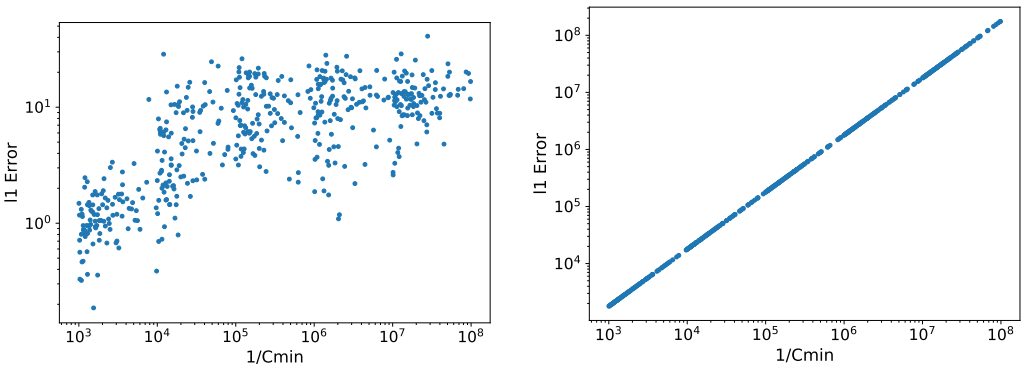
(a) empirisch aus den Daten **(b)** Nach der Schranke

Abbildung 4.5: l1-Fehler des MLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 4×4-Gitter in log-log-scale.

Dadurch, dass die Parametervektoren so gewählt wurden, dass ihre dazugehörigen C_{min} gleichmäßig über ein Intervall verteilt sind, sieht der Boxplot für die obere Schranke sehr gleichmäßig aus. Die Varianz, die durch die Antennen erkennbar ist, ist durch die logarithmische Skalierung verzehrt, sodass die Abweichung nach unten größer erscheint als nach oben, obwohl dies nicht unbedingt der Fall ist. Dafür lässt sich mithilfe dieser Skalierung der Zusammenhang zwischen der Schranke und dem Schätzer erkennen.

Es lässt sich hier jedoch trotzdem an der y-Achse erkennen, dass der Fehler für die empirischen Schätzungen deutlich kleiner ist als der Fehler den wir erhalten, wenn wir die Stichprobenkomplexität 3.1.2 nach dem Fehler ϵ umstellen.

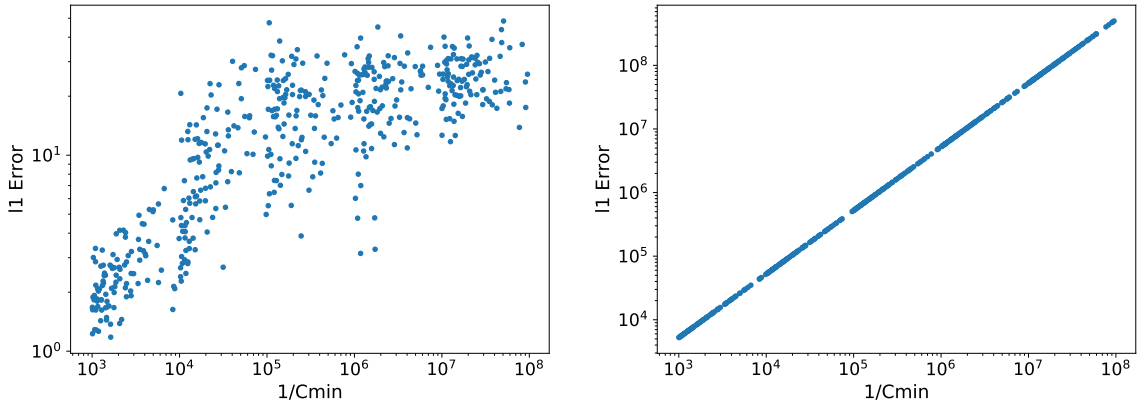
Da die Stichprobenkomplexität vom minimalen Eigenwert der Hessematrix der Log-Likelihood C_{min} abhängt, betrachten wir diese Abhängigkeit auf unseren empirischen Daten.



(a) geschätzter l1-Fehler

(b) l1-Fehler-Schranke für die Wahrscheinlichkeit $1 - \delta = 0.67$

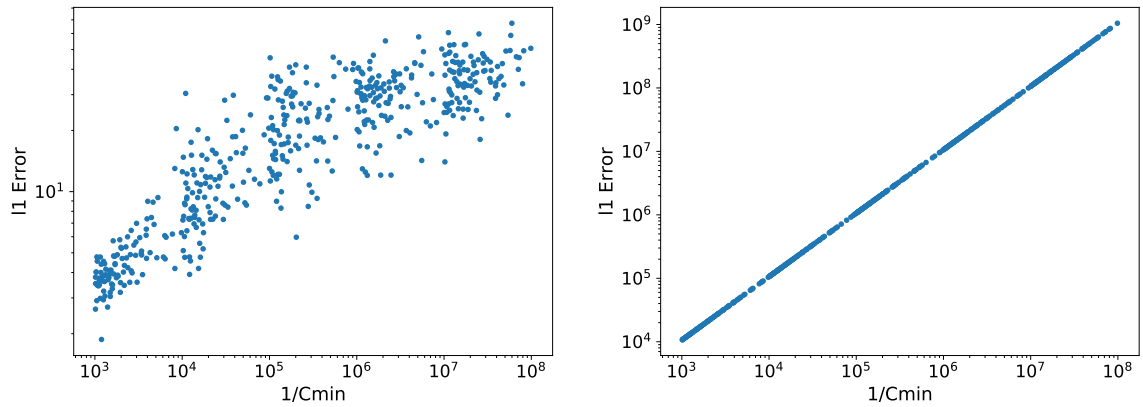
Abbildung 4.6: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 2×2-Gitter in log-log-scale.



