Evaluierung numerischer Optimierungsverfahren für die robuste Spracherkennung nach dem REMOS-Konzept

Masterarbeit

von

Roland Maas

Betreuer

PD Dr. Martin Gugat

Prof. Dr. Walter Kellermann

Dipl.-Ing. Armin Sehr

17. September 2009

Friedrich-Alexander-Universität Erlangen-Nürnberg

Department Mathematik

Lehrstuhl für Angewandte Mathematik II

Lehrstuhl für Multimediakommunikation und Signalverarbeitung

Lehrstuhl für Multimediakommunikation
und Signalverarbeitung
Prof. Dr.-Ing. W. Kellermann

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Masterarbeit

für

Herrn cand. rer. nat. Roland Maas

Evaluierung numerischer Optimierungsverfahren für die robuste Spracherkennung nach dem REMOS-Konzept

Robuste Spracherkennung im Freisprechmodus ist für viele Anwendungen äußerst wünschenswert. Aufgrund der Mehrwegeausbreitung in realen akustischen Umgebungen nimmt das Mikrofon bei größeren Abständen zum Sprecher nicht nur das gewünschte Signal, sondern auch dessen Nachhall auf. Dadurch wird die Fehlerrate heutiger Spracherkennungssysteme erheblich erhöht. Da der Nachhall einen dispersiven Effekt auf die zur Erkennung verwendeten Merkmalfolgen hat, erreichen konventionelle Signalverbesserungs- und Modelladaptions-Algorithmen nur eingeschränkte Verbesserungen in halligen Umgebungen.

Das neuartige REMOS-Konzept (REverberation MOdeling for Speech recognition) setzt erstmals eine Kombination eines Hidden Markov Models (HMM) und eines Nachhallmodells ein und erzielt auch in stark verhallten Umgebungen vielversprechende Ergebnisse. Das HMM modelliert die unverhallte Sprache, während das Nachhallmodell den Effekt des Nachhalls direkt im Merkmalbereich beschreibt. Für die Spracherkennung wird eine erweiterte Version des Viterbi-Algorithmus eingesetzt, bei der in jedem Iterationsschritt eine innere Optimierung die wahrscheinlichsten Beiträge des HMMs und des Nachhallmodells zum aktuellen verhallten Merkmalvektor ermittelt. Bisher wurde das Verfahren nur für „mel-spectral"-Merkmale implementiert und ausgiebig verifiziert.

Bei der Erweiterung des REMOS-Konzepts auf leistungsfähigere Sprachmerkmale, wie logarithmische „mel-spectral"-Merkmale und „Mel-Frequency Cepstral Coefficients" (MFCCs), kann die innere Optimierung nur durch numerische Verfahren gelöst werden. Die innere Optimierung kann dabei als ein Problem mit einer quadratischen Zielfunktion und nichtlinearen Nebenbedingungen oder als ein Problem mit einer nichtlinearen Zielfunktion und linearen Nebenbedingungen formuliert werden. Im Rahmen dieser Arbeit sollen verschiedene Formulierungen und verschiedene Verfahren zur numerischen Lösung des inneren Optimierungsproblems evaluiert werden. Zu der Evaluierung der Verfahren gehört insbesondere die Darstellung einer Konvergenztheorie der verwendeten Algorithmen. Um die dazugehörigen Voraussetzungen zu prüfen, ist eine Untersuchung der Struktur der Zielfunktion erforderlich, um damit zum Beispiel die Existenz und die Stabilität einer Lösung zu gewährleisten. Da bei dem behandelten Problem die Lösung nicht eindeutig bestimmt ist, sind Aussagen über die Anzahlen der auftretenden kritischen Punkte (zum Beispiel obere und untere Schranken) interessant. Durch die Implementierung der numerischen Methoden soll der vorhandene C-Code für einen HMM Toolkit (HTK)-basierten Erkenner mit möglichst effizienten Algorithmen zur Lösung des inneren Optimierungsproblems erweitert werden.

Ausgabe: 01.04.2009
Abgabe: 30.09.2009

Betreuer: Priv.-Doz. Dr. Martin Gugat (AM2)
Prof. Dr.-Ing. Walter Kellermann (LMS)
Dipl.-Ing. Armin Sehr (LMS)

(Priv.-Doz. Dr. M. Gugat) (Prof. Dr.-Ing. W. Kellermann)

Inhaltsverzeichnis

Kapitel 1

Einführung

Die automatische Spracherkennung ist ein Forschungsgebiet der Informatik und Ingenieurwissenschaften, welches in den 1960er Jahren entstand. Nach anfänglichen Schwierigkeiten schritt die Entwicklung in den 1980er Jahren dank neuer technischer Möglichkeiten und komplexerer statistischer Modelle und Algorithmen weiter voran (vgl. [6]). Heutzutage haben Spracherkennungssysteme auf verschiedene Weise Einzug in unseren Alltag erhalten: beispielsweise im Rahmen von Diktiersystemen, Telefonhotlines oder Mobiltelefonsteuerungen. In vielen dieser Anwendungsszenarien werden bereits sehr gute Spracherkennungsraten erzielt, meist muss sich dazu jedoch das Mikrofon nahe am Mund des Sprechers befinden. Es gibt allerdings auch zahlreiche Anwendungsfelder, in denen das Tragen eines Mikrofons oder Headsets entweder nicht möglich oder mit erheblichen Komforteinbußen für den Benutzer verbunden ist. Beispielsweise bei der automatischen Erstellung von Sitzungsprotokollen ist es durchaus aufwändig, jeden Konferenzteilnehmer mit einem eigenen Mikrofon auszustatten. Auch die Sprachsteuerung medizintechnischer Geräte während einer Operation oder die Steuerung eines Fernsehgerätes mittels Sprachbefehlen wäre deutlich praktikabler, installierte man die verwendeten Mikrofone etwa am Gerät selbst und nicht am jeweiligen Sprecher.

Mit steigendem Abstand des Mikrofons zum Sprecher tritt allerdings ein Effekt immer weiter in den Vordergrund, welcher heutigen Spracherkennungssystemen große Probleme bereitet: Ab einer Distanz von etwa einem Meter gelangt nicht mehr nur das direkte Sprachsignal vom Mund des Sprechers zum Mikrofon, sondern neben Hintergrundgeräuschen insbesondere auch ein nicht mehr zu vernachlässigender Anteil an Nachhall der eigenen Sprache, welcher durch z. B. Reflexion des Schalls an Wänden und Möbeln entsteht (vgl. Abbildung 1.1).

Um die dabei auftretende Problematik besser nachvollziehen zu können, betrachten wir zunächst die Funktionsweise von Spracherkennungssystemen und lehnen uns dabei in weiten Teilen an [9] an.

Ein erster wichtiger Verarbeitungsschritt in einem Spracherkenner ist die Extraktion bestimmter Merkmale aus dem Sprachsignal.

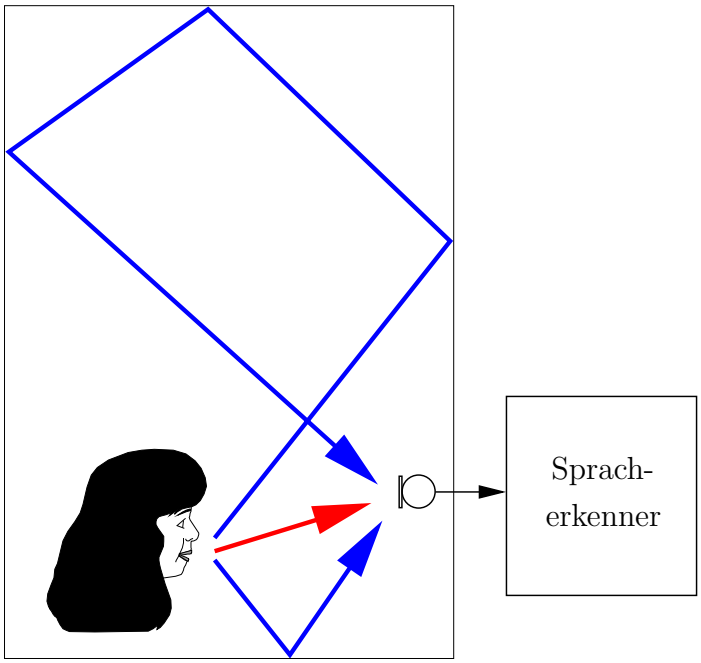


Abbildung 1.1: Schematische Darstellung von Nachhall.

In Abbildung 1.2 ist zu sehen, welche Etappen nötig sind, um aus einem ursprünglichen Sprachsignal die so genannten MFCC-Merkmalvektoren (Mel Frequency Cepstral Coefficients) zu berechnen.

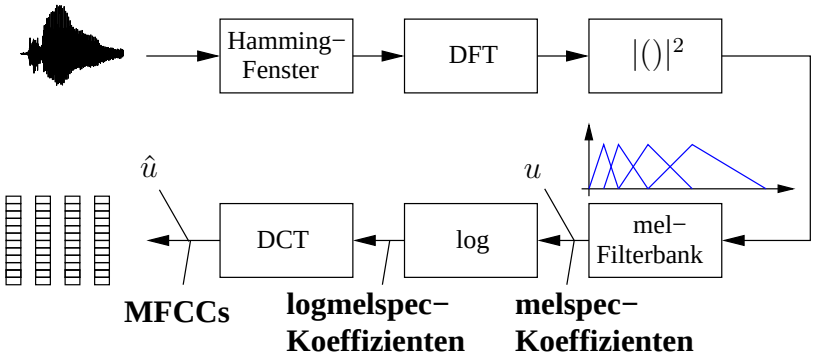


Abbildung 1.2: Merkmalextraktion nach [9].

In der Trainingsphase des Spracherkenners wird nun für jedes Wort (oder jede Silbe oder jedes Phonem) ein so genanntes Hidden Markov Modell (HMM) erstellt, das beschreibt, mit welcher Wahrscheinlichkeit das jeweilige Trainingswort eine bestimmte Merkmalvektorfolge erzeugt. Durch Verbinden mehrerer HMMs zu einem HMM-Netzwerk lassen sich mithin ganze Wortfolgen (oder Folgen von Silben oder

Phonemen) modellieren.

Soll nun in einer praktischen Anwendung eine sprachliche Äußerung erkannt werden, so vergleicht das System die Merkmalvektoren des Sprachsignals mit den HMMs des Erkennerwortschatzes. Mit Hilfe des Viterbi-Algorithmus wird dabei die wahrscheinlichste Wortfolge zur gegebenen Sprachäußerung ermittelt.

Zur Illustration ist in Abbildung 1.3 die logmelspec-Merkmalvektorfolge der Äußerung „one, one, eight“ dargestellt, welche mit einem Mikrofon nahe am Mund des Sprechers aufgezeichnet wurde.

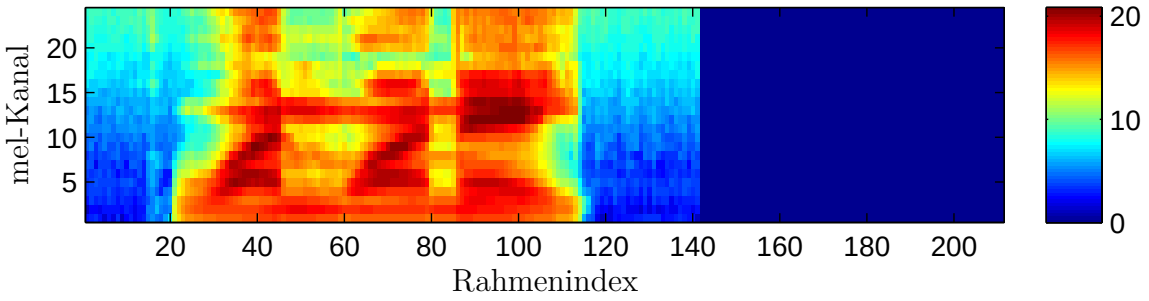


Abbildung 1.3: Logmelspec-Darstellung der (unverhallten) Äußerung „one, one, eight“.

Zum Vergleich zeigt Abbildung 1.4 die Merkmalvektorfolge derselben Äußerung, diesmal allerdings mit einem mehrere Meter entfernten Mikrofon aufgenommen. Der Nachhall ist deutlich zu sehen: Er äußert sich in einer zeitlichen Verschleifung der Sprachmerkmale. Die Wort-/Phonemgrenzen sind undeutlicher zu erkennen, da jeder Vektor nun nicht mehr nur Direktschall repräsentiert, sondern zusätzlich den Nachhall der vorangehenden Vektoren.

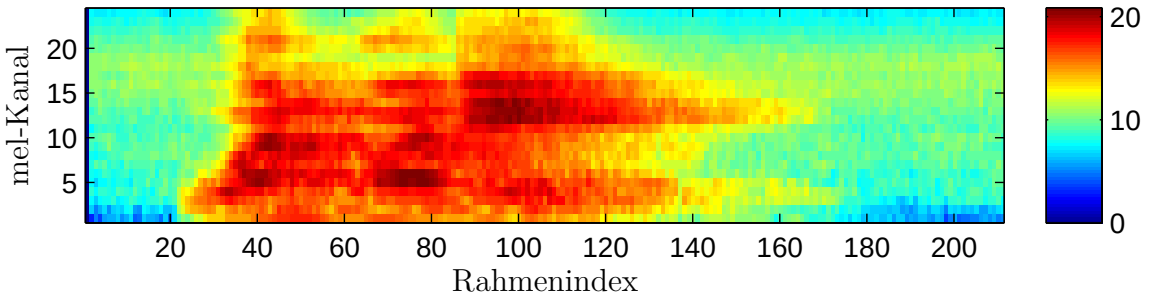


Abbildung 1.4: Logmelspec-Darstellung der (verhallten) Äußerung „one, one, eight“.

Wie man sich leicht vorstellen kann, steigt die Fehlerrate erheblich, wenn ein Spracherkenner, der mit klaren (unverhallten) Testdaten, wie in Abbildung 1.3, trainiert wurde, auf ein verhalltes Signal, wie in Abbildung 1.4, angewendet wird.

Ein möglicher Lösungsansatz wäre natürlich, den Spracherkenner von vorne herein mit verhallten Daten zu trainieren. Dieses Verfahren wird in der Praxis auch tatsächlich

mitunter angewendet, birgt aber den signifikanten Nachteil, dass der Erkenner beim Wechsel der Umgebung (etwa Raumwechsel) neu trainiert werden muss, da sich die Nachhallcharakteristik ändert.

Ein anderer Lösungsansatz, der diesem Problem begegnet, ist das so genannte REMOS-Konzept (REverberation MOdeling for Speech recognition) [9]. Dieses Verfahren erweitert herkömmliche HMM-basierte Spracherkennungssysteme um ein Nachhallmodell, das eine statistische Beschreibung der Raumimpulsantwort im Merkmalbereich darstellt. Dabei wird die verhallte Merkmalvektorsequenz als Faltung der klaren Merkmalvektorsequenz mit der melspec-Repräsentation der Raumimpulsantwort angesehen. Im Rahmen des Spracherkennungssystems wird dann versucht, diese Faltung wieder rückgängig zu machen, um eine Schätzung für die klare Merkmalvektorfolge zu erhalten, welche dann mit den Modellen des Erkennerwortschatzes verglichen wird. Das ermöglicht es, den Spracherkenner einmalig mit unverhallten Testdaten zu trainieren und erfordert dabei lediglich die Kenntnis des Nachhallmodells. Soll der Spracherkenner in einer anderen Umgebung eingesetzt werden, so ist es nicht nötig, selben neu zu trainieren, sondern es genügt, das Nachhallmodell neu zu schätzen.