(a) geschätzter l1-Fehler

(b) l1-Fehler-Schranke für die Wahrscheinlichkeit $1-\delta = 0.67$

Abbildung 4.7: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 3×3-Gitter in log-log-scale.



(a) geschätzter l1-Fehler

(b) l1-Fehler-Schranke für die Wahrscheinlichkeit $1-\delta = 0.67$

Abbildung 4.8: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 4×4-Gitter in log-log-scale.

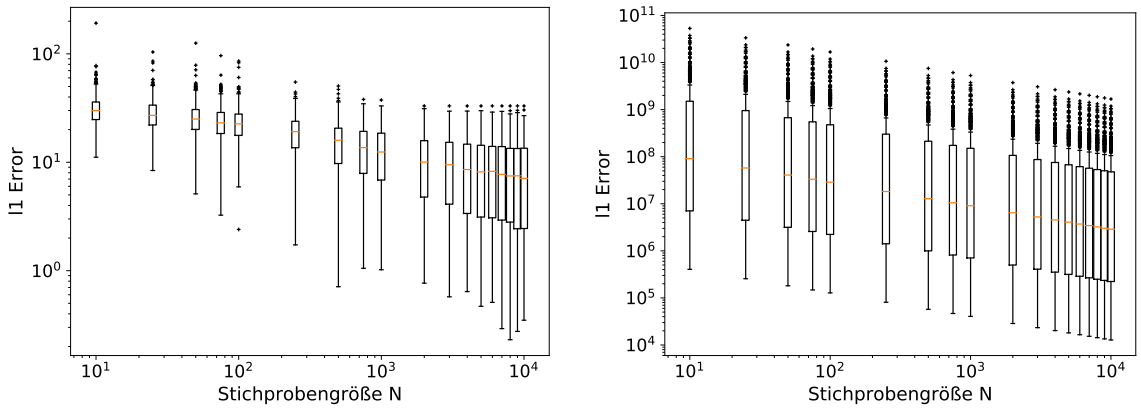
Für jede Gittergröße sind zwei Diagramme nebeneinander zu sehen. Auf der linken Seite ist der l1-Fehler des MLE bei einer Stichprobengröße von $N = 10000$ in Abhängigkeit von C_{min} für jeden Parametervektor als Punkt zu sehen. Auf der rechten Seite ist der Fehler, für die Wahrscheinlichkeit $1 - \delta = 0.67$ und $l = 1/2$, aus der Schranke zu sehen.

Es lässt sich erkennen, dass der Fehler durch die Maximum-Likelihood, den wir aus den synthetischen Daten berechnet haben, und die Fehler-Schranke weit voneinander entfernt sind. Der l1-Fehler aus den Daten ist deutlich niedriger als die Schranke, aber das Verhalten bezüglich C_{min} ist tatsächlich ähnlich.

4.2 MPLE

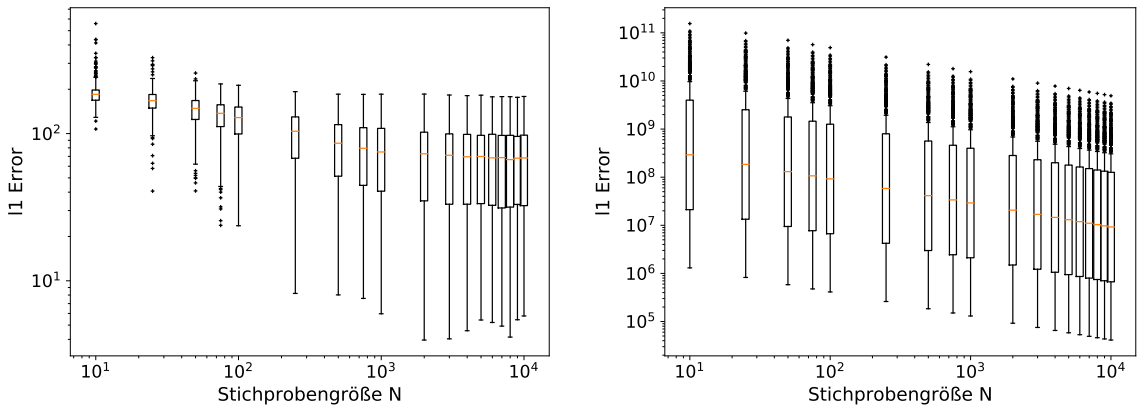
Für den MPLE ist jede Zufallsvariable im Zufallsvektor eine Komponente. Dadurch, dass in der minimalen Statistik von Ising-Modellen für jeden Knoten es einen Eintrag gibt, sodass $\rho_t = \Lambda_{\min}(\nabla^2 \hat{\ell}_{CL}(\boldsymbol{\theta}^*)_t)$ gilt, sind ρ_{min} und C_{min} gleich. Deswegen ist die Stichprobenkomplexität in diesem Fall nur von C_{min} abhängig. Auf Grund der anderen Definition von C_{min} (vgl. (3.58)) für den MPLE, wurde C_{min} gemäß dieser Definition für alle Parametervektoren neu berechnet.

Die Experimente sind ähnlich zu denen bei der Betrachtung des MLEs, mit identischen Parameterverktoren und verwendeten Daten, sowie δ und l. Die neu berechneten C_{min} lagen auch auf einem ähnlichen Intervall.



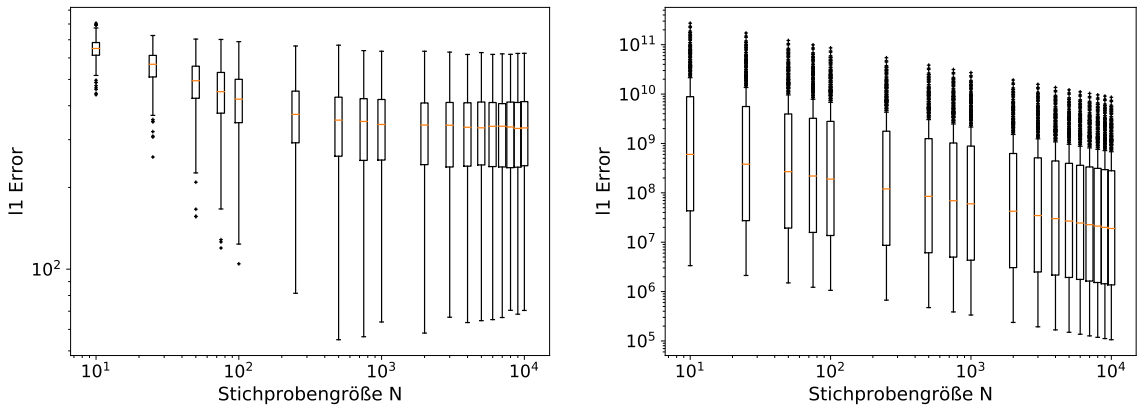
(a) Empirisch aus den Daten **(b)** Nach der Schranke

Abbildung 4.9: l1-Fehler des MPLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 2×2-Gitter in log-log-scale.



(a) Empirisch aus den Daten **(b)** Nach der Schranke

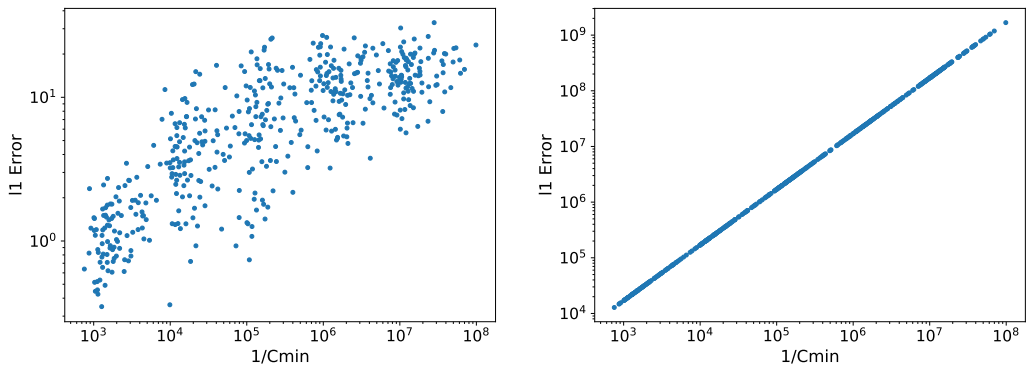
Abbildung 4.10: l1-Fehler des MPLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 3×3-Gitter in log-log-scale.



(a) Empirisch aus den Daten **(b)** Nach der Schranke

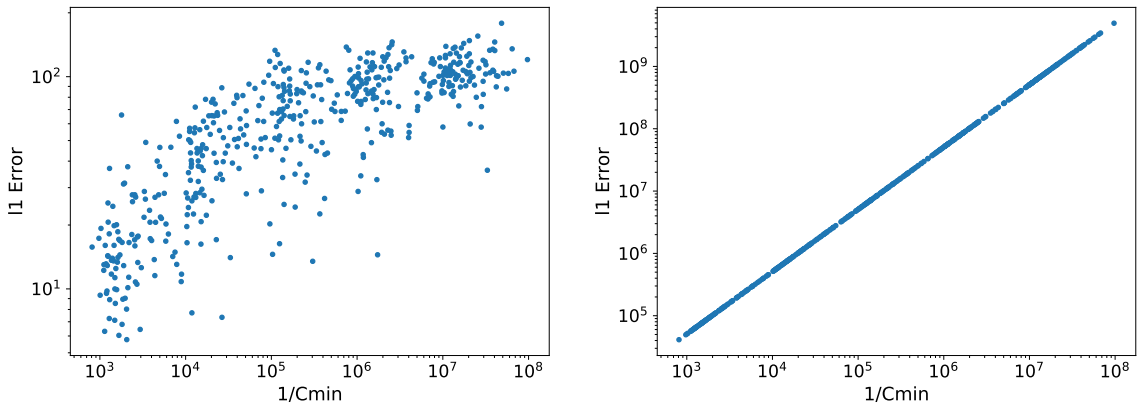
Abbildung 4.11: l1-Fehler des MPLE über 500 Parameter in Abhängigkeit von der Stichprobengröße bei dem 4×4-Gitter in log-log-scale.

Auch beim MPLE können wir beobachten, dass der Fehler deutlich kleiner ist als die berechnete Schranke andeutet. Obwohl der Fehler kleiner ist, scheint er jedoch in Abhängigkeit von N, vor allem für das 3×3- und 4×4-Gitter langsamer zu fallen als die Schranke.