Die eigentliche Herausforderung bei der Umsetzung des REMOS-Konzeptes ist die Durchführung der Entfaltung, welche wir im Folgenden genauer betrachten:
Es sei dazu $u(k) \in \mathbb{R}^N$ der k. melspec-Merkmalvektor des verhallten Sprachsignals, $h(m,k) \in \mathbb{R}^N$, $m = 0, ..., M-1$, die die Raumimpulsantwort beschreibende Vektorfolge und $s(k) \in \mathbb{R}^N$ der k. melspec-Merkmalvektor des klaren Sprachsignals. Im Rahmen des REMOS-Konzeptes nimmt man an, dass sich die verhallte Merkmalvektorfolge $u(k)$ näherungsweise durch folgende Faltung beschreiben lässt:

$$u(k) = \sum_{m=0}^{M-1} h(m,k) \odot s(k-m),$$

wobei $\odot$ für die elementweise Vektormultiplikation steht.

In herkömmlichen HMM-basierten Spracherkennungsalgorithmen wird jeder Merkmalvektor $u(k)$ der aufgezeichneten Äußerung mit den HMMs eines jeden Wortes des Erkennerwortschatzes verglichen. Beim REMOS-Konzept vergleicht man, vereinfacht gesagt, das entfaltete Signal $s(k)$ mit den HMMs des Erkenners. Um die Entfaltung dabei so effizient wie möglich durchführen zu können, folgen nun noch einige vereinfachende Annahmen, die sich in der Praxis kaum negativ auf die Erkennungsraten auswirken, aber die Geschwindigkeit des REMOS-Konzeptes deutlich steigern.

Wir zerlegen $u(k)$ wie folgt:

$$\begin{aligned} u(k) &= h(0,k) \odot s(k) + \sum_{m=1}^{M-1} h(m,k) \odot s(k-m) \\ &\approx \underbrace{h(0,k) \odot s(k)}_{=:v(k)} + a(k) \cdot \underbrace{\sum_{m=1}^{M-1} \mu_h(m) \odot s(k-m)}_{:=r(k)}, \end{aligned}$$

wobei $\mu_h(m) \in \mathbb{R}^N$ für die Mittelwerte des unabhängig identisch verteilten Zufallsprozesses $h(m,k)$ steht (vgl. [7]). Wir bezeichnen $v(k) \in \mathbb{R}^N$ als Schätzung des Direktschalls, $r(k) \in \mathbb{R}^N$ als Schätzung des Nachhalls und $a(k) \in \mathbb{R}^N$ als Korrekturfaktor. Die Vektoren $s(k-m)$, für $m = 1, ..., M-1$, wurden dabei bereits in vorhergehenden Schritten geschätzt. Die einzig unbekannten Größen sind also $h(0,k)$ $s(k)$ (resp. $v(k)$) sowie $a(k)$.

Man geht davon aus, dass es sich bei $s(k)$ und $h(0,k)$ um statistisch unabhängige log-normalverteilte Zufallsprozesse handelt. Unter Kenntnis des Vektors $v(k)$ schätzt man die Vektoren $s(k)$ und $h(0,k)$ durch Maximierung der Verbundwahrscheinlichkeitsdichte von $s(k)$ und $h(0,k)$ unter der Nebenbedingung

$$v(k) = h(0,k) \odot s(k).$$

Dieses Optimierungsproblem lässt sich analytisch explizit lösen und soll uns daher nicht weiter beschäftigen.

Wesentlich anspruchsvoller ist demgegenüber die Bestimmung von $v(k)$, auf die wir uns nun konzentrieren werden. Zur besseren Lesbarkeit schreiben wir dabei nur noch die Vektorbuchstaben ohne Argument k.

Zunächst sind die melspec-Merkmalvektoren durch komponentenweises Ziehen des Logarithmus in den logmelspec-Bereich zu überführen. Durch eine zusätzliche Multiplikation mit einer orthogonalen Matrix $S^T \in \mathbb{R}^{N \times N}$, welche eine diskrete Kosinustransformation (DCT) beschreibt, ergeben sich die Merkmalvektoren im MFCC-Bereich:

$$\hat{u} := S^T \log(u) = S^T \log(v + a \cdot r).$$

Mit

$$\hat{v} := S^T \log(v),\ \hat{r} := S^T \log(r) \text{ und } \hat{a} := \log(a)$$

erhält man:

$$e^{S\hat{u}} = e^{S\hat{v}} + e^{S\hat{a}+S\hat{r}}, \tag{1.1}$$

wobei die Anwendung der Exponentialfunktion wieder komponentenweise zu verstehen ist.

In der Praxis nimmt man $\hat{v}$ sowie $\hat{a}$ als statistisch unabhängige gemeinsam normalverteilte Zufallsprozesse mit folgender Verbundichte an:

$$f_{\hat{v}\hat{a}}(\hat{v},\hat{a}) = K \cdot \exp\left(-\frac{1}{2} \cdot (\hat{v}-\mu_{\hat{v}})^T C_{\hat{v}\hat{v}}^{-1}(\hat{v}-\mu_{\hat{v}}) - \frac{1}{2} \cdot (\hat{a}-\mu_{\hat{a}})^T C_{\hat{a}\hat{a}}^{-1}(\hat{a}-\mu_{\hat{a}})\right).$$

$C_{\hat{v},\hat{v}} \in \mathbb{R}^{N \times N}$ bzw. $C_{\hat{a},\hat{a}} \in \mathbb{R}^{N \times N}$ stehen dabei für die diagonalen Kovarianzmatrizen und $\mu_{\hat{v}} \in \mathbb{R}^N$ bzw. $\mu_{\hat{a}} \in \mathbb{R}^N$ für die Vektoren der Mittelwerte von $\hat{v}$ bzw. $\hat{a}$. Bei K handelt es sich um eine positive Normierungskonstante.

Die Schätzung von $\hat{v}$ und $\hat{a}$ erfolgt durch die Maximierung der Verbundwahrscheinlichkeitsdichte unter der Nebenbedingung (1.1):

$$\begin{aligned} &\max_{\hat{v}\hat{a} \in \mathbb{R}^N} f_{\hat{v},\hat{a}}(\hat{v},\hat{a}) \\ &\text{s. t. } e^{S\hat{u}} = e^{S\hat{v}} + e^{S\hat{a}+S\hat{r}}, \end{aligned} \tag{1.2}$$

dabei steht „s. t.“ für „subject to“ und bedeutet „unter der Nebenbedingung“.

Um eine etwas anschaulichere Version dieses Optimierungsproblems zu erhalten, führen wir noch eine Koordinatentransformation durch:

$$\begin{aligned} x &:= S\hat{v} - S\hat{u}, \\ y &:= S\hat{a} + S\hat{r} - S\hat{u}, \\ c &:= S\mu_{\hat{v}} - S\hat{u}, \\ d &:= S\mu_{\hat{a}} + S\hat{r} - S\hat{u}, \\ A &:= SC_{\hat{v}\hat{v}}^{-1}S^T, \\ B &:= SC_{\hat{a}\hat{a}}^{-1}S^T. \end{aligned}$$

Man beachte, dass es sich bei A und B per Definition um symmetrisch positiv definite Matrizen handelt, da die diagonalen Kovarianzmatrizen $C_{\hat{v},\hat{v}}$ und $C_{\hat{a},\hat{a}}$ ausschließlich strikt positive Einträge besitzen.

Statt nun die Funktion $f_{\hat{v}\hat{a}}$ zu maximieren, minimieren wir $|\log(1/K \cdot f_{\hat{v}\hat{a}})|$ und erhalten schließlich folgendes zu (1.2) äquivalentes Optimierungsproblem:

$$\textbf{(P)} \quad \begin{cases} \min\limits_{x,y\in\mathbb{R}^N} & \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d) \\ & \text{s. t. } e^x + e^y = 1. \end{cases}$$

Ziel dieser Masterarbeit soll es sein, numerische Optimierungsverfahren zur Lösung von (P) zu untersuchen. Entscheidend für die Praxistauglichkeit ist dabei insbesondere die Laufzeit der Verfahren. In einer reellen Anwendung ist das Problem (P) mehrere Male pro Sekunde zu lösen (abhängig von der Wortschatzgröße des Spracherkenners), da Endverbraucher meist sehr reaktive Systeme erwarten und in diesen Fällen Spracherkenner Äußerungen „sofort“, d. h. in Echtzeit, verarbeiten müssen.

In Kapitel 2 gehen wir dazu zunächst auf einige analytische Aspekte von (P) ein, bevor wir im 3. Kapitel zwei speziell auf (P) zugeschnittene Algorithmen entwickeln. Zum Abschluss wenden wir diese Algorithmen in Kapitel 4 auf ein echtes Sprachsignal an und vergleichen die Ergebnisse mit denen herkömmlicher State-of-the-Art-Optimierungsalgorithmen.

Kapitel 2

Das exakte Optimierungsproblem

In diesem Kapitel widmen wir uns der analytischen Untersuchung des soeben hergeleiteten Optimierungsproblems. Um den gemeinsamen Sprachgebrauch abzustimmen, führen wir dazu erst einige Notationen ein.

2.1 Notationen

Von nun an sei $N \geq 1$ eine natürliche Zahl und $\Omega_N := \{1, ..., N\}$. Außerdem gelten die nachfolgenden Vereinbarungen.

Definition 2.1.

- $\mathbb{N}$ *bezeichne die Menge der natürlichen Zahlen ohne Null, wohingegen* $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

- $\mathbb{1}$ *bezeichne die Einheitsmatrix des euklidischen* $\mathbb{R}^{N \times N}$.

- *Für eine Matrix* $M \in \mathbb{R}^{N \times N}$ *und Indizes* $i, j \in \Omega_N$ *bezeichnen* $(M)_{ij}, m_{ij}, (M)_{i,j}$ *sowie* $m_{i,j}$ *den Eintrag in der* i*. Zeile und* j*. Spalte der Matrix* M*. Wir definieren zusätzlich* $m_{ij} := 0$ *für* $(i, j) \notin (\Omega_N \times \Omega_N)$*. Darüber hinaus stehe* $M_i \in \mathbb{R}^N$ *für die* i*. Spalte von* M.

- *Im Folgenden sei jedes Element* v *des* $\mathbb{R}^N$ *ein Spaltenvektor mit den Komponenten* v_i $(i \in \Omega_N)$*. Wir definieren zusätzlich* $v_i := 0$ *für* $i \notin \Omega_N$.
  Von dieser Notation abweichend sei $e_i := \mathbb{1}_i$ *der* i*. Einheitsvektor des euklidischen* $\mathbb{R}^N$.

- *Für zwei Vektoren* $u, v \in \mathbb{R}^N$ *stehe* $\langle u, v \rangle = u^T v$ *für das euklidische Skalarprodukt des* $\mathbb{R}^N$.

- *Ausdrücke der Form* e^x *und* $\ln(x)$ *mit* $x \in \mathbb{R}^N$ *seien komponentenweise zu*

verstehen, d.h.

$$e^x = \begin{pmatrix} e^{x_1} \\ \vdots \\ e^{x_N} \end{pmatrix} \text{ und } \ln(x) = \begin{pmatrix} \ln(x_1) \\ \vdots \\ \ln(x_N) \end{pmatrix}.$$

Es sei noch darauf hingewiesen, dass wir als „globale Lösung“ oder schlicht „Lösung“ eines Optimierungsproblems eine globale Optimalstelle (d. h. ein Argument) und nicht den Optimalwert der Zielfunktion bezeichnen.

Im Folgenden werden wir oft auch von „lokalen Lösungen“ sprechen, womit lokale Optimalstellen gemeint sind.

Wir unterscheiden ferner zwei Operatoren: Während

$$\min f(x)$$

den Optimalwert der Zielfunktion zurück gibt, ist das Ergebnis von

$$\arg\min f(x)$$

die/eine Optimalstelle.

Man beachte, dass ein Optimierungsproblem mehrere (globale) Optimalstellen haben kann, aber höchstens einen (globalen) Optimalwert.

2.2 Analytische Untersuchung

Das exakte Optimierungsproblem lautet, wie bereits gesehen:

$$\textbf{(P)} \quad \begin{cases} \min\limits_{x,y\in\mathbb{R}^N} \ \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d) \\ \text{s. t. } \forall i \in \Omega_N : e^{x_i} + e^{y_i} = 1 \end{cases}$$

mit $A, B \in \mathbb{R}^{N\times N}$ symmetrisch positiv definit und $c, d \in \mathbb{R}^N$.
Wir definieren die Zielfunktion von (P) hierbei wie folgt:

$$f : \mathbb{R}^N \times \mathbb{R}^N \longrightarrow \mathbb{R},$$

$$(x,y) \longmapsto f(x,y) := \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d).$$

Bevor wir uns der näheren Untersuchung von (P) widmen, stellen wir fest, dass wir durch Einsetzen der Nebenbedingung in die Zielfunktion weitere zu (P) äquivalente

Optimierungsprobleme formulieren können. Beispielsweise durch Auflösen der Nebenbedingung nach y erhält man:

$$(\mathbf{P_x}) \quad \left\{ \min_{x\in(-\infty,0)^N} \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(\ln(1-e^x)-d)^T B(\ln(1-e^x)-d) \right. .$$

Die Zielfunktion definieren wir dabei folgendermaßen:

$$\hat{f} : (-\infty,0)^N \longrightarrow \mathbb{R},$$
$$x \longmapsto \hat{f}(x) := \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(\ln(1-e^x)-d)^T B(\ln(1-e^x)-d).$$

Da (P) und ($\mathrm{P_x}$) äquivalent sind, gelten alle nachfolgenden Erkenntnisse gleichermaßen für beide Optimierungsprobleme. Aufgrund der Form von ($\mathrm{P_x}$), ziehen wir dieses aber zunächst für die analytische Untersuchung heran.

Es folgt nun ein Lemma, um den Gradienten und die Hessematrix der Zielfunktion $\hat{f}$ von ($\mathrm{P_x}$) zu berechnen.

Lemma 2.2. *Es sei $M \in \mathbb{R}^{N\times N}$ symmetrisch, $U \subseteq \mathbb{R}$ offen, $p \in \mathcal{C}^2(U)$ und*

$$h : U^N \longrightarrow \mathbb{R},\ x \longmapsto h(x) := \frac{1}{2}\cdot \vec{p}(x)^T M \vec{p}(x)$$

mit

$$\vec{p}(x) := \begin{pmatrix} p(x_1) \\ \vdots \\ p(x_N) \end{pmatrix}.$$

Dann gilt:

$$\nabla h(x) = \begin{pmatrix} p'(x_1) & & \\ & \ddots & \\ & & p'(x_N) \end{pmatrix} M\vec{p}(x)$$

und

$$\nabla^2 h(x) = \begin{pmatrix} p'(x_1) & & \\ & \ddots & \\ & & p'(x_N) \end{pmatrix} M \begin{pmatrix} p'(x_1) & & \\ & \ddots & \\ & & p'(x_N) \end{pmatrix}$$
$$+ \begin{pmatrix} p''(x_1)\cdot\langle M_1, \vec{p}(x)\rangle & & \\ & \ddots & \\ & & p''(x_N)\cdot\langle M_N, \vec{p}(x)\rangle \end{pmatrix}.$$

Der Beweis ergibt sich unmittelbar unter Anwendung elementarer Ableitungsregeln. Wir können nun leicht den folgenden Satz zeigen.