(a) geschätzter l1-Fehler **(b)** l1-Fehler-Schranke für die Wahrscheinlichkeit $1 - \delta = 0.67$

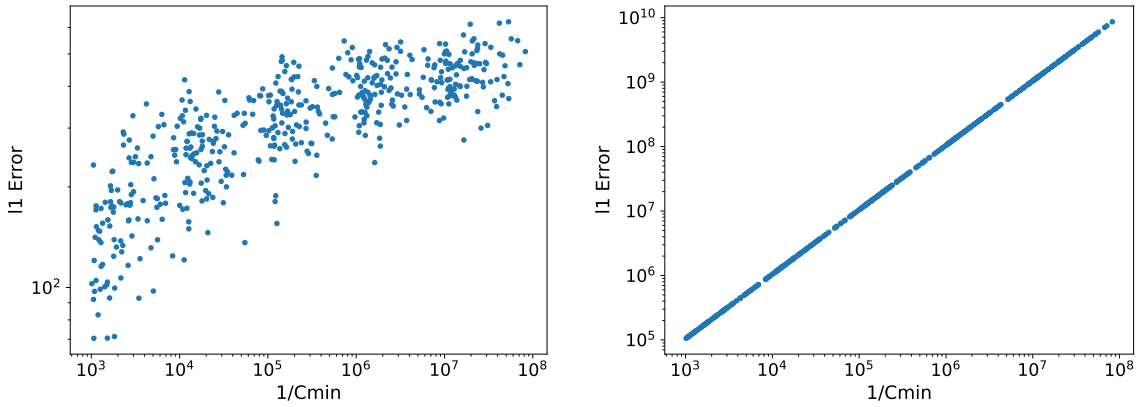
Abbildung 4.12: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 2×2-Gitter in log-log-scale.



(a) geschätzter l1-Fehler

(b) l1-Fehler-Schranke für die Wahrscheinlichkeit $1-\delta = 0.67$

Abbildung 4.13: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 3×3-Gitter in log-log-scale.



(a) geschätzter l1-Fehler

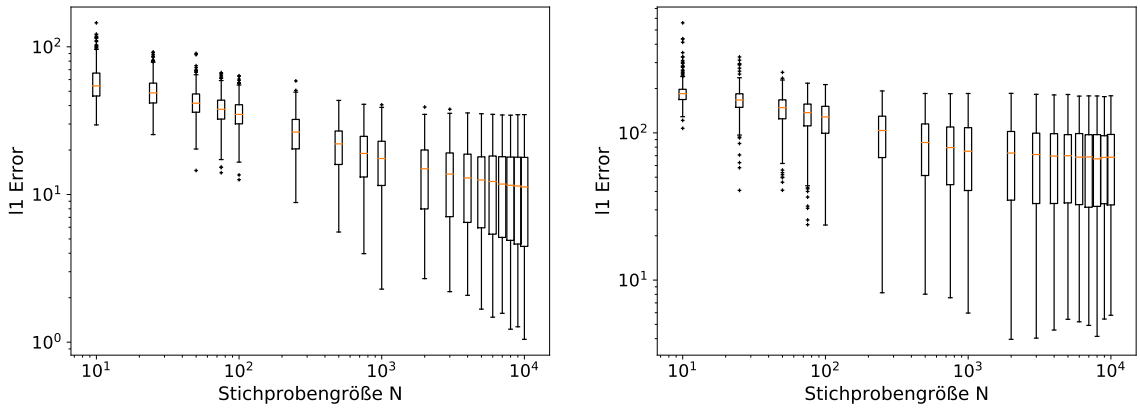
(b) l1-Fehler-Schranke für die Wahrscheinlichkeit $1-\delta = 0.67$

Abbildung 4.14: Vergleich zwischen dem l1-Fehler über 500 Parameter in Abhängigkeit von C_{min} bei einer Stichprobengröße von $N = 10000$ für die 4×4-Gitter in log-log-scale.

Auch hier lässt sich ein deutlicher Zusammenhang zwischen dem minimalen Eigenwert und dem l1-Fehler erkennen. Wenn die Boxplots bei $N = 10000$ für den MPLE betrachtet werden, dann lässt sich erkennen, dass die Antennen sehr weit nach unten reichen. An den Grafiken nach C_{min} ist erkennbar, dass dies die Parametervektoren sind, deren Hessematrix einen größeren minimalen Eigenwert haben.

4.3 Vergleich

Im Vergleich der MPL-Schätzung und der ML-Schätzung, ist auffällig, dass der MPLE für kleinere Stichproben einen deutlich größeren Fehler aufweist.



(a) MLE **(b)** MPLE

Abbildung 4.15: Vergleich 3×3-Gitter

Weiterhin lässt sich feststellen, dass der Median des MPLE in Abhängigkeit von N langsamer sinkt. Der Unterschied zwischen dem Fehler für die unterschiedlich großen Modelle ist für dem MPLE auch deutlich größer als beim MPLE. Während der l1-Fehler des MLE im Median für eine Stichprobengröße von 10000, für das 2×2-Gitter bei ca. 4, das 3×3-Gitter bei ca. 11 und das 4×4-Gitter bei ca. 21 liegt, liegt dieser Median beim MPLE in der gleichen Ordnung bei ungefähr 7, 68 und 329. Die höhere Effizienz des MLE ist deutlich erkennbar.
Weiterhin ist der MPLE für diese Experimente von der Laufzeit langsamer als der MLE, da die Zustandsräume dieser Ising-Modelle klein sind und die Partitionsfunktion für jeden Parametervektor nur ein Mal berechnet werden muss. Bei dem MPLE hängen die Partitionsfunktionen der Komponenten von der Stichprobe ab, sodass sie für jede Stichprobe neu berechnet werden müssen. Die Laufzeit des MPLE hängt weniger vom Zustandsraum und mehr von der Stichprobengröße ab. Dadurch, dass der Zustandsraum in unseren Experimenten klein ist, liefert der MPLE keinen Vorteil bezüglich der Laufzeit. Für größere Modelle ist dies nicht der Fall.