Satz 2.3 (Notwendiges Optimalitätskriterium für $(\mathrm{P_x})$)**.** *Jede (lokale) Lösung x von $(\mathrm{P_x})$ mit $x \in (-\infty, 0)^N$ genügt folgender Bedingung:*

$$A(x-c) + \begin{pmatrix} \frac{e^{x_1}}{1-e^{x_1}} & & \\ & \ddots & \\ & & \frac{e^{x_N}}{1-e^{x_N}} \end{pmatrix} B(d - \ln(1-e^x)) = 0. \tag{2.1}$$

Beweis. *Mit Lemma 2.2 erhalten wir zum einen*

$$\nabla \left[\frac{1}{2}(x-c)^T A(x-c) \right] = A(x-c)$$

und zum anderen

$$\nabla \left[\frac{1}{2}(\ln(1-e^x) - d)^T B(\ln(1-e^x) - d) \right] =$$

$$= \nabla \left[\frac{1}{2}(d - \ln(1-e^x))^T B(d - \ln(1-e^x)) \right]$$

$$= \begin{pmatrix} \frac{e^{x_1}}{1-e^{x_1}} & & \\ & \ddots & \\ & & \frac{e^{x_N}}{1-e^{x_N}} \end{pmatrix} B(d - \ln(1-e^x)).$$

Insgesamt gilt also:

$$\nabla \hat{f}(x) = A(x-c) + \begin{pmatrix} \frac{e^{x_1}}{1-e^{x_1}} & & \\ & \ddots & \\ & & \frac{e^{x_N}}{1-e^{x_N}} \end{pmatrix} B(d - \ln(1-e^x)),$$

Die Aussage des Satzes folgt nun unmittelbar aus dem bekannten Kriterium (vgl. [3]):

$$\nabla \hat{f}(x) \stackrel{!}{=} 0.$$

□

Es gibt noch beliebig viele weitere äquivalente Formulierungen zu dem hier vorgestellten Optimalitätskriterium (2.1). Insbesondere führt das nachfolgende Lagrange-System (KKT) von (P) auf die gleiche Bedingung (2.1), wenn man das Gleichungssystem (KKT) derart umformt, dass die Lagrange-Multiplikatoren $\lambda_i \in \mathbb{R}$ $(i = 1, ..., N)$ und die zweite Variable $y \in \mathbb{R}^N$ eliminiert werden.

$$\text{(KKT)} \quad \begin{cases} \langle A_i, (x - c)\rangle + \lambda_i \cdot e^{x_i} = 0, \ \forall i = 1, ..., N \\ \langle B_i, (y - d)\rangle + \lambda_i \cdot e^{y_i} = 0, \ \forall i = 1, ..., N \\ e^{x_i} + e^{y_i} = 1, \ \forall i = 1, ..., N \end{cases} .$$

Für den nächsten Satz, der uns ein hinreichendes Optimalitätskriterium liefern wird, benötigen wir vorweg noch ein kurzes Lemma über Summen und Produkte symmetrisch positiv definiter Matrizen.

Lemma 2.4. *$U, V \in \mathbb{R}^{N\times N}$ seien symmetrisch positiv definite Matrizen und ferner $W \in \mathbb{R}^{N\times N}$ eine symmetrisch positiv semi-definite Matrix, dann sind*

$$VUV$$

sowie

$$U + W$$

ebenfalls symmetrisch positiv definit.

Beweis. *Die Symmetrie von VUV und $U + W$ ist offensichtlich. Positive Definitheit liegt vor, da für alle $x \in \mathbb{R}^N, x \neq 0$ gilt:*

$$\langle x, VUVx\rangle = \langle \underbrace{Vx}_{=:y}, U \underbrace{Vx}_{=:y}\rangle = \langle y, Uy\rangle > 0.$$

y ist dabei sicher ungleich Null, da V symmetrisch positiv definit ist und damit insbesondere

$$\text{Kern(V)} = \{0\}.$$

Es ist ferner

$$\langle x, (U + W)x\rangle = \underbrace{\langle x, Ux\rangle}_{>0} + \underbrace{\langle x, Wx\rangle}_{\geq 0} > 0.$$

□

Satz 2.5 (Hinreichendes Optimalitätskriterium für $(\mathrm{P_x})$)**.** *Gilt für eine Stelle $x^* \in (-\infty, 0)^N$ neben der Bedingung aus Satz 2.3 zusätzlich*

$$\forall i \in \Omega_N : \sum_{j=1}^{N} b_{ij} \cdot \left[d_j - \ln(1 - e^{x_j^*}) \right] \geq 0, \tag{2.2}$$

so ist x^* eine (zumindest lokale) Lösung von $(\mathrm{P_x})$.
Gilt zudem

$$\forall x \in (-\infty, 0)^N \; \forall i \in \Omega_N : \sum_{j=1}^{N} b_{ij} \cdot [d_j - \ln(1 - e^{x_j})] \geq 0, \tag{2.3}$$

dann ist x^* sogar die einzige (und damit globale) Lösung von $(\mathrm{P_x})$.

Beweis. *Wir betrachten noch einmal die Funktion $\nabla \hat{f}$ aus dem Beweis von Lemma 2.3:*

$$\nabla \hat{f}(x) = A(x - c) + \underbrace{\begin{pmatrix} \frac{e^{x_1}}{1-e^{x_1}} & & \\ & \ddots & \\ & & \frac{e^{x_N}}{1-e^{x_N}} \end{pmatrix}}_{=:Q(x)} B(d - \ln(1 - e^x)).$$

Wendet man Lemma 2.2 nun auf beide Summanden an, ergibt sich:

$$\nabla^2 \hat{f}(x) = A + Q(x)BQ(x)$$

$$+ \underbrace{\begin{pmatrix} \frac{e^{x_1}}{(1-e^{x_1})^2} \cdot \langle B_1,\ d - \ln(1 - e^x) \rangle & & \\ & \ddots & \\ & & \frac{e^{x_N}}{(1-e^{x_N})^2} \cdot \langle B_N,\ d - \ln(1 - e^x) \rangle \end{pmatrix}}_{=:U(x)}.$$

Gemäß [8] ist eine kritische Stelle $x \in (-\infty, 0)^N$ dann eine Minimalstelle von $\hat{f}$, wenn die Matrix $\nabla^2 \hat{f}(x)$ positiv definit ist. Da

$$\forall x \in (-\infty, 0)^N \; \forall i \in \Omega_N : \frac{e^{x_i}}{1 - e^{x_i}} > 0,$$

ist die Matrix $Q(x)$ sicher positiv definit und damit gemäß Lemma 2.4 auch die Matrizen $Q(x)BQ(x)$ sowie $A + Q(x)BQ(x)$. Um ein hinreichendes Optimalitätskriterium zu erhalten, fehlt uns also nur noch eine Bedingung für die positive Semi-Definitheit von $U(x)$. Offensichtlich ist

$$\forall x \in (-\infty, 0)^N \; \forall i \in \Omega_N : \frac{e^{x_i}}{(1 - e^{x_i})^2} > 0,$$

womit $U(x)$ genau dann positiv semi-definit für ein $x^* \in (-\infty, 0)^N$ ist, wenn

$$\forall i \in \Omega_N : \langle B_i,\ d - \ln(1 - e^{x^*})\rangle = \sum_{j=1}^{N} b_{ij} \cdot \left[d_j - \ln(1 - e^{x_j^*})\right] \geq 0.$$

Ist die entsprechende Ungleichung (2.3) für alle $x \in (-\infty, 0)^N$ erfüllt, so ist $\hat{f}$ global strikt konvex und hat nach [8] genau ein Minimum. □

Ist das Kriterium (2.2) aus Satz 2.5 verletzt, kann man allerdings ohne Weiteres keine Aussage darüber treffen, ob ein Minimum vorliegt oder nicht. Der Grund hierfür ist, dass selbst bei einer negativ definiten Matrix $U(x)$ aus dem vorangehenden Beweis die Hessematrix $\nabla^2 \hat{f}(x)$ positiv definit sein kann.

Um eine etwas bessere Vorstellung von unserem Optimierungsproblem zu erlangen, ist in Abbildung 2.1 eine Niveaulinien-Zeichnung für den Fall $N = 1$ zu sehen.

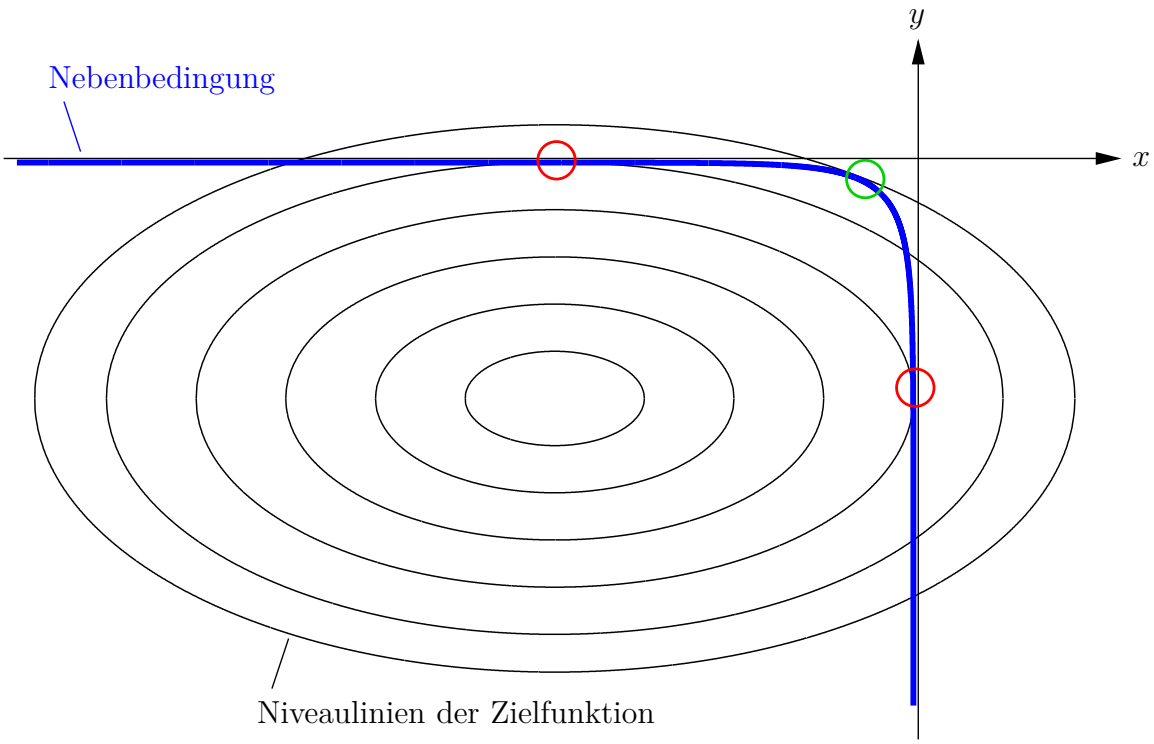


Abbildung 2.1: Zielfunktion und Nebenbedingung von (P) für $N = 1$

Die drei Kreise weisen dabei auf Stellen hin, an denen sich die Niveaulinien der Zielfunktion und der Nebenbedingung „berühren“, d. h. deren Gradienten linear abhängig sind. Es handelt sich also um Stellen, an denen das notwendige (Lagrange-) Optimalitätskriterium erfüllt ist. An den roten Kreisen liegen hier Minima und am grünen Kreis ein Maximum vor. Der folgende Satz zeigt, dass es im Fall $N = 1$ tatsächlich höchstens zwei (lokale) Minima geben kann.

Satz 2.6 (Obere Schranke für die Anzahl lokaler Lösungen)**.** *Handelt es sich bei A und B um Diagonalmatrizen, so hat* (P) *höchstens 2^N lokale Lösungen.*

Beweis. *Sind A, B in Diagonalform, so zerfällt* (P) *in N unabhängige Teilprobleme:*

$$\begin{aligned} \min_{x_i, y_i \in \mathbb{R}} \ & \tfrac{1}{2} a_{ii}(x_i - c_i)^2 + \tfrac{1}{2} b_{ii}(y_i - d_i)^2 \\ & s.\ t.\ e^{x_i} + e^{y_i} = 1, \end{aligned} \tag{2.4}$$

wobei $i \in \Omega_N$. Jedes dieser Teilprobleme ist äquivalent zu (P) *im Falle $N = 1$, weshalb wir uns im Rahmen dieses Beweises zunächst auf die Untersuchung von* (P) *für $N = 1$ beschränken können und daher kurz schreiben:*

$$\begin{aligned} \min_{x, y \in \mathbb{R}} \ & \tfrac{1}{2} a(x - c)^2 + \tfrac{1}{2} b(y - d)^2 \\ & s.\ t.\ e^{x} + e^{y} = 1 \end{aligned} \tag{2.5}$$

mit $a, b > 0$.
Das zugehörige Problem ($\mathrm{P_x}$) *lautet:*

$$\min_{x \in (-\infty, 0)} \ \tfrac{1}{2} a(x - c)^2 + \tfrac{1}{2} b(\ln(1 - e^x) - d)^2 \ , \tag{2.6}$$

wobei die Zielfunktion $\hat{f}$ dann entsprechend die Form

$$\hat{f}(x) = \frac{1}{2} a(x - c)^2 + \frac{1}{2} b(\ln(1 - e^x) - d)^2 \tag{2.7}$$

annimmt. In diesem Beweis interessieren wir uns für die zweite Ableitung von $\hat{f}$ und im Besonderen für das Vorzeichen von $\hat{f}''(x)$, da dieses Rückschlüsse auf die Anzahl der lokalen Minima von (2.7) zulässt.
Die zweite Ableitung von $\hat{f}$ nach x lautet, wie bereits im Beweis von Satz 2.5 allgemein berechnet:

$$\hat{f}''(x) = a + b \cdot \left(\frac{e^x}{1 - e^x} \right)^2 + b \cdot \frac{e^x}{(1 - e^x)^2} \cdot (d - \ln(1 - e^x)).$$

Für die Grenzwerte von $\hat{f}''(x)$ gilt:

$$\lim_{x \to -\infty} \hat{f}''(x) = \lim_{x \to -\infty} \left[a + b \cdot \underbrace{\left(\frac{e^x}{1 - e^x} \right)^2}_{\to 0} + b \cdot \underbrace{\frac{e^x}{(1 - e^x)^2}}_{\to 0} \cdot \underbrace{(d - \ln(1 - e^x))}_{\to d} \right] = a > 0,$$

$$\lim_{x \nearrow 0} \hat{f}''(x) = \lim_{x \nearrow 0} \left[a + b \cdot \underbrace{\left(\frac{e^x}{1 - e^x} \right)^2}_{\to +\infty} + b \cdot \underbrace{\frac{e^x}{(1 - e^x)^2}}_{\to +\infty} \cdot \underbrace{(d - \ln(1 - e^x))}_{\to +\infty} \right] = +\infty > 0.$$

Um das Vorzeichen (bzw. die Vorzeichenwechsel) von $\hat{f}''$ genauer zu untersuchen, unterscheiden wir zwei Fälle:

1. Fall: $d \geq 0$
In diesem Fall ist $\hat{f}''(x)$ für alle $x \in (-\infty, 0)$ strikt positiv, denn:

$$\hat{f}''(x) = \underbrace{a}_{>0} + \underbrace{b \cdot \left(\frac{e^x}{1-e^x}\right)^2}_{>0} + \underbrace{b \cdot \frac{e^x}{(1-e^x)^2}}_{>0} \cdot \underbrace{(\underbrace{d}_{\geq 0} - \underbrace{\ln(1-e^x)}_{<0 \text{ da } x<0})}_{>0} > 0.$$

Damit ist $\hat{f}$ strikt konvex auf $(-\infty, 0)$ und hat mithin genau ein Minimum (vgl. [8]).