Kapitel 5

Fazit

In dieser Arbeit wurde der Beweis aus [4] aufgearbeitet und auf MRFs angewendet. Dadurch wurde eine PAC-Schranke und eine Schranke für die Stichprobenkomplexiät für den regularisierten MLE und den regularisierten konsistenten MCLE aufgestellt. Beide Schranken gelten für Exponentialfamilien in minimaler Darstellung. Der Großteil des Beweises bleibt, im Vergleich zu [4] unverändert, es werden jedoch einige Details genauer erklärt.
Die PAC-Schranken von Bradley und Guestrin lassen sich zum Ende des Beweises durch andere Überlegungen und durch die Wahl eines kleineren Regularisierungsparameters λ verbessern, diese Verbesserung kann jedoch nicht genutzt werden um eine bessere Stichprobenkomplexität zu berechnen. Deswegen sind sich die Stichprobenkomplexitäten sehr ähnlich.
Die Stichprobenkomplexität des MLE ist kleiner als die des MPLE, was an der größeren Wahrscheinlichkeit der PAC-Schranke und den unterschiedlichen Parametern C_{min} liegt, wobei C_{min} für den MLE als $\Lambda_{\min}(\nabla^2 \ell_L(\boldsymbol{\theta}^*))$ definiert ist und für den MCLE als $\min_{j\in\{1,\dots,m\}} \nabla^2 \ell_{CL}(\boldsymbol{\theta}^*)_{A_j}$ definiert ist. Beide C_{min} hängen von dem wahren Parameter $\boldsymbol{\theta}^*$ ab und sind somit unter realen Umständen nicht bekannt. Dies ist das größte Problem des Beweises und der Grund warum nur synthetische Experimente zum Prüfen der Schranke möglich sind. Ein weiteres Problem ist, dass der Regularisierungsparameter λ für die Gültigkeit der PAC-Schranke sehr begrenzt ist.
In den Experimenten wurden Ising-Modelle benutzt, um die Schranken mit dem Fehler des, aus den synthetischen Daten resultierenden, Schätzers zu vergleichen. Ising-Modelle haben den Vorteil, dass die minimale suffiziente Statistik bekannt ist.
Es ist aus den Experimenten erkennbar, dass eine deutlich kleinere Anzahl an Daten für die ausgewählten kleinen Modelle reicht, um einen hinreichend kleinen Fehler zu erzielen, als die Schranke vorgibt. Die Abhängigkeit des Fehlers von C_{min} konnte für den MLE und MPLE, genauso wie bei Bradley und Guestrin, auch in den synthetischen Experimenten erkannt werden.

Anhang A

Weitere Informationen

A.1 Weitere genutzte Sätze und Definitionen der Wahrscheinlichkeitstheorie

A.1.1 Theorem (Bonferroni-Ungleichung [12]). *Seinen E_i, mit $i \in 1, ..., n$ beliebige Ereignisse in einem Wahrscheinlichkeitsraum, dann gilt*

$$\mathbb{P}(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} \mathbb{P}(E_i) \; . \tag{A.1}$$

A.1.2 Theorem (Hoeffding Ungleichung [14]). *Seien $X_1, ..., X_N$ unabhängige Zufallsvariablen und $\overline{X} = \frac{1}{N}(X_1 + ... + X_N)$. Sei der Wert jeder Zufallsvariable X_i mit $i \in \{1, ..., N\}$ durch das Intervall $[a_i, b_i]$ begrenzt, dann gilt*

$$\mathbb{P}(|\overline{X} - \mathbb{E}[\overline{X}]| \geq t) \leq 2 \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \tag{A.2}$$

A.1.3 Definition (Erwartungstreue[12]). Ein Schätzer $\hat{\boldsymbol{\theta}}_N$ von $\boldsymbol{\theta}$ über N Realisierungen einer Zufallsvariable $\boldsymbol{X}$ wird erwartungstreu genannt, wenn $\mathbb{E}(\hat{\boldsymbol{\theta}}_N) = \boldsymbol{\theta}^*$. Wobei $\boldsymbol{\theta}^*$ den Parameter der Verteilung bezeichnet, aus der die N Realisierungen erzeugt wurden.

A.1.4 Definition (Konsistenz [12]). Ein Schätzer $\hat{\boldsymbol{\theta}}_N$ von $\boldsymbol{\theta}^*$ über N Beobachtungen einer Zufallsvariable $\boldsymbol{X}$ wird konsistent genannt, wenn bei wachsendem N $\hat{\boldsymbol{\theta}}_N$ gegen $\boldsymbol{\theta}^*$ konvergiert.

A.2 Mathematische Grundlagen

A.2.1 Eigenwerte

A.2.1 Theorem (Satz von Courant-Fischer [15]). *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit den Eigenwerten $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ und sei $\mathcal{S}^k$ mit $k \in \{1, 2, \ldots, n\}$ die Menge aller k-dimensionaler Untervektorraum von $\mathbb{R}^{n\times n}$, dann gilt für $k \in \{1, 2, \ldots, n\}$*

$$\lambda_k = \min_{S\in\mathcal{S}^k} \max_{x:0\neq x\in S} \frac{\langle x, Ax\rangle}{\langle x, x\rangle} = \max_{S:S\in\mathcal{S}^{n-k+1}} \min_{x:0\neq x\in S} \frac{\langle x, Ax\rangle}{\langle x, x\rangle} \tag{A.3}$$

Dies bedeutet für $k = 1$:

$$\lambda_1 = \min_{x\neq 0} \frac{x^T Ax}{\langle x, x\rangle}$$

Und für $k = n$:

$$\lambda_n = \max_{x\neq 0} \frac{x^T Ax}{\langle x, x\rangle}$$

Da $\langle x, x\rangle = \|x\|_2^2 = \|x\|_2\|x\|_2$ gilt, können im Zähler x und x^T durch $\|x\|_2$ geteilt werden, weswegen $\max_{x\neq 0} \frac{x^T Ax}{\langle x,x\rangle} = \max_{\|x\|_2=1} x^T Ax$ und $\min_{x\neq 0} \frac{x^T Ax}{\langle x,x\rangle} = \min_{\|x\|_2=1} x^T Ax$ geschrieben werden kann.