2. Fall: $d < 0$
Die entscheidende Frage ist die nach den Nullstellen von $\hat{f}''(x)$:

$$\hat{f}''(x) = a + b \cdot \left(\frac{e^x}{1-e^x}\right)^2 + b \cdot \frac{e^x}{(1-e^x)^2} \cdot (d - \ln(1-e^x)) \stackrel{!}{=} 0. \tag{2.8}$$

Zur leichteren Untersuchung von (2.8) führen wir folgende Substitution ein:

$$z := \frac{e^x}{1-e^x}$$

mit

$$z \in (0, +\infty), \text{ da } x \in (-\infty, 0)$$

und definieren

$$\begin{aligned}
s(z) &:= \frac{1}{b} \cdot \hat{f}''(x) = \frac{1}{b} \cdot \hat{f}''(x(z)) \\
&= \frac{a}{b} + \left(\frac{e^x}{1-e^x}\right)^2 + \frac{e^x}{(1-e^x)^2} \cdot (d - \ln(1-e^x)) \\
&= \frac{a}{b} + \underbrace{\left(\frac{e^x}{1-e^x}\right)^2}_{=z^2} + \underbrace{e^{-x}}_{=\frac{1+z}{z}} \cdot \underbrace{\frac{e^{2x}}{(1-e^x)^2}}_{=z^2} \cdot (d - \ln(\underbrace{1-e^x}_{=1-\frac{z}{1+z}=\frac{1}{1+z}})) \\
&= \frac{a}{b} + z^2 + \frac{1+z}{z} \cdot z^2 \cdot \left(d - \ln\left(\frac{1}{1+z}\right)\right) \\
&= \frac{a}{b} + z^2 + (1+z) \cdot z \cdot (\ln(1+z) + d) \\
&= \frac{a}{b} + z^2 + \ln(1+z) \cdot z + \ln(1+z) \cdot z^2 + d \cdot z + d \cdot z^2.
\end{aligned}$$

Man beachte, dass $z(x)$ im Definitionsbereich streng monoton steigend in x und bijektiv ist, denn:

$$x_1 < x_2 \in (-\infty, 0) \Longrightarrow z(x_1) = \frac{\overbrace{e^{x_1}}^{<e^{x_2}}}{\underbrace{1-e^{x_1}}_{>1-e^{x_2}}} < \frac{e^{x_2}}{1-e^{x_2}} = z(x_2)$$

und

$$z = \frac{e^x}{1-e^x} \iff z - z \cdot e^x = e^x \iff z = (1+z) \cdot e^x \iff x = \ln\left(\frac{z}{1+z}\right).$$

Damit ist ebenso $x(z)$ im Definitionsbereich streng monoton steigend in z. $s(z)$ ist also eine reskalierte Variante von $\hat{f}''(x)$ mit allerdings denselben Vorzeichenwechseln. Genau diese Vorzeichenwechsel möchten wir als nächstes untersuchen und betrachten dazu die Ableitungen von s:

$$\begin{aligned}
s(z) &= \frac{a}{b} + z^2 + \ln(1+z) \cdot z + \ln(1+z) \cdot z^2 + d \cdot z + d \cdot z^2, \\
s'(z) &= 2z + \frac{z}{1+z} + \ln(1+z) + \frac{z^2}{1+z} + 2z \cdot \ln(1+z) + d + 2d \cdot z \\
&= 2z + \frac{z+z^2}{1+z} + \ln(1+z) + 2z \cdot \ln(1+z) + d + 2d \cdot z \\
&= 2z + z \cdot \frac{1+z}{1+z} + \ln(1+z) + 2z \cdot \ln(1+z) + d + 2d \cdot z \\
&= 3z + \ln(1+z) + 2z \cdot \ln(1+z) + d + 2d \cdot z, \\
s''(z) &= 3 + \frac{1}{1+z} + \frac{2z}{1+z} + 2 \cdot \ln(1+z) + 2d \\
&= 3 + \frac{2z+1}{1+z} + 2 \cdot \ln(1+z) + 2d, \\
s'''(z) &= \frac{1}{(1+z)^2} + \frac{2}{1+z} > 0.
\end{aligned}$$

Da nun aber $s'''(z)$ strikt positiv ist, muss $s'(z)$ strikt konvex sein.
Ferner gilt für die Grenzwerte von $s'(z)$:

$$\lim_{z \searrow 0} s'(z) = d < 0,$$
$$\lim_{z \to +\infty} s'(z) = +\infty,$$

woraus folgt, dass $s'(z)$ genau eine Nullstelle hat und zwar mit einem Vorzeichenwechsel von $-$ nach $+$, d. h. es existiert genau ein $z_0 \in (0, +\infty)$ mit $s'(z_0) = 0$ und

$$\forall z \in (0, z_0) : s'(z) < 0,$$
$$\forall z \in (z_0, +\infty) : s'(z) > 0.$$

Das bedeutet für die Funktion $s(z)$, dass sie zuerst fällt und dann steigt. Da zudem für die Grenzwerte von $s(z)$

$$\lim_{z \searrow 0} s(z) = \frac{a}{b} > 0$$

und

$$\lim_{z\to+\infty} s(z) = +\infty > 0$$

gilt, kann s höchstens zwei Nullstellen haben. Mithin kann $\hat{f}''$ höchstens zwei Nullstellen und daher $\hat{f}$ höchstens zwei Wendepunkte haben. Mit

$$\lim_{x\to-\infty} \hat{f}(x) = \lim_{x\to-\infty} \left[\frac{1}{2}a\underbrace{(x-c)^2}_{\to+\infty} + \frac{1}{2}b(\underbrace{\ln(1-e^x)}_{\to 0} - d)^2\right] = +\infty,$$

$$\lim_{x\nearrow 0} \hat{f}(x) = \lim_{x\nearrow 0} \left[\frac{1}{2}a\underbrace{(x-c)^2}_{\to c^2} + \frac{1}{2}b(\underbrace{\ln(1-e^x)}_{\to-\infty} - d)^2\right] = +\infty$$

folgt schließlich, dass $\hat{f}$ höchstens zwei Minima haben kann.
Bevor wir nun die Untersuchung des 2. Falles (nämlich $d < 0$) abschließen, sei noch auf einen Aspekt hingewiesen:
Wählt man a/b hinreichend groß im Vergleich zu d, so hat s gar keine Nullstelle, womit $\hat{f}$ global strikt konvex ist und daher genau ein Minimum besitzt. Es gibt sogar einen Grenzfall für genau eine Wahl von a/b im Vergleich zu d, in dem s genau eine Nullstelle (und zwar ohne Vorzeichenwechsel) hat. Unter diesen Umständen hätte $\hat{f}$ ebenfalls genau ein Minimum mit zudem verschwindender zweiter Ableitung.

Insgesamt gibt es also für jedes Teilproblem (2.4) höchstens 2 Minima. Nachdem es N dieser unabhängigen Teilprobleme gibt, können bis zu 2^N lokale Lösungen für (P) *existieren.* □

Kapitel 3

Das approximierte Optimierungsproblem

In diesem Kapitel befassen wir uns mit einer Approximation unseres exakten Optimierungsproblems und werden dabei in den Kapiteln 3.4 und 3.5 zwei Lösungsalgorithmen entwickeln.

Doch vorweg zunächst wieder einige wichtige Konventionen.

3.1 Notationen

Von nun an gelten zusätzlich zu den Festlegungen aus Definition 2.1 die nachfolgenden Notationen.

Definition 3.1.

- *Das Komplement $\bar{U}$ einer Teilmenge $U \subseteq V$ sei definiert als $\bar{U} := V \setminus U$.*
- *$|U|$ bezeichne die Kardinalität einer Teilmenge $U \subseteq V$ und ∂U den Rand von U.*
- *Für eine Matrix $M \in \mathbb{R}^{N \times N}$ und eine Teilmenge $I \subseteq \mathbb{N}_0$ seien M_I und $M_{I,I}$ wie folgt definiert:*

$$(M_I)_{i,j} := \begin{cases} 0, \ \textit{falls } i \in (I \cap \Omega_N) \\ (M)_{i,j}, \ \textit{sonst} \end{cases},$$

$$(M_{I,I})_{i,j} := \begin{cases} 1, \ \textit{falls } (i \in (I \cap \Omega_N) \ \vee \ j \in (I \cap \Omega_N)) \wedge i = j \\ 0, \ \textit{falls } (i \in (I \cap \Omega_N) \vee j \in (I \cap \Omega_N)) \wedge i \neq j \\ (M)_{i,j}, \ \textit{sonst} \end{cases}.$$

Für die Inverse von $M_{I,I}$ schreiben wir kurz:

$$M_{I,I}^{-1} := (M_{I,I})^{-1}.$$

- *Für $M \in \mathbb{R}^{N\times N}$ und $I, J \subseteq \Omega_N$ stehe $M^I \in \mathbb{R}^{(N-|I|)\times N}$ für die Matrix M, wobei alle Zeilen $i \in I$ gestrichen sind. $M^{I,J} \in \mathbb{R}^{(N-|I|)\times(N-|J|)}$ bezeichne die Matrix M, bei der alle Zeilen $i \in I$ und Spalten $j \in J$ gestrichen sind.*

- *Für ein Element $i \in \Omega_N$ und eine Teilmenge $I \subseteq \Omega_N$ gelten folgende Schreibweisen:*

  - *Der „Vorgänger“ $v_I(i) \in I$ von i sei das bzgl. i nächstkleinere Element in I, d. h. $v_I(i) < i \ \wedge \ \nexists j \in I : v_I(i) < j < i$. Existiert kein solcher Vorgänger, setzen wir $v_I(i) := 0 \notin I$.*
  - *Der „Nachfolger“ $n_I(i) \in I$ von i sei das bzgl. i nächstgrößere Element in I, d. h. $i < n_I(i) \ \wedge \ \nexists j \in I : i < j < n_I(i)$. Existiert kein solcher Nachfolger, setzen wir $n_I(i) := 0 \notin I$.*

- *Für $i, j \in \mathbb{N}$ gelten folgende Schreibweisen:*

$$\varepsilon_{i,j} := \begin{cases} 0, & \text{falls } i = j = 0 \\ 1, & \text{falls } i = 0 \ \dot\vee \ j = 0 \\ 2, & \text{falls } i \neq 0 \ \wedge \ j \neq 0 \end{cases},$$

$$\varepsilon_i := \varepsilon_{i,0},$$

wobei $\dot\vee$ für die exklusive Disjunktion steht.

Es folgen nun einige Beispiele, um die gerade definierten Schreibweisen zu veranschaulichen.

Beispiel 3.2. *Für eine Matrix $M \in \mathbb{R}^{3\times3}$, $I = \{2\}$ und $J = \{3\}$ ist:*

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}, \quad M_I = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ 0 & 0 & 0 \\ m_{31} & m_{32} & m_{33} \end{pmatrix}, \quad M_{I,I} = \begin{pmatrix} m_{11} & 0 & m_{13} \\ 0 & 1 & 0 \\ m_{31} & 0 & m_{33} \end{pmatrix},$$

$$M^I = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}, \quad M^{I,J} = \begin{pmatrix} m_{11} & m_{12} \\ m_{31} & m_{32} \end{pmatrix}, \quad M^{J,I} = \begin{pmatrix} m_{11} & m_{13} \\ m_{21} & m_{23} \end{pmatrix}.$$

Beispiel 3.3. *Für $N = 5$ und $I = \{1, 3\}$ ist*

$$\mathbb{1}^I = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

und

$$\mathbb{1}_I = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \mathbb{1}^I(\mathbb{1}^I)^T.$$

Beispiel 3.4. *Für $N = 5$ und $I = \{1, 3, 5\} \subseteq \Omega_N$ gilt:*

$$\Omega_N = \{1, 2, 3, 4, 5\},$$

$$\bar{I} = \{2, 4\}$$

und

$$v_I(3) = 1,\ n_I(3) = 5,$$
$$v_{\bar{I}}(3) = 2,\ n_{\bar{I}}(3) = 4.$$

3.2 Grundlagen

In diesem Kapitel wenden wir uns einer Approximation unseres exakten Optimierungsproblems zu.

Zur Wiederholung nachfolgend nochmals das exakte Optimierungsproblem.

$$\textbf{(P)} \begin{cases} \min\limits_{x,y \in \mathbb{R}^N} \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d) \\ \text{s. t. } \forall i \in \Omega_N : e^{x_i} + e^{y_i} = 1 \end{cases}$$

Wir richten den Blick nun auf die exakte Nebenbedingung für ein beliebiges $i \in \Omega_N$ (vgl. Abbildung 3.1) und nähern diese abschnittsweise linear durch die negativen Koordinatenhalbachsen an (vgl. Abbildung 3.2).

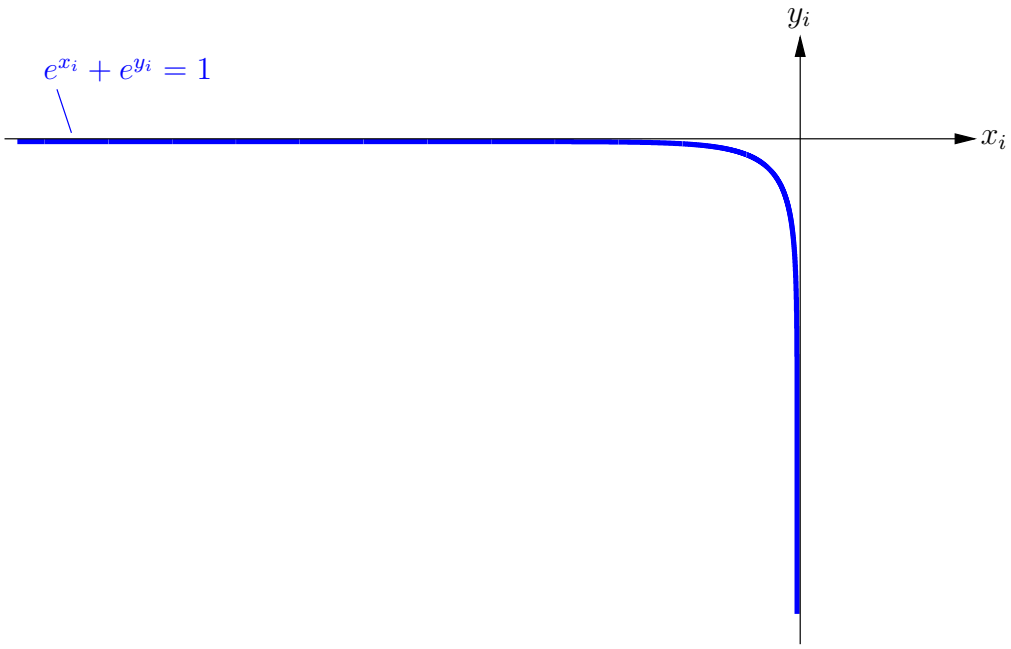


Abbildung 3.1: Exakte Nebenbedingung für $i \in \Omega_N$

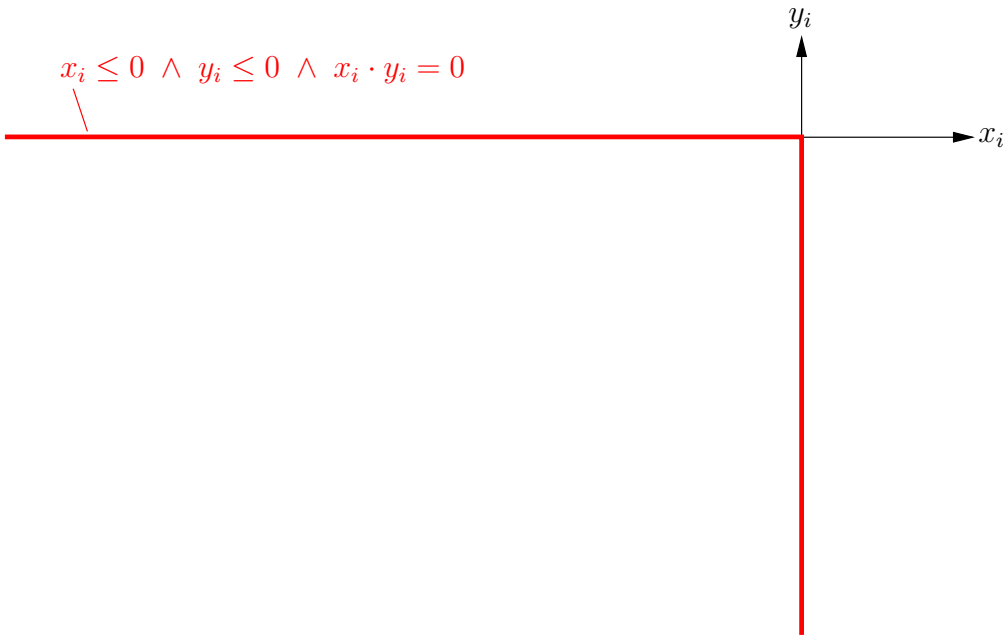


Abbildung 3.2: Approximierte Nebenbedingung für $i \in \Omega_N$

Unser approximiertes Optimierungsproblem lautet dementsprechend:

$$\textbf{(AP1)} \quad \begin{cases} \min\limits_{x,y\in\mathbb{R}^N} \ \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d) \\ \text{s. t. } \forall i \in \Omega_N : x_i \leq 0 \ \wedge \ y_i \leq 0 \ \wedge x_i \cdot y_i = 0 \end{cases}.$$

Diese Approximation der Nebenbedingung stellt für herkömmliche Optimierungsalgorithmen (wie etwa SQP-Verfahren) erheblich Probleme dar und ist u. U. sogar numerisch instabiler als das exakte Problem. Der Grund hierfür liegt darin, dass sich der nun entstandene „Knick“ in der Nebenbedingung sehr schlecht durch abschnittsweise differenzierbare Funktionen annähern lässt, was aber im Rahmen vieler herkömmlicher Verfahren versucht wird.