A.2.2 Normen

A.2.2 Theorem (Hölder-Ungleichung [15]). *Für alle messbaren Funktionen f, g und für alle $p, q \in \mathbb{R}$ mit $\frac{1}{p} + \frac{1}{q} = 1$ gilt*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \,, \tag{A.4}$$

mit einer Gleichheit nur wenn $|f|^p$ und $|g|^q$ linear von einander abhängig sind.

A.2.3 Definition (Frobeniusnorm [24]). Sei $A \in \mathbb{R}^{n\times m}$ mit $n, m \in \mathbb{N}$ eine Matrix, dann ist die Frobeniusnorm $\|A\|_F$ der Matrix A definiert als

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2} \tag{A.5}$$

A.3 Ableitungen

A.3.1 Log-Likelihood

Erste Ableitung

Jeder Eintrag im Gradient an der Stelle $t \in 1, ..., d$ ist $\nabla[\hat{\ell}_L(\boldsymbol{\theta})]_t = \frac{\partial}{\partial \boldsymbol{\theta}_t}\hat{\ell}_L(\boldsymbol{\theta})$:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_t}\hat{\ell}_L(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N}\Big(-\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}^i)\rangle + A(\boldsymbol{\theta})\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(-\phi(\boldsymbol{x}^i)_t + \frac{\partial}{\partial \boldsymbol{\theta}_i}A(\boldsymbol{\theta})\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(-\phi(\boldsymbol{x}^i)_t + \frac{\partial}{\partial \boldsymbol{\theta}_t}\log(\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(-\phi(\boldsymbol{x}^i)_t + \frac{1}{(\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle)}\left(\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle)\phi(\boldsymbol{x})_t\right)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(-\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\boldsymbol{\theta}}\left[\phi(\boldsymbol{X}^j)_t\right]\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N} -\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\boldsymbol{\theta}}\left[\phi(\boldsymbol{X})_t\right]
\end{aligned}
$$

Zweite Ableitung

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_t \boldsymbol{\theta}_k}\hat{\ell}_L(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_k}\frac{1}{N}\sum_{i=1}^{N}\Big(-\boldsymbol{x}_t^i + \mathbb{E}_{\boldsymbol{\theta}}\left[\phi(\boldsymbol{X})_t\right]\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial}{\partial \boldsymbol{\theta}_k}\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\theta}))\phi(\boldsymbol{x})_t\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\theta}))\phi(\boldsymbol{x})_t(\phi(\boldsymbol{x})_k - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k])\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{\boldsymbol{x}}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\theta}))(\phi(\boldsymbol{x})_t\phi(\boldsymbol{x})_k) - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}[\phi(\boldsymbol{X})_k])\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_k] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]) \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_k] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]
\end{aligned}
$$

Dritte Ableitung

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_s}\nabla^2\hat{\ell}_L(\boldsymbol{\theta})_{t,k} =& \frac{\partial}{\partial \boldsymbol{\theta}_s}\left(\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_k] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X}^j)_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\right)\\
=& \frac{\partial}{\partial \boldsymbol{\theta}_s}\left(\sum_{\boldsymbol{x}}\exp(\langle\boldsymbol{\theta},\phi(x)\rangle - A(\boldsymbol{\theta}))\phi(\boldsymbol{x})_r\phi(\boldsymbol{x})_k - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\right)\\
=& \sum_{\boldsymbol{x}}(\exp(\langle\boldsymbol{\theta},\phi(x)\rangle - A(\boldsymbol{\theta}))(\phi(\boldsymbol{x})_s - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s])\phi(\boldsymbol{x})_t\phi(\boldsymbol{x})_k)\\
&- \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,(\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k\phi(\boldsymbol{X})_s] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\,\mathbb{E}[\phi(\boldsymbol{X})_s])\\
&- \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\,(\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_s] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s])\\
=& \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s\phi(\boldsymbol{X})_k\phi(\boldsymbol{X})_t] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s]\\
&+ 2\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\\
&- \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k\phi(\boldsymbol{X})_s] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_k]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_t\phi(\boldsymbol{X})_s] \qquad \text{(A.6)}
\end{aligned}$$

Für $\frac{\partial}{\partial \boldsymbol{\theta}_s}\nabla^2\hat{\ell}_L(\boldsymbol{\theta})$ gilt nun:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_s}\nabla^2\hat{\ell}_L(\boldsymbol{\theta}) =& \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s\phi(\boldsymbol{X})^{\otimes}] - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})^{\otimes}]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s] + 2\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})_s]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})]^{\otimes}\\
&- \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]^T - \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})_s]\,\mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{x})]^T \qquad \text{(A.7)}
\end{aligned}$$