Wie wir in Kapitel 3.4 jedoch sehen werden, kann diese Art der abschnittsweisen Linearisierung der Nebenbedingung auch immense Vorteile mit sich bringen — und zwar insbesondere dann, wenn, wie in unserem Anwendungsfall, die Laufzeit eines Lösungsalgorithmus im Vordergrund steht.

Wir möchten zunächst etwas mehr über das Problem (AP1) erfahren und stellen uns daher die Frage nach dem Approximationsfehler, den wir begehen.

Satz 3.5 (Approximationsfehler im Diagonalfall)**.** *Sind A und B Diagonalmatrizen, (x^*, y^*) eine globale Lösung von* (P) *und $(\hat{x}^*, \hat{y}^*)$ eine globale Lösung von* (AP1)*, dann lässt sich das exakte Optimum summandenweise für alle $i \in \Omega_N$ durch das approximierte nach oben abschätzen:*

$$f_i(x_i^*, y_i^*) \leq \begin{cases} f_i(\hat{x}_i^*, \hat{y}_i^*), & \text{falls } e^{c_i} + e^{d_i} \leq 1 \\ f_i(\hat{x}_i^* - \ln 2, \hat{y}_i^* - \ln 2), & \text{falls } e^{c_i} + e^{d_i} \geq 1 \end{cases}.$$

Ist $c_i, d_i \in \mathbb{R} \setminus (-\ln 2, 0)$, so gilt darüber hinaus folgende Abschätzung nach unten:

$$f_i(x_i^*, y_i^*) \geq \begin{cases} f_i(\hat{x}_i^*, \hat{y}_i^*) & \text{falls } c_i, d_i \geq 0 \\ f_i(\hat{x}_i^* - \ln 2, \hat{y}_i^* - \ln 2) & \text{falls } c_i, d_i \leq -\ln 2 \end{cases}.$$

f_i ist dabei der i. Summand der Zielfunktion f, d. h.

$$f(x, y) = \sum_{i=1}^{N} f_i(x_i, y_i)$$

mit

$$f_i(x_i, y_i) := \frac{1}{2} a_{ii}(x_i - c_i)^2 + \frac{1}{2} b_{ii}(y_i - d_i)^2.$$

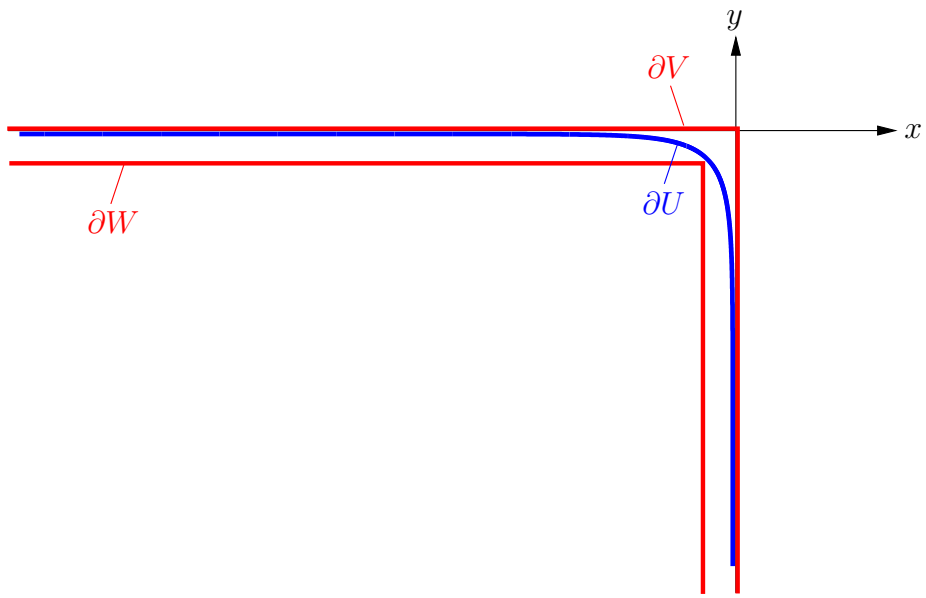


Abbildung 3.3: Schachtelung der exakten Nebenbedingung

Beweis. *Da* (P) *für diagonale Matrizen* A *und* B *in* N *unabhängige Teilprobleme zerfällt, konzentrieren wir uns im Folgenden auf die Untersuchung von* (P) *im eindimensionalen Fall. Für die Gesamtheit dieses Beweis sei also* $N = 1$ *und*

$$f(x,y) = \frac{1}{2}a(x-c)^2 + \frac{1}{2}b(y-d)^2$$

mit $a, b > 0$.

f *ist streng monoton fallend entlang eines jeden geradlinigen Weges zum Scheitel* (c,d), *d. h. mit*

$$h(\lambda) := f((x,y) + \lambda(c-x, d-y)),\ x,y \in \mathbb{R}$$

gilt für alle $0 \leq \lambda_1 \leq \lambda_2 \leq 1$*:*

$$h(\lambda_1) \geq h(\lambda_2), \tag{3.1}$$

was sich durch Berechnen der ersten Ableitung von h *leicht nachprüfen lässt.*

Wir definieren die folgenden Mengen, deren Ränder sich auch in Abbildung 3.3 wieder finden:

$$\begin{aligned} U &:= \{(x,y) \in \mathbb{R}^2 \mid e^x + e^y \leq 1\}, \\ V &:= \{(x,y) \in \mathbb{R}^2 \mid x \leq 0 \ \wedge\ y \leq 0\}, \\ W &:= \{(x,y) \in \mathbb{R}^2 \mid x \leq -\ln 2 \ \wedge\ y \leq -\ln 2\}. \end{aligned}$$

Man beachte, dass ∂U *die zulässige Menge von* (P), ∂V *die zulässige Menge von* (AP1)

und ∂W die zulässige Menge von

$$\begin{aligned} &\min_{x,y\in\mathbb{R}} \ \tfrac{1}{2}a(x-c)^2 + \tfrac{1}{2}b(y-d)^2 \\ &s.\ t.\ x \leq -\ln 2 \ \wedge \ y \leq -\ln 2 \ \wedge (x+\ln 2)(y+\ln 2) = 0 \end{aligned} \tag{3.2}$$

ist. Durch Koordinatentransformation erhält man dabei unmittelbar:

$$\arg\min_{x,y\in\partial W} \ f(x,y) = \arg\min_{x,y\in\partial V} \ f(x-\ln 2, y-\ln 2).$$

Offensichtlich ist

$$W \subseteq U \subseteq V$$

und dementsprechend

$$\bar{W} \supseteq \bar{U} \supseteq \bar{V}.$$

Für die Abschätzungen nach oben und unten unterscheiden wir nun jeweils zwei Fälle:

1. Fall: $(c,d) \in (\bar{U} \cup \partial U)$
Ist $(x^*, y^*) \in \partial U$ eine globale Lösung von (P)*, dann gilt für alle $(x,y) \in \partial W \subseteq U$:*

$$f(x^*, y^*) \leq f(x,y).$$

Zum Beweis nehmen wir an, es gäbe ein $(x,y) \in U$ mit

$$f(x^*, y^*) > f(x,y).$$

Da $(x,y) \in U$ und $(c,d) \in (\bar{U} \cup \partial U)$, existierte dann aber ein $(x^{**}, y^{**}) \in \partial U$ auf dem geradlinigen Weg zwischen (x,y) und (c,d). Damit wäre nach (3.1)

$$f(x^{**}, y^{**}) \leq f(x,y) < f(x^*, y^*)$$

und (x^*, y^*) keine globale Lösung von (P)*.*

Alle weiteren Fälle lassen sich nach völlig analogem Muster beweisen.

2. Fall: $(c,d) \in U$
Ist $(x^*, y^*) \in \partial U$ eine globale Lösung von (P)*, dann gilt für alle $(x,y) \in \partial V \subseteq \bar{U}$:*

$$f(x^*, y^*) \leq f(x,y).$$

3. Fall: $(c,d) \in (\bar{V} \cup \partial V)$
Ist $(x^*, y^*) \in \partial V$ eine globale Lösung von (AP1)*, dann gilt für alle $(x,y) \in \partial U \subseteq V$:*

$$f(x^*, y^*) \leq f(x,y).$$

4. Fall: $(c,d) \in W$
Ist $(x^*, y^*) \in \partial W$ eine globale Lösung von (3.2), dann gilt für alle $(x,y) \in \partial U \subseteq \bar{W}$:

$$f(x^*, y^*) \leq f(x,y).$$

□

In Abbildung 3.2 kann man bereits erahnen, dass sich das Problem (AP1) auch als so genanntes „disjunktives“ Optimierungsproblem auffassen lässt. Dies bedeutet, dass für jedes $i \in \Omega_N$ die Entscheidung getroffen werden muss, ob $x_i = 0$ oder $y_i = 0$ zu setzen ist. Hat man diese Entscheidung jedoch für alle $i \in \Omega_N$ getroffen, so vereinfacht sich das verbleibende Optimierungsproblem zu einem linearen Gleichungssystem (vgl. Satz 3.15), welches man mit heutigen Algorithmen sehr effizient lösen kann.

Die elementare Frage, die sich nun stellt und die auch gleichzeitig einen entscheidenden Teil dieser Masterarbeit ausmachen wird, ist die nach einem sinnvollen Entscheidungskriterium für die Disjunktion in der Nebenbedingung von (AP1).

Um uns nun genau dieser Frage anzunähern, definieren wir ein weiteres Optimierungsproblem, dessen Nebenbedingung nur noch aus besagter Disjunktion besteht und keine Negativitätsrestriktionen mehr aufweist:

$$\textbf{(AP2)} \quad \begin{cases} \min\limits_{x,y\in\mathbb{R}^N} \ \frac{1}{2}(x-c)^T A(x-c) + \frac{1}{2}(y-d)^T B(y-d) \\ \text{s. t. } \forall i \in \Omega_N : x_i = 0 \ \vee \ y_i = 0 \end{cases}.$$

Wir zerlegen (AP2) weiter in zwei Teilprobleme.

Definition 3.6. *Für $I, J \subseteq \Omega_N$ definieren wir*

$$\Phi_1(I) := \min_{x\in\mathbb{R}^N} \ \frac{1}{2}(x-c)^T A(x-c)$$

$$s.\ t.\ \forall i \in I : x_i = 0,$$

und

$$\Phi_2(J) := \min_{y\in\mathbb{R}^N} \ \frac{1}{2}(y-d)^T B(y-d)$$

$$s.\ t.\ \forall j \in J : y_j = 0.$$

Wie wir später zeigen werden, handelt es sich bei Φ_1 und Φ_2 tatsächlich um Abbildungen von $\mathcal{P}(\Omega_N)$ nach $\mathbb{R}$, da die in deren Definition vorkommenden Minimierungsprobleme für alle $I, J \subseteq \Omega_N$ genau eine Lösung besitzen (vgl. Satz 3.15). Φ_1 und Φ_2 sind somit also wohldefiniert.

Damit führen wir ein weiteres Optimierungsproblem ein

$$\textbf{(AP2-I)} \quad \begin{cases} \min\limits_{I\subseteq\Omega_N} \ \Phi_1(I) + \Phi_2(\bar{I}) \end{cases}$$

und formulieren sogleich den folgenden Satz.

Satz 3.7. *Es gilt:*

$$\text{(AP2)} \iff \text{(AP2-I)}.$$

Beweis. *Zu allererst stellen wir fest, dass die Zielfunktionen von* (AP2) *und* (AP2-I) *identisch sind.*

„$\Longrightarrow$“:
Es sei (x^*, y^*) eine globale Lösung von (AP2) *und $I \subseteq \Omega_N$ bzw. $J \subseteq \Omega_N$ die Menge aller Indizes, für welche die Restriktion $x_i = 0$ bzw. $y_i = 0$ aktiv war. Offensichtlich muss $I \cup J = \Omega_N$ gelten, da andernfalls die Nebenbedingung von* (AP2) *für wenigstens eine Komponente verletzt gewesen wäre. Es ist also $J \supseteq \bar{I}$ und mithin I eine globale Lösung von* (AP2-I).

„$\Longleftarrow$“:
Es sei $I \subseteq \Omega_N$ eine globale Lösung von (AP2-I) *und (x^*, y^*) die zugehörige Optimalstelle der Zielfunktion f. Dann ist $x_i^* = 0$ $(\forall i \in I)$ und $y_i^* = 0$ $(\forall i \in \bar{I})$, was*

$$\forall i \in \Omega_N : \; x_i^* = 0 \; \vee \; y_i^* = 0.$$

impliziert. Damit ist (x^*, y^*) eine globale Lösung von (AP2). □

(AP2) und (AP2-I) sind also äquivalente Optimierungsprobleme mit dem Unterschied, dass (AP2-I) das eigentliche Kernproblem besser aufzeigt, nämlich das Finden einer Indexmenge $I \subseteq \Omega_N$, mit anderen Worten also das Finden einer optimalen Entscheidung bzgl. der oben angesprochenen Disjunktion.

Anmerkung 3.8 (Hinweis zum Sprachgebrauch)**.** *Die Bezeichnung „Lösung von* (AP2-I)*“ stehe je nach Kontext für die optimale Menge I oder für die zugehörige Stelle (x, y). Dementsprechend wird jede Menge $I \subseteq \Omega_N$ als eine (zumindest lokale) Lösung von* (AP2-I) *angesehen.*

Bevor wir uns aber nunmehr der näheren Untersuchung von (AP2-I) hingeben, bleibt noch der Bogen zu (AP1) zu spannen, weshalb wir ein letztes Problem definieren müssen:

$$\textbf{(AP2-II)} \quad \begin{cases} \min\limits_{\substack{I,J \subseteq \Omega_N \\ I \cup J = \Omega_N}} \Phi_1(I) + \Phi_2(J) \; . \end{cases}$$

Lemma 3.9. *Für alle $I_1 \subseteq I_2 \subseteq \Omega_N$ gilt:*

$$\Phi_1(I_1) \leq \Phi_1(I_2)$$

und gleichermaßen

$$\Phi_2(I_1) \leq \Phi_2(I_2).$$

Beweis. *Die zulässige Menge des Minimierungsproblems zu $\Phi_1(I_1)$ ist*

$$U := \{x \in \mathbb{R}^N \mid \forall i \in I_1 : x_i = 0\}$$

und die zu $\Phi_1(I_2)$ gehörende zulässige Menge ist

$$V := \{x \in \mathbb{R}^N \mid \forall i \in I_2 : x_i = 0\}$$

Die Aussage folgt nun unmittelbar, da offensichtlich $V \subseteq U$ gilt. □

Gemäß nachfolgendem Satz sind auch (AP2-I) und (AP2-II) äquivalent, was auf den ersten Blick darauf hindeutete, dass wir durch die Einführung von (AP2-II) nichts gewonnen hätten. Allerdings besitzt (AP2-II) im Vergleich zu (AP2-I) zusätzliche lokale Lösungen, deren Rolle nun verdeutlicht werden soll.