A.3.2 Log-Composite-Likelihood

Erste Ableitung

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_t}\hat{\ell}_{CL}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\frac{\partial}{\partial \boldsymbol{\theta}_t}\left(-\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x}^i)\rangle) + \log\left(\sum_{\boldsymbol{x'}_{A_j}}\exp(\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\rangle)\right)\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\left(-\phi_{B_j}(\boldsymbol{x}^i)_t + \frac{\frac{\partial}{\partial \boldsymbol{\theta}_t}\left(\sum_{x'_{A_j}}\exp(\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\rangle)\right)}{\left(\sum_{x'_{A_j}}\exp(\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\rangle)\right)}\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\left(-\phi(\boldsymbol{x}^i)_t + \frac{\sum_{x'_{A_j}}\exp(\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\rangle)\phi(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_t}{\left(\sum_{x'_{A_j}}\exp(\langle\boldsymbol{\theta}_{B_j},\phi_{B_j}(\boldsymbol{x'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})\rangle)\right)}\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j:t\in B_j}\left(-\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X'}_{A_j}|\boldsymbol{x}^i_{\setminus A_j})}[\phi_{B_j}(\boldsymbol{X'}_{A_j},\boldsymbol{x}^i_{\setminus A_j})_t]\right)
\end{aligned}$$

Zweite Ableitung

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}_t \boldsymbol{\theta}_k} \hat{\ell}_{CL}(\boldsymbol{\theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}_k} \sum_{j:\boldsymbol{\theta}_t \in \boldsymbol{\theta}_{B_j}} \left(-\phi(\boldsymbol{x}^i)_t + \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \left(\frac{\partial}{\partial \boldsymbol{\theta}_k} - \phi(\boldsymbol{x}^i)_t + \frac{\partial}{\partial \boldsymbol{\theta}_k} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \frac{\partial}{\partial \boldsymbol{\theta}_k} \frac{\sum_{\boldsymbol{x}'_{A_j}} \left(\exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \phi(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t \right)}{\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right)}
\end{aligned}
$$

An dieser Stelle nutzen wir die Produktregel und führen zum verkürzen der Formeln die Notation $\phi^{(i)}(\boldsymbol{x}_{A_j}) = \phi_{B_j}(\boldsymbol{x}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})$ ein.

$$
\begin{aligned}
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \left(\frac{\sum_{\boldsymbol{x}'_{A_j}} \left(\exp(\langle \boldsymbol{\theta}_{B_j}, \phi^{(i)}(\boldsymbol{x}'_{A_j}) \rangle) \phi^{(i)}(\boldsymbol{x}'_{A_j})_k \phi^{(i)}(\boldsymbol{x}'_{A_j})_t \right)}{\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right)} \right. \\
&\left. - \frac{\sum_{\boldsymbol{x}'_{A_j}} \left(\exp(\langle \boldsymbol{\theta}_{B_j}, \phi^{(i)}(\boldsymbol{x}'_{A_j}) \rangle) \phi^{(i)}(\boldsymbol{x}'_{A_j})_t \right)}{\left(\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right)^2} \sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi^{(i)}(\boldsymbol{x}'_{A_j}) \rangle) \phi^{(i)}(\boldsymbol{x}'_{A_j})_k \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \left(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k \phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right. \\
&\left. - \frac{\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi^{(i)}(\boldsymbol{x}'_{A_j}) \rangle) \phi^{(i)}(\boldsymbol{x}'_{A_j})_t}{\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right)} \frac{\sum_{\boldsymbol{x}'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi^{(i)}(\boldsymbol{x}'_{A_j}) \rangle) \phi^{(i)}(\boldsymbol{x}'_{A_j})_k}{\left(\sum_{x'_{A_j}} \exp(\langle \boldsymbol{\theta}_{B_j}, \phi_{B_j}(\boldsymbol{x}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j}) \rangle) \right)} \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \left(\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k \phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right. \\
&\left. - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k] \, \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}'_{A_j} | \boldsymbol{x}^i_{\setminus A_j})} [\phi(\boldsymbol{X}'_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right)
\end{aligned} \tag{A.8}
$$

Dritte Ableitung

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}_s} \nabla^2 \hat{\ell}_{CL}(\boldsymbol{\theta})_{k,t} = \\
&\frac{\partial}{\partial \boldsymbol{\theta}_s} \frac{1}{N} \sum_{i=1}^{N} \sum_{j:t,k \in B_j} \left(\mathbb{E}^i_j [\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k \phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] - \mathbb{E}^i_j [\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k] \, \mathbb{E}^i_j [\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t] \right)
\end{aligned}
$$

Da die Ableitung von $\mathbb{E}_j^i[\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k]$ durch vorherige Berechnungen schon bekannt ist (A.8), können wie die Ableitung für $\mathbb{E}_j^i[\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k \phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t]$ daraus herleiten und für den Term $\mathbb{E}_j^i[\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_k]\,\mathbb{E}_j^i[\phi(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})_t]$ kann einfach die Produktregel benutzt werden. Weiterhin werden wir von hier aus wieder die abgekürzte Notation $\phi^{(i)}(\boldsymbol{X}_{A_j}) = \phi_{B_j}(\boldsymbol{X}_{A_j}, \boldsymbol{x}^i_{\setminus A_j})$ nutzen.

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t,k\in B_j}\Big(\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s] - \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\\
&-(\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k](\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_s] - \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s])\\
&+\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t](\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s] - \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]))\Big)\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t,k\in B_j}\Big(\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s] - \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\\
&-\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_s] + \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\\
&-\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s] + \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\Big)\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{j:t,k\in B_j}\Big(\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s] - \mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\\
&-\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k\phi^i(\boldsymbol{X}_{A_j})_s] + 2\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_k]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_s]\\
&-\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t]\,\mathbb{E}_j^i[\phi^i(\boldsymbol{X}_{A_j})_t\phi^i(\boldsymbol{X}_{A_j})_s])
\end{aligned}
$$

Abbildungsverzeichnis

Literaturverzeichnis

[1] Christoph Beierle und Gabriele Kern-Isberner. *Methoden wissensbasierter Systeme: Grundlagen, Algorithmen, Anwendungen*. Computational Intelligence. Springer Vieweg, 5 edition, 2014. ISBN 978-3-8348-1896-6. doi: 10.1007/978-3-8348-2300-7.