Satz 3.10. *Es gilt:*

1. (AP2-I) $\Longleftrightarrow$ (AP2-II).
2. *Jede globale Lösung von* (AP2)*, die der Nichtpositivitätsforderung von* (AP1) *genügt, ist auch eine globale Lösung von* (AP1).
3. *Jede lokale Lösung von* (AP1) *ist auch eine lokale Lösung von* (AP2-II). *Umgekehrt ist auch jede lokale Lösung von* (AP2-II)*, die der Nichtpositivitätsforderung von* (AP1) *genügt, eine lokale Lösung von* (AP1).

Beweis. *Zu 1.:*
Es sei $I \subseteq \Omega_N$ eine globale Lösung von (AP2-I)*, dann ist $(I, \bar{I})$ eine globale Lösung von* (AP2-II)*, da $I \cup \bar{I} = \Omega_N$ und nach Lemma 3.9*

$$\Phi_2(\bar{I}) \leq \Phi_2(J)$$

für alle $J \supseteq \bar{I}$, d. h. für alle J mit $I \cup J = \Omega_N$ gilt.

Es sei (I, J) eine globale Lösung von (AP2-I)*, dann ist I eine globale Lösung von* (AP2-I)*, da $\bar{I} \subseteq J$ und nach Lemma 3.9 wieder*

$$\Phi_2(\bar{I}) \leq \Phi_2(J)$$

ist.

Zu 2.:
Es sei (x^*, y^*) eine globale Lösung von (AP2)*, U die zulässige Menge von* (AP2) *und V die zulässige Menge von* (AP1)*. Gilt $(x^*, y^*) \in V$, so ist (x^*, y^*) auch Lösung von* (AP1)*, da $V \subseteq U$.*

Zu 3.:
Es sei (x^*, y^*) eine lokale Lösung von (AP1) *und I, J wie folgt definiert:*

$$I := \{i \in \Omega_N \mid x_i^* = 0\},$$

$$J := \{i \in \Omega_N \mid y_i^* = 0\}.$$

Aufgrund der Disjunktion in der Nebenbedingung von (AP1) *gilt:*

$$I \cup J = \Omega_N. \tag{3.3}$$

Diese spezielle Optimalstelle (x^*, y^*) kann man daher auch als Lösung folgender Probleme ansehen:

$$\begin{aligned} x^* &= \arg\min_{x \in \mathbb{R}^N} \ \tfrac{1}{2}(x-c)^T A(x-c) \\ \text{s. t.: } &\forall i \in I: \ h_i(x) := x_i = 0, \ \forall i \in \bar{I}: \ g_i(x) := x_i \leq 0, \end{aligned} \tag{3.4}$$

$$\begin{aligned} y^* &= \arg\min_{y\in\mathbb{R}^N} \ \tfrac{1}{2}(y-d)^T B(y-d) \\ \text{s. t.: } &\forall i \in J: \ y_i = 0, \ \forall i \in \bar{J}: \ y_i \leq 0. \end{aligned} \tag{3.5}$$

Wir betrachten nun die Lagrange-Funktion zu (3.4):

$$L(x) = \frac{1}{2}(x-c)^T A(x-c) + \sum_{i\in\bar{I}} \lambda_i g_i(x) + \sum_{i\in I} \mu_i h_i(x), \ \lambda_i, \mu_i \in \mathbb{R}$$

und das zugehörige KKT-System (vgl. [2]):

$$\begin{aligned} \nabla L(x) &= 0, \\ h_i(x) &= 0 \ (\forall i \in I), \\ \lambda_i &\geq 0 \ (\forall i \in \bar{I}), \\ g_i(x) &\leq 0 \ (\forall i \in \bar{I}), \\ \sum_{i\in\bar{I}} \lambda_i g_i(x) &= 0. \end{aligned}$$

Definitionsgemäß gilt nun aber für alle $i \in \bar{I}$:

$$x_i < 0,$$

da andernfalls $i \in I$ wäre. Um die KKT-Bedingung

$$\sum_{i\in\bar{I}} \lambda_i g_i(x) = \sum_{i\in\bar{I}} \lambda_i x_i \stackrel{!}{=} 0 \ \textit{mit} \ \lambda_i \geq 0$$

zu erfüllen, müssen somit alle $\lambda_i = 0$ sein. Für $\nabla L(x)$ impliziert dies:

$$\nabla L(x) = A(x-c) + \sum_{i\in\bar{I}} 0 \cdot e_i + \sum_{i\in I} \mu_i \cdot e_i.$$

Damit ist x^* Lösung des folgenden linearen Gleichungssystems:

$$\begin{aligned} A(x-c) + \sum_{i\in I} \mu_i \cdot e_i &= 0, \\ \forall i \in I: x_i &= 0, \end{aligned}$$

was äquivalent zum Lagrange-System von $\Phi_1(I)$ ist. Völlig analog lässt sich zeigen, dass y^* die Lösung des Lagrange-System zu $\Phi_2(J)$ ist. Mit (3.3) folgt schließlich, dass (x^*, y^*) bzw. (I, J) eine lokale Lösung von (AP2-II) *ist.*

Umgekehrt kann man sich natürlich auch eine beliebige lokale Lösung (I, J) von (AP2-II)*, d. h. also beliebige Mengen $I, J \subseteq \Omega_N$ mit $I \cup J = \Omega_N$ vorgeben. Ist die zu $\Phi_1(I)$ und $\Phi_2(J)$ gehörige Lösung (x^*, y^*) nicht-positiv, dann ist sie auch Lösung von (3.4) und (3.5) und mithin eine lokale Lösung von* (AP1)*. Man beachte, dass die Lagrange-Systeme zu $\Phi_1(I)$ und $\Phi_2(J)$ immer eindeutig lösbar sind (vgl. Satz 3.15) und zumindest für $I = J = \Omega_N$ auf eine für* (AP1) *zulässige Lösung führen: nämlich $(x^*, y^*) = (0, 0)$.* □

Anmerkung 3.11. *Dieser Satz 3.10 ist entscheidend für spätere auf* (AP2-I) *und* (AP2-II) *fußende Lösungsalgorithmen für* (AP1)*, denn er besagt:*

1. *Finde eine globale Lösung von* (AP2-I)*. Ist sie zulässig für* (AP1)*, so ist sie auch eine globale Lösung von* (AP1)*.*
2. *Wenn nicht, dann suche unter allen lokalen Lösungen von* (AP2-II)*, die zulässig für* (AP1) *sind, eine mit kleinstem Zielfunktionswert. Eine solche lokale Lösung existiert sicher und ist auch eine globale Lösung von* (AP1)*.*

Anmerkung 3.12. *Hat eine Lösung von* (AP2) *positive Einträge, so impliziert dies jedoch nicht, dass man durch Null Setzen dieser Einträge eine Lösung von* (AP1) *erhält. Als einfaches Beispiel betrachte man folgendes dreidimensionales Problem:*

$$\min_{x,y,z\in\mathbb{R}} g(x,y,z) = \begin{pmatrix} x-3.96 \\ y+2.99 \\ z+4.03 \end{pmatrix}^T \begin{pmatrix} 3.04 & 0.29 & -0.42 \\ 0.29 & 1.98 & 1.23 \\ -0.42 & 1.23 & 0.99 \end{pmatrix} \begin{pmatrix} x-3.96 \\ y+2.99 \\ z+4.03 \end{pmatrix} \tag{3.6}$$

$$\textit{subject to:}$$

$$x \leq 0 \ \wedge \ y \leq 0 \ \wedge \ x \cdot y = 0$$

$$z \leq 0.$$

Minimierte man die Zielfunktion g nun einzig unter der Nebenbedingung:

$$x = 0,$$

so erhielte man als Lösung:

$$v_x = \begin{pmatrix} 0 \\ 4.13 \\ -14.55 \end{pmatrix}.$$

Zu Null Setzen der unzulässigen y-Komponente ergäbe:

$$\hat{v}_x = \begin{pmatrix} 0 \\ 0 \\ -14.55 \end{pmatrix}$$

mit $g(\hat{v}_x) = 55.70$*. Minimierte man g hingegen unter der Restriktion*

$$y = 0,$$

lautete das Ergebnis:

$$v_y = \begin{pmatrix} 3.11 \\ 0 \\ -8.10 \end{pmatrix}.$$

Setzte man wieder schlicht die unzulässige x-Komponente gleich Null, entstünde:

$$\hat{v}_y = \begin{pmatrix} 0 \\ 0 \\ -8.10 \end{pmatrix}$$

mit $g(\hat{v}_y) = 31.42$. *Die tatsächliche Lösung* v^* *von (3.6) ist aber weder* $\hat{v}_x$ *noch* $\hat{v}_y$. *Sie ergibt sich, wenn man als Nebenbedingung*

$$x = 0 \ \wedge \ y = 0$$

wählt:

$$v^* = \begin{pmatrix} 0 \\ 0 \\ -9.42 \end{pmatrix}.$$

Hierbei ist $g(v^*) = 29.70$.

Es kann gemäß Punkt 2 aus Anmerkung 3.11 u. U. nötig sein, alle lokalen Lösungen von (AP2-II), die zulässig für (AP1) sind, zu untersuchen, um eine globale Lösung für (AP1) zu ermitteln. Wir stellen uns daher zum Abschluss dieses Kapitels noch die Frage, wie viele solcher lokaler Lösungen existieren können.

Satz 3.13 (Anzahl lokaler Lösungen von (AP1))**.** (AP1) *hat höchstens* 2^N *lokale Lösungen.*

Beweis. *Es sei* (x^*, y^*) *eine lokale Lösung von* (AP1) *und*

$$I := \{i \in \Omega_N \mid x_i^* = 0\}.$$

Damit kann man diese spezielle Optimalstelle (x^*, y^*) *auch als Lösung folgender Präzisierung von* (AP1) *ansehen:*

$$\begin{gathered} (x^*, y^*) = \arg \min_{x,y \in \mathbb{R}^N} \ \tfrac{1}{2}(x-c)^T A(x-c) + \tfrac{1}{2}(y-d)^T B(y-d) \\ \textit{subject to:} \\ \forall i \in \bar{I} : x_i \leq 0, \ \forall i \in I : y_i \leq 0, \\ \forall i \in I : x_i = 0, \ \forall i \in \bar{I} : y_i = 0. \end{gathered} \tag{3.7}$$

Definiert man zusätzlich

$$J := \{i \in \Omega_N \mid y_i^* = 0\},$$

so ist (x^*, y^*) *aber auch Lösung dieser Präzisierung von* (AP1)*:*

$$\begin{gathered}(x^*, y^*) = \arg\min_{x,y\in\mathbb{R}^N} \ \tfrac{1}{2}(x-c)^T A(x-c) + \tfrac{1}{2}(y-d)^T B(y-d) \\ \textit{subject to:} \\ \forall i \in \bar{I} : x_i \leq 0, \forall i \in \bar{J} : y_i \leq 0, \\ \forall i \in I : x_i = 0, \ \forall i \in J : y_i = 0.\end{gathered} \tag{3.8}$$

Im Rahmen des Beweises zu Satz 3.10 haben wir bereits bemerkt, dass (3.8) genau (x^*, y^*) *als Lösung hat. Da die zulässige Menge* V *von (3.8) offensichtlich eine Teilmenge der zulässigen Menge* U *von (3.7) ist, ist jede Lösung von (3.7), die zudem in* V *liegt, auch eine Lösung von (3.8). Damit kann (3.7) neben* (x^*, y^*) *keine weitere lokale Lösung besitzen. Denn gäbe es eine solche weitere lokale Lösung* $(\tilde{x}, \tilde{y})$ *von (3.7), dann läge sie in* $U \setminus V$*, womit ein* $i \in I \cap J$ *exsitierte, für das*

$$\tilde{y}_i < 0$$

gälte. Für ein solches $\tilde{y}_i$ *wäre aber dann keine der Nebenbedingungen aktiv gewesen, was in diesem Fall*

$$\left.\frac{\partial f}{\partial y_i}\right|_{(x,y)=(\tilde{x},\tilde{y})} = 0$$

implizierte. Wegen der strikten Konvexität von f *folgte daraus, dass sich der Funktionswert durch Variation der* i*. Komponente von* $\tilde{y}$ *echt vergrößern ließe, d. h.*

$$\forall \varepsilon \in \mathbb{R}, \varepsilon \neq 0 : \ f(\tilde{x}, \ \tilde{y}) < f(\tilde{x}, \ \tilde{y} + \varepsilon \cdot e_i).$$

Damit kann es sich bei der ursprünglichen Lösung (x^*, y^*) *um keine lokale Minimalstelle von (3.7) handeln, da man sich durch Variation der* i*. Komponente von* y^* *weiter verbessert hätte, ohne dabei unzulässig zu werden. Dies steht im Widerspruch zur Voraussetzung. Somit hat (3.7) genau* (x^*, y^*) *als Lösung.*

Da (3.7) schließlich für jede Wahl von $I \subseteq \Omega_N$ *genau eine Lösung besitzt und*

$$|\mathcal{P}(\Omega_N)| = 2^N,$$

hat (AP1) *höchstens* 2^N *lokale Lösungen.* □

Von jetzt an wenden wir uns verstärkt der Untersuchung von (AP2-I) und (AP2-II) zu und interessieren uns dabei insbesondere für deren Zielfunktionen Φ_1 und Φ_2.

3.3 Eigenschaften von Φ_1 und Φ_2

Wir beginnen zur Einführung mit einem Satz über Eigenschaften symmetrisch positiv definiter Matrizen.

Satz 3.14. *Es sei $M \in \mathbb{R}^{N\times N}$ symmetrisch positiv definit, dann gilt:*

1. *Alle Hauptminoren von M sind positiv.*
2. *M ist invertierbar und ihre Inverse M^{-1} ist ebenfalls symmetrisch positiv definit.*
3. *Die Diagonalelemente von M sind strikt positiv.*
4. *Für alle $I \subseteq \Omega_N$ ist $M_{I,I}$ symmetrisch positiv definit.*

Beweis. *Zum Beweis von 1. sei auf [1] verwiesen.*

Zu 2.:
Nach dem Satz der Hauptachsentransformation (vgl. [1]) existiert eine orthogonale Matrix $S \in \mathbb{R}^{N\times N}$, so dass

$$M = SDS^T,$$

wobei $D \in \mathbb{R}^{N\times N}$ eine Diagonalmatrix mit den Eigenwerten $\lambda_i(M)$ $(i = 1, ..., N)$ von M auf der Hauptdiagonalen ist. Nach [1] sind symmetrische Matrizen genau dann positiv definit, wenn ihre Eigenwerte strikt positiv sind. Dies impliziert in unserem Fall die Existenz von D^{-1}. Mithin ist

$$M^{-1} = SD^{-1}S^T,$$

denn

$$MM^{-1} = SDS^TSD^{-1}S^T = \mathbb{1} = M^{-1}M.$$

Schließlich ist noch

$$M^{-T} = (SD^{-1}S^T)^T = SD^{-T}S^T = SD^{-1}S^T = M^{-1}$$

und

$$\lambda_i(M^{-1}) = \frac{1}{\lambda_i(M)} > 0,$$

was wieder gemäß [1] die positive Definitheit von M^{-1} nach sich zieht.

Zu 3.:
Wäre beispielsweise $m_{ii} \leq 0$ für ein $i \in \Omega_N$, dann folgte für den i. Einheitsvektor $e_i \in \mathbb{R}^N$:

$$\langle e_i, Me_i\rangle = m_{ii} \leq 0,$$

was ein Widerspruch zur positiven Definitheit von M wäre.