[2] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):pp. 179–195, 1975. ISSN 00390526. URL `http://www.jstor.org/stable/2987782`.

[3] Stephen Boyd und Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

[4] Joseph Bradley und Carlos Guestrin. Sample complexity of composite likelihood. In Neil D. Lawrence und Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 136–160, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL `http://proceedings.mlr.press/v22/bradley12.html`.

[5] Lawrence D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986. ISSN 07492170. URL `http://www.jstor.org/stable/4355554`.

[6] Andrei Bulatov und Martin Grohe. The complexity of partition functions. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, und Donald Sannella, editors, *Automata, Languages and Programming*, pages 294–306, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27836-8.

[7] Marijtje Duijn, Krista J Gile, und Mark S Handcock. A framework for the comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Social networks*, 31:52–62, 01 2009. doi: 10.1016/j.socnet.2008.10.003.

[8] J.J. Duistermaat und J.A.C. Kolk. *Distributions: Theory and Applications*. Cornerstones. Birkhäuser Boston, 2010. ISBN 978-0-8176-4675-2. doi: 10.1007/978-0-8176-4675-2.

[9] B. Gidas. Consistency of maximum likelihood and pseudo-likelihood estimators for gibbs distributions. In Wendell Fleming und Pierre-Louis Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 129–145, New York, NY, 1988. Springer New York. ISBN 978-1-4613-8762-6.

[10] Linus Hamilton, Frederic Koehler, und Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. *CoRR*, abs/1705.11107, 2017. URL http://arxiv.org/abs/1705.11107.

[11] Steve Hanneke. The optimal sample complexity of PAC learning. *CoRR*, abs/1507.00473, 2015. URL http://arxiv.org/abs/1507.00473.

[12] Georgii Hans-Otto. *Stochastik, Einführung in die Wahrscheinlichkeitstheorie und Statistik*. De Gruyter, 2015. ISBN 978-3-11-035969-5. URL https://www.degruyter.com/view/product/428666.

[13] Trevor Hastie, Robert Tibshirani, und Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer, 2009. ISBN 9780387848570. URL http://www.worldcat.org/oclc/300478243.

[14] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

[15] Roger A Horn und Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012. ISBN 978-0-521-83940-2.

[16] Anil K. Jain und Sateesha G. Nadabar. *Markov Random Field Applications in Image Analysis*, pages 39–50. Springer US, Boston, MA, 1992. ISBN 978-1-4615-3388-7. doi: 10.1007/978-1-4615-3388-7_5. URL https://doi.org/10.1007/978-1-4615-3388-7_5.

[17] Adam R. Klivans und Raghu Meka. Learning graphical models using multiplicative weights. *CoRR*, abs/1706.06274, 2017. URL http://arxiv.org/abs/1706.06274.

[18] Percy Liang und Michael I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 584–591, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390230. URL http://doi.acm.org/10.1145/1390156.1390230.

[19] Thomas Liebig, Nico Piatkowski, Christian Bockermann, und Katharina Morik. Dynamic route planning with real-time traffic predictions. *Information Systems*, 64: 258 – 265, 2017. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2016.01.007. URL http://www.sciencedirect.com/science/article/pii/S0306437916000181.

[20] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988. URL http://dx.doi.org/10.1090/conm/080/999014.

[21] Christopher D. Manning und Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

[22] Benjamin M. Marlin und Nando de Freitas. Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. *CoRR*, abs/1202.3746, 2012. URL http://arxiv.org/abs/1202.3746.

[23] Yariv Dror Mizrahi, Misha Denil, und Nando de Freitas. Distributed parameter estimation in probabilistic graphical models. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1700–1708, 2014. URL http://papers.nips.cc/paper/5317-distributed-parameter-estimation-in-probabilistic-graphical-models.

[24] K. B. Petersen und M. S. Pedersen. The matrix cookbook, nov 2012. URL http://www2.imm.dtu.dk/pubdb/p.php?3274. Version 20121115.

[25] Nico Piatkowski. *Exponential families on resource-constrained systems*. PhD thesis, Technical University of Dortmund, Germany, 2018. URL http://hdl.handle.net/2003/36877.

[26] Pradeep Ravikumar, Martin J. Wainwright, und John D. Lafferty. High-dimensional ising model selection using l 1 -regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 06 2010. doi: 10.1214/09-AOS691. URL https://doi.org/10.1214/09-AOS691.

[27] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

[28] Martin J. Wainwright und Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

[29] Ingo Wegener. *Theoretische Informatik: Eine algorithmenorientierte Einführung*. XLeitfäden der Informatik. Vieweg+Teubner Verlag, 1993. ISBN 978-3-322-94004-9. doi: 10.1007/978-3-322-94004-9.

Eidesstattliche Versicherung
(Affidavit)

Scheftelowitsch, Olga
Name, Vorname
(Last name, first name)

183885
Matrikelnr.
(Enrollment number)

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/Masterarbeit* mit dem folgenden Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present Bachelor's/Master's* thesis with the following title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution.

Titel der Bachelor-/Masterarbeit*:
(Title of the Bachelor's/ Master's* thesis):

Statistische Effizienz der Parameterschätzung für Exponentialfamilien

*Nichtzutreffendes bitte streichen
(Please choose the appropriate)

Dortmund, 09.09.2019
Ort, Datum
(Place, date)

Olga Scheftelowitsch
Unterschrift
(Signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to €50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, section 63, subsection 5 of the North Rhine-Westphalia Higher Education Act (*Hochschulgesetz*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:**

Dortmund, 09.09.2019
Ort, Datum
(Place, date)

Olga Scheftelowitsch
Unterschrift
(Signature)

**Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.