Zu 4.:
Die Symmetrie von $M_{I,I}$ folgt unmittelbar aus der Symmetrie von M und der Definition von $M_{I,I}$ (siehe Definition 3.1). Zum Nachweis der positiven Definitheit von $M_{I,I}$ sei zunächst $I = \{r\} \subseteq \Omega_N$ und $v \in \mathbb{R}^N$ ein beliebiger Vektor, den wir wie folgt zerlegen:

$$v = \hat{v} + v_r e_r,\ \hat{v} \in \mathbb{R}^N.$$

Damit ist insbesondere $\hat{v}_r = 0$. Es sei nochmals darauf hingewiesen, dass nach Definition 2.1 zwar $v_r \in \mathbb{R}$ aber $e_r \in \mathbb{R}^N$. Es folgt nun:

$$\begin{aligned}\langle v, M_{I,I} v\rangle &= \langle \hat{v} + v_r e_r, M_{I,I}(\hat{v} + v_r e_r\rangle \\ &= \langle \hat{v}, M_{I,I}\hat{v}\rangle + 2\langle \hat{v}, M_{I,I} v_r e_r\rangle + \langle v_r e_r, M_{I,I} v_r e_r\rangle \\ &= \underbrace{\langle \hat{v}, M\hat{v}\rangle}_{>0} + 2\underbrace{\langle \hat{v}, v_r \underbrace{(M_{I,I})_r}_{=e_r}\rangle}_{=0} + v_r^2 \underbrace{(M_{I,I})_{r,r}}_{=1} > 0\end{aligned}$$

Hierbei steht $(M_{I,I})_r$ für die r. Spalte von $M_{I,I}$, welche definitionsgemäß gleich e_r ist. Die Aussage des Satzes für beliebige $I \subseteq \Omega_N$ folgt induktiv. □

Damit können wir nun den Satz beweisen, auf den wir im Vorfeld schon mehrfach verwiesen hatten.

Satz 3.15. *Für jedes $I \subseteq \Omega_N$ hat $\Phi_1(I)$ genau eine Lösung x^*. Analog hat $\Phi_2(J)$ für jedes $j \subseteq \Omega_N$ genau eine Lösung y^* und es gilt:*

$$x^* = A_{I,I}^{-1} A_I c,$$

$$y^* = B_{I,I}^{-1} B_I d.$$

Beweis. *Wir zeigen die Aussage des Satzes exemplarisch für $\Phi_1(I)$, da der Beweis $\Phi_2(J)$ völlig analog verläuft. Wir betrachten zunächst die Lagrange-Bedingungen für $\Phi_1(I)$. Nach [4] sind diese hier nicht nur notwendig, sondern auch hinreichend, da die Zielfunktion zu $\Phi_1(I)$ strikt konvex und die Nebenbedingungen affin sind:*

$$A(x - c) + \sum_{i \in I} \lambda_i \cdot e_i = 0, \tag{3.9}$$

$$\forall i \in I : x_i = 0, \tag{3.10}$$

wobei $e_i \in \mathbb{R}^N$ der i. Einheitsvektor ist und $\lambda_i \in \mathbb{R}$ ($\forall i \in I$).

Einzig die Zeilen $i \notin I$ des Gleichungssystems (3.9) sind für die Lösung x von Bedeutung. Unter Vernachlässigung der Zeilen $i \in I$ von (3.9) erhält man daher:

$$\forall i \notin I : \langle A_i, x\rangle = \langle A, c\rangle, \tag{3.11}$$

$$\forall i \in I : x_i = 0. \tag{3.12}$$

Wir reformulieren die Bedingung (3.12) mit dem euklidischen Skalarprodukt:

$$\forall i \notin I : \langle A_i, x\rangle = \langle A, c\rangle, \tag{3.13}$$

$$\forall i \in I : \langle e_i, x\rangle = \langle 0, c\rangle. \tag{3.14}$$

Durch Zusammenführen beider Gleichungssysteme (3.13) und (3.14) in Matrixschreibweise ergibt sich:

$$A_{I,I}x = A_I c. \tag{3.15}$$

Nach Satz 3.14 ist $A_{I,I}$ invertierbar, womit (3.15) eindeutig lösbar ist:

$$x^* = A_{I,I}^{-1} A_I c.$$

□

Wir wollen als nächstes eine vereinfachte Darstellung von Φ_1 und Φ_2 angeben. Dazu vorweg ein kurzes Lemma.

Lemma 3.16. *Für $M \in \mathbb{R}^{N\times N}$ symmetrisch positiv definit und $I \subseteq \Omega_N$ gilt:*

$$P^T(PMP^T)^{-1}PM = M_{I,I}^{-1} M_I, \tag{3.16}$$

wobei $P := \mathbb{1}^I \in \mathbb{R}^{(N-|I|)\times N}$.

Beweis. *Es sei zunächst $I = \{r\} \subseteq \Omega_N$ und M der Form:*

$$M = \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix}$$

mit

$$m_{r-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{r-1,r} \end{pmatrix} \in \mathbb{R}^{l_1},\ m_{r+1} = \begin{pmatrix} m_{r+1,r} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{l_2},$$

$$M_1 \in \mathbb{R}^{l_1\times l_1},\ M_2 \in \mathbb{R}^{l_2\times l_2},$$

$$l_1, l_2 \in \mathbb{N},\ l_1 + l_2 + 1 = N.$$

Es gilt:

$$PM = \begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & 0 & \mathbb{1} \end{pmatrix} \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix} = \begin{pmatrix} M_1 & m_{r-1} & \\ & m_{r+1} & M_2 \end{pmatrix}$$

und gleichermaßen

$$PMP^T = \begin{pmatrix} M_1 & m_{r-1} & \\ & m_{r+1} & M_2 \end{pmatrix} \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \\ 0 & \mathbb{1} \end{pmatrix} = \begin{pmatrix} M_1 & \\ & M_2 \end{pmatrix}.$$

Somit ist:

$$P^T(PMP^T)^{-1} = \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \\ 0 & \mathbb{1} \end{pmatrix} \begin{pmatrix} M_1^{-1} & \\ & M_2^{-1} \end{pmatrix} = \begin{pmatrix} M_1^{-1} & \\ 0 & 0 \\ & M_2^{-1} \end{pmatrix}.$$

Für die linke Seite von (3.16) folgt schließlich:

$$P^T(PMP^T)^{-1}PM = \begin{pmatrix} \mathbb{1} & M_1^{-1}m_{r-1} & \\ 0 & 0 & 0 \\ & M_2^{-1}m_{r+1} & \mathbb{1} \end{pmatrix},$$

was gleich der rechten Seite von (3.16) ist, denn:

$$M_{I,I}^{-1}M_I = \begin{pmatrix} M_1^{-1} & 0 & \\ 0 & 1 & 0 \\ & 0 & M_2^{-1} \end{pmatrix} \begin{pmatrix} M_1 & m_{r-1} & \\ 0 & 0 & 0 \\ & m_{r+1} & M_2 \end{pmatrix} = \begin{pmatrix} \mathbb{1} & M_1^{-1}m_{r-1} & \\ 0 & 0 & 0 \\ & M_2^{-1}m_{r+1} & \mathbb{1} \end{pmatrix}.$$

Die Aussage des Lemmas für beliebige $I \subseteq \Omega_N$ folgt nun induktiv. □

Satz 3.17. *Die Abbildungen Φ_1 und Φ_2 aus Definition 3.6 haben die Form*

$$\Phi_1(I) = \frac{1}{2}c^T A(\mathbb{1} - A_{I,I}^{-1}A_I)c,\ I \subseteq \Omega_N,$$

$$\Phi_2(J) = \frac{1}{2}d^T B(\mathbb{1} - B_{J,J}^{-1}B_J)d,\ J \subseteq \Omega_N.$$

Beweis. *Wir betrachten x^* aus Satz 3.15 und wenden Lemma 3.16 an, wobei wir wieder statt $\mathbb{1}^I$ kurz P schreiben:*

$$x^* = A_{I,I}^{-1} A_I c = P^T (PAP^T)^{-1} PAc.$$

Einsetzen von x^* in die Zielfunktion ergibt:

$$\Phi_1(I) = f_1(x^*) = \frac{1}{2}(P^T(PAP^T)^{-1}PAc - c)^T A(P^T(PAP^T)^{-1}PAc - c).$$

Ausklammern von $-c$ und einige weitere Umformungen führen direkt auf die Behauptung:

$$\begin{aligned}
\Phi_1(I) &= \frac{1}{2}c^T(\mathbb{1} - P^T(PAP^T)^{-1}PA)^T A(\mathbb{1} - P^T(PAP^T)^{-1}PA)c \\
&= \frac{1}{2}c^T(A - AP^T(PAP^T)^{-1}PA)(\mathbb{1} - P^T(PAP^T)^{-1}PA)c \\
&= \frac{1}{2}c^T[A - AP^T(PAP^T)^{-1}PA + AP^T(PAP^T)^{-1}PA \\
&\quad + AP^T \underbrace{(PAP^T)^{-1}PAP^T}_{=\mathbb{1}}(PAP^T)^{-1}P^TA]c \\
&= \frac{1}{2}c^T[A - 2AP^T(PAP^T)^{-1}PA + AP^T(PAP^T)^{-1}PA]c \\
&= \frac{1}{2}c^T[A - AP^T(PAP^T)^{-1}PA]c = \frac{1}{2}c^TA[\mathbb{1} - P^T(PAP^T)^{-1}PA]c \\
&= \frac{1}{2}c^TA(\mathbb{1} - A_{I,I}^{-1}A_I)c.
\end{aligned}$$

□

Im Falle, dass I, J einelementige Teilmengen von Ω_N sind, lässt sich die soeben bewiesene Form von Φ_1 und Φ_2 sogar noch weiter präzisieren.

Satz 3.18. *Im Besonderen gilt für $r \in \Omega_N$:*

$$\Phi_1(\{r\}) = \frac{1}{2} \cdot \frac{c_r^2}{(A^{-1})_{r,r}},$$

$$\Phi_2(\{r\}) = \frac{1}{2} \cdot \frac{d_r^2}{(B^{-1})_{r,r}}.$$

Beweis. *Wir betrachten das Problem*

$$\begin{aligned}
x^* = \arg\min_{x \in \mathbb{R}^N} \ & \tfrac{1}{2}(x-c)^T A(x-c) \\
s.\ t.\ & x_r = 0,
\end{aligned} \tag{3.17}$$

Das zu (3.17) gehörige Lagrange-System ist wie schon im Beweis zu Satz 3.15 gezeigt eindeutig lösbar und lautet:

$$A(x-c)+\lambda\cdot e_r=0, \tag{3.18}$$

$$x_r=0 \tag{3.19}$$

mit $\lambda\in\mathbb{R}$. Auflösen von (3.18) nach x ergibt:

$$x=c-\lambda\cdot A^{-1}e_r.$$

Nun muss λ nur noch so gewählt werden, dass die Nebenbedingung $x_r=0$ erfüllt ist. Dies ist offensichtlich für

$$\lambda=\frac{c_r}{(A^{-1})_{r,r}}$$

der Fall. Damit ergibt sich für die Lösung von (3.17):

$$x^*=c-\frac{c_r}{(A^{-1})_{r,r}}\cdot A^{-1}e_r.$$

Mithin folgt die Aussage des Satzes:

$$\begin{aligned}\Phi_1(\{r\})=f_1(x^*)&=\frac{1}{2}(c-\frac{c_r}{(A^{-1})_{r,r}}\cdot A^{-1}e_r-c)^TA(c-\frac{c_r}{(A^{-1})_{r,r}}\cdot A^{-1}e_r-c)\\&=\frac{1}{2}\cdot\frac{c_r^2}{(A^{-1})_{r,r}^2}\cdot e_r^TA^{-1}\underbrace{AA^{-1}}_{=\mathbb{1}}e_r=\frac{1}{2}\cdot\frac{c_r^2}{(A^{-1})_{r,r}^2}\cdot(A^{-1})_{r,r}\\&=\frac{1}{2}\cdot\frac{c_r^2}{(A^{-1})_{r,r}}.\end{aligned}$$

□

Anmerkung 3.19. *Das Ergebnis des vorigen Satzes lässt sich auch geometrisch interpretieren. Bei den Niveaumengen von quadratischen Formen handelt es sich um mehrdimensionale Ellipsoiden. Die k-Niveaumenge von f ist durch folgende Gleichung beschrieben:*

$$H:\ \frac{1}{2}(x-c)^TA(x-c)+\frac{1}{2}(y-d)^TB(y-d)=k.$$

Der Wert

$$\Delta_{x_r}:=2\cdot\sqrt{2}\cdot\sqrt{k}\cdot\sqrt{(A^{-1})_{r,r}}$$

entspricht dabei der Ausdehnung des Ellipsoiden H in Normalenrichtung zur Koordinatenebene

$$E_{x_r}:\ x_r=0.$$

Der Optimalwert k^* des Problems [1]

$$k^*=\min_{x,y\in\mathbb{R}^N}\ \tfrac{1}{2}(x-c)^TA(x-c)\ \left[+\tfrac{1}{2}(y-d)^TB(y-d)\right]$$

$$s.\ t.\ x_r=0$$

[1] Der Summand in eckigen Klammern hat hier keinen Einfluss auf den Optimalwert, da für y keine Restriktion existiert und mithin $y=d$ gewählt werden kann.

ist geometrisch also nichts anderes als das Quadrat des Faktors, um den man den Ellipsoiden H strecken muss, damit er genau die Koordinatenebene E_{x_r} berührt. Zur Berührung kommt es offensichtlich genau dann, wenn die halbe Ausdehnung von H gleich dem Abstand $|c_r|$ des Scheitels von der Ebene E_{x_r} ist (vgl. Abbildung 3.4):

$$\begin{aligned} \frac{1}{2} \cdot \Delta_{x_r} &\stackrel{!}{=} |c_r| \\ \iff \sqrt{2} \cdot \sqrt{k} \cdot \sqrt{(A^{-1})_{r,r}} &= |c_r| \\ \iff k &= \frac{1}{2} \cdot \frac{c_r^2}{(A^{-1})_{r,r}}. \end{aligned}$$

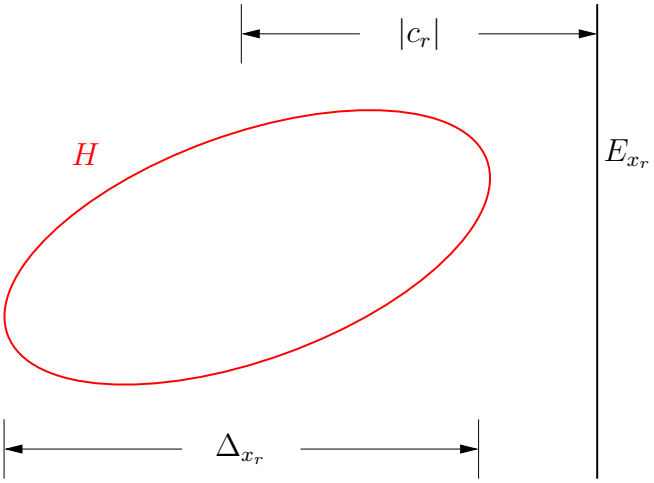


Abbildung 3.4: Ausdehnung eines Ellipsoiden

Nun, da wir die wichtigsten Werkzeuge hergeleitet haben, werden wir in den beiden folgenden Kapiteln zwei Lösungsstrategien für (AP1) zusammenstellen. Ziel wird es hierbei immer auch sein, den Aufwand der Verfahren nach Möglichkeit gering zu halten. Es lohnt sich daher, bevor man mit der Entwicklung eines allgemeinen Lösungsalgorithmus beginnt, einen Blick auf die tatsächliche Struktur der Matrizen A und B zu werfen. Man stellt dabei fest, dass diese in der Praxis stark diagonaldominant sind und die Einträge mit steigender Nebendiagonale betragsmäßig stets rasch abfallen. Diese aus praktischen Messungen resultierende Beobachtung motiviert unser weiteres Vorgehen:
So beschreiben wir in Kapitel 3.4 einen sehr schnellen Algorithmus, welcher exakt ist, falls A und B Diagonalmatrizen sind, und in der Praxis eine zumindest sehr gute lokale Lösung ermittelt.
In einer etwas besseren Approximation der tatsächlichen Gestalt von A und B gehen wir in Kapitel 3.5 von tridiagonalen Matrizen aus und entwickeln darauf basierend

eine weitere Lösungsstrategie, welche zwar deutlich aufwändiger ist als erstere, im Gegenzug aber, auf die tatsächlichen Matrizen der Praxis angewendet, einer globalen Lösung noch näher kommt.

3.4 Der Diagonal-Fall

In diesem Kapitel seien A und B also Diagonalmatrizen. Unter diesen Umständen lässt sich die Disjunktion in der Nebenbedingung von (AP2) leicht mittels des folgenden Entscheidungskriteriums auflösen.

Satz 3.20. *Eine Teilmenge $I_0 \subseteq \Omega_N$ ist genau dann Lösung von* (AP2-I)*, wenn für jedes $i \in I_0$*

$$\Phi_1(\{i\}) \leq \Phi_2(\{i\}) \tag{3.20}$$

und für jedes $j \in \bar{I}_0$

$$\Phi_1(\{j\}) \geq \Phi_2(\{j\}) \tag{3.21}$$

gilt.

Beweis. *Sind A, B in Diagonalform, so zerfällt* (AP2-I) *in N unabhängige Teilprobleme:*

$$\begin{gathered} \min_{x_i, y_i \in \mathbb{R}} \ \tfrac{1}{2} a_{ii}(x_i - c_i)^2 + \tfrac{1}{2} b_{ii}(y_i - d_i)^2 \\ s.\ t.\ x_i = 0 \ \vee \ y_i = 0, \end{gathered} \tag{3.22}$$

wobei $i \in \Omega_N$. Der Optimalwert von (3.22) ist offensichtlich gleich

$$\min\left(\frac{1}{2} a_{ii} c_i^2,\ \frac{1}{2} b_{ii} d_i^2\right).$$

Da im Diagonalfall

$$a_{ii} = \frac{1}{(A^{-1})_{ii}} \ \textit{und}\ b_{ii} = \frac{1}{(B^{-1})_{ii}},$$

ist mit Satz 3.18

$$\min\left(\frac{1}{2} a_{ii} c_i^2,\ \frac{1}{2} b_{ii} d_i^2\right) = \min\left(\Phi_1(\{i\}), \Phi_2(\{i\})\right).$$

□

Anmerkung 3.21. *Ist für alle $i \in \Omega_N$*

$$\Phi_1(\{i\}) \neq \Phi_2(\{i\}),$$

werden aus den Bedingungen im vorigen Satz 3.20 strikte Ungleichungen, womit es genau eine Menge $I_0 \subseteq \Omega_N$ gibt, die Lösung von (AP2-I) *ist.*

Somit lassen sich globale Lösungen für (AP2-I) konstruieren. Sind nun alle Komponenten einer solchen Lösung nicht-positiv, ist sie auch eine globale Lösung für (AP1). Andernfalls tritt Punkt 2 aus Anmerkung 3.11 in Kraft, welchem wir mit folgendem Satz genügen:

Satz 3.22. *Es sei* $(\hat{x}, \hat{y})$ *eine globale Lösung von* (AP2) *und* $U \subseteq \Omega_N$ *die Menge aller für* (AP1) *unzulässigen Komponenten, d. h.*

$$U := \{i \in \Omega_N \mid \hat{x}_i > 0 \ \vee \ \hat{y}_i > 0\},$$

dann lässt sich eine globale Lösung von (AP1) *konstruieren, indem man*

$$\hat{x}_i = \begin{cases} 0, & \text{falls } c_i > 0 \\ c_i, & \text{sonst} \end{cases}$$

und

$$\hat{y}_i = \begin{cases} 0, & \text{falls } d_i > 0 \\ d_i, & \text{sonst} \end{cases}$$

setzt.

Beweis. *Wir betrachten das Problem* (AP2-II)*, welches aufgrund der Diagonalgestalt von* A *und* B *in* N *unabhängige Teilprobleme zerfällt:*

$$\begin{aligned} &\min_{x_i, y_i \in \mathbb{R}} \ \tfrac{1}{2} a_{ii}(x_i - c_i)^2 + \tfrac{1}{2} b_{ii}(y_i - d_i)^2 \\ &s.\ t.\ x_i = 0 \ \vee \ y_i = 0 \ \vee \ x_i = y_i = 0. \end{aligned} \tag{3.23}$$

Wir konzentrieren uns hierbei nur auf $i \in U$*. Nach Anmerkung 3.11 ist also unter den lokalen Lösungen von* (AP2-II)*, die zulässig für* (AP1) *sind, eine optimale zu suchen. Die lokalen Lösungen von (3.23) sind:*

$$v_x := \begin{pmatrix} 0 \\ \\ d_i \end{pmatrix}, \ v_y := \begin{pmatrix} c_i \\ \\ 0 \end{pmatrix} \ \text{sowie } v_{xy} := \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}.$$

Es gibt nun folgende Fälle zu unterscheiden:
- Ist v_x *zulässig (d. h.* $d_i \leq 0$*), dann ist* v_y *unzulässig, da nach Voraussetzung* $i \in U$*. In diesem Fall ist* v_x *optimal für* (AP1)*, da der Zielfunktionswert zu* v_{xy} *nach Lemma 3.9 sicher nicht kleiner ist.*
- Ist v_y *zulässig (d. h.* $c_i \leq 0$*), dann ist* v_x *unzulässig, da nach Voraussetzung* $i \in U$*. In diesem Fall ist* v_y *optimal für* (AP1)*, da der Zielfunktionswert zu* v_{xy} *nach Lemma 3.9 sicher nicht kleiner ist.*
- Sind sowohl v_x *als auch* v_y *unzulässig, so ist* v_{xy} *optimal für* (AP1)*.* □

Beide Sätze zusammen ergeben unmittelbar folgenden Algorithmus, um eine Lösung (x, y) für (AP1) im Diagonalfall zu ermitteln:

BEGIN AP1DIAG
  for $i = 1$ to N **do**
    if $c_i \leq 0$ **and** $d_i \leq 0$ **then**
      if $\Phi_1(\{i\}) \leq \Phi_2(\{i\})$ **then**
        $I_0 \leftarrow i$;
      else
        $J_0 \leftarrow i$;
      end if
    else
      if $c_i > 0$ **then**
        $I_0 \leftarrow i$;
      end if
      if $d_i > 0$ **then**
        $J_0 \leftarrow i$;
      end if
    end if
  end for
  Berechne Lösung x zu $\Phi_1(I_0)$;
  Berechne Lösung y zu $\Phi_2(J_0)$;

END AP1DIAG

3.5 Der Tridiagonal-Fall

In diesem Kapitel seien A und B Tridiagonalmatrizen. Wir möchten, ähnlich wie im vorigen Kapitel, einen Algorithmus zur Lösung von (AP1) angeben und werden dazu insbesondere ein hinreichendes Optimalitätskriterium für (AP2-I) ermitteln.

3.5.1 Herleitung eines Optimalitätskriteriums

Ein erster wichtiger Schritt ist, eine geschlossene Darstellung für $\Phi_1(I)$ und $\Phi_2(J)$ für beliebige Teilmengen $I, J \subseteq \Omega_N$ zu finden. Dazu bedarf es jedoch erst noch einiger Vorbereitungen.

Es seien von nun an $M \in \mathbb{R}^{N\times N}$, $\tilde{M} \in \mathbb{R}^{l\times l}$, $M_1 \in \mathbb{R}^{l_1\times l_1}$, $M_2 \in \mathbb{R}^{l_2\times l_2}$, $M_3 \in \mathbb{R}^{l_3\times l_3}$ beliebige symmetrisch positiv definite Tridiagonalmatrizen mit $l, l_1, l_2, l_3 \in \mathbb{N}_0$. Ferner seien M und $\tilde{M}$, sofern nichts Gegenteiliges erwähnt ist, von nachfolgender Gestalt.

$$\tilde{M} = \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix},$$

$$M = \begin{pmatrix} \tilde{M} & m_{s-1} & \\ m_{s-1}^T & m_{s,s} & m_{s+1}^T \\ & m_{s+1} & M_3 \end{pmatrix}$$

mit

$$m_{r-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{r-1,r} \end{pmatrix} \in \mathbb{R}^{l_1},\ m_{r+1} = \begin{pmatrix} m_{r,r+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{l_2},$$

$$m_{s-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{s-1,s} \end{pmatrix} \in \mathbb{R}^{l},\ \ m_{s+1} = \begin{pmatrix} m_{s,s+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{l_3}.$$

und

$$r, s \in \mathbb{N}_0,\ r \leq s.$$

Lemma 3.23. *Es seien U, V zwei invertierbare Matrizen und*

$$W := \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix},$$

dann ist auch W invertierbar und es gilt:

$$W^{-1} = \begin{pmatrix} U^{-1} & 0 \\ 0 & V^{-1} \end{pmatrix}.$$

Das nachfolgende Lemma werden wir im späteren Verlauf heranziehen, um in Φ_1 bzw. Φ_2 auftretende Terme auf Diagonalelemente von A^{-1} bzw. B^{-1} zurückzuführen.

Lemma 3.24. *Es gilt:*

$$\frac{1}{(\tilde{M}^{-1})_{r,r}} = \begin{cases} m_{r,r} - m_{r+1}^T M_2^{-1} m_{r+1}, & \text{falls } r = 1 \\ m_{r,r} - m_{r-1}^T M_1^{-1} m_{r-1} - m_{r+1}^T M_2^{-1} m_{r+1}, & \text{falls } 1 < r < N \\ m_{r,r} - m_{r-1}^T M_1^{-1} m_{r-1}, & \text{falls } r = N \end{cases} .$$

Beweis. *Durch Gleichsetzen der Aussagen der Sätze 3.17 und 3.18 erhalten wir:*

$$\frac{1}{2} \cdot w^T \tilde{M} (\mathbb{1} - \tilde{M}_{\{r\},\{r\}}^{-1} \tilde{M}_{\{r\}}) w = \frac{1}{2} \cdot \frac{w_i^2}{(\tilde{M}^{-1})_{r,r}} \tag{3.24}$$

für beliebiges $w \in \mathbb{R}^N$. Wir untersuchen die linke Seite der Gleichung nun etwas genauer:

$$\tilde{M}_{\{r\},\{r\}}^{-1} \tilde{M}_{\{r\}} = \begin{pmatrix} M_1^{-1} & 0 & \\ 0 & 1 & 0 \\ & 0 & M_2^{-1} \end{pmatrix} \begin{pmatrix} M_1 & m_{r-1} & \\ 0 & 0 & 0 \\ & m_{r+1} & M_2 \end{pmatrix} = \begin{pmatrix} \mathbb{1} & M_1^{-1} m_{r-1} & \\ 0 & 0 & 0 \\ & M_2^{-1} m_{r+1} & \mathbb{1} \end{pmatrix} .$$

Damit folgt:

$$\tilde{M} \tilde{M}_{\{r\},\{r\}}^{-1} \tilde{M}_{\{r\}} = \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix} \begin{pmatrix} \mathbb{1} & M_1^{-1} m_{r-1} & \\ 0 & 0 & 0 \\ & M_2^{-1} m_{r+1} & \mathbb{1} \end{pmatrix}$$

$$= \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & (m_{r-1}^T M_1^{-1} m_{r-1} + m_{r+1}^T M_2^{-1} m_{r+1}) & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix} .$$

Und somit:

$$\tilde{M}(\mathbb{1} - \tilde{M}_{\{r\},\{r\}}^{-1} \tilde{M}_{\{r\}}) = \tilde{M} - \tilde{M} \tilde{M}_{\{r\},\{r\}}^{-1} \tilde{M}_{\{r\}}$$

$$= \begin{pmatrix} 0 & 0 & \\ 0 & m_{r,r} - (m_{r-1}^T M_1^{-1} m_{r-1} + m_{r+1}^T M_2^{-1} m_{r+1}) & 0 \\ & 0 & 0 \end{pmatrix} .$$

Setzen wir nun dieses Ergebnis in (3.24) ein:

$$\frac{1}{2} \cdot w^T \tilde{M}(\mathbb{1} - \tilde{M}^{-1}_{\{r\},\{r\}} \tilde{M}_{\{r\}})w = \frac{1}{2} \cdot \frac{w_i^2}{(\tilde{M}^{-1})_{r,r}}$$

$$\iff w_i^2(m_{r,r} - m_{r-1}^T M_1^{-1} m_{r-1} - m_{r+1}^T M_2^{-1} m_{r+1}) = \frac{w_i^2}{(\tilde{M}^{-1})_{r,r}}$$

$$\iff m_{r,r} - m_{r-1}^T M_1^{-1} m_{r-1} - m_{r+1}^T M_2^{-1} m_{r+1} = \frac{1}{(\tilde{M}^{-1})_{r,r}}.$$

Die Aussage des Satzes für $r = 1$ folgt aus gleichem Räsonnement für $l_1 = 0$. Man setzt dazu also $M_1 = 0,\ m_{r-1} = 0$. Für den Beweis des Falles $r = N$ ist analog $l_2 = 0,\ M_2 = 0,\ m_{r+1} = 0$ zu setzen. □

Hiermit können wir nun eine übersichtliche Darstellung der in Satz 3.17 vorkommenden Matrix herleiten.

Satz 3.25. *Es sei $I \subseteq \Omega_N$ und $U \in \mathbb{R}^{N \times N}$ wie folgt definiert:*

$$U := M - M M^{-1}_{I,I} M_I.$$

Dann gilt:

$$u_{i,j} = \begin{cases} 1/(M^{-1}_{\{v_I(i),n_I(i)\},\{v_I(i),n_I(i)\}})_{i,i}, & \text{falls } i = j \ \wedge\ i,j \in I \\ m_{i,j}, & \text{falls } |i-j| = 1 \ \wedge\ i,j \in I \\ -m_{i+1,i} m_{j-1,j} (M^{-1}_{\{i,j\},\{i,j\}})_{i+1,j-1}, & \text{falls } |i-j| > 1 \ \wedge\ j = n_I(i) \ \wedge\ i,j \in I \\ 0, & \text{sonst} \end{cases},$$

wobei $u_{i,j} = u_{j,i}$ und o. B. d. A. $i \leq j$.

Beweis. *Für die Menge I gelte zunächst: $I \subseteq \{1, ..., l\}$ mit Kardinalität $|I| = k \leq l$. $r \leq l$ sei ihr größtes Element. Wir beweisen die Aussage mittels vollständiger Induktion über k.*

Induktionsanfang: $k = 1$
Für eine einelementige Teilmenge $I = \{r\} \subseteq \Omega_N$ haben wir bereits im Rahmen des Beweises von Lemma 3.24 gezeigt, dass für $U = M - M M^{-1}_{I,I} M_I$ nur der Eintrag $u_{r,r} = 1/(M^{-1})_{r,r}$ ungleich Null ist. Man beachte, dass $v_I(r) = 0 = n_I(r)$ und damit

$$u_{r,r} = 1/(M^{-1})_{r,r} = 1/(M^{-1}_{\{0,0\},\{0,0\}})_{r,r} = 1/(M^{-1}_{\{v_I(r),n_I(r)\},\{v_I(r),n_I(r)\}})_{r,r}$$

gilt.

Induktionsschritt: $k \rightsquigarrow k+1$
Im Rahmen der Induktionsannahme gehen wir davon aus, dass $\tilde{U} = \tilde{M} - \tilde{M}\tilde{M}_{I,I}^{-1}\tilde{M}_I$ die Aussage des Satzes erfülle. Wir betrachten nun $U = M - MM_{I\cup\{s\},I\cup\{s\}}^{-1}M_{I\cup\{s\}}$ mit $r < s$.

1. Fall: $r < s-1$

$$M_{I\cup\{s\},I\cup\{s\}}^{-1}M_{I\cup\{s\}} = \begin{pmatrix} \tilde{M}_{I,I}^{-1} & 0 & \\ 0 & 1 & 0 \\ & 0 & M_3^{-1} \end{pmatrix} \begin{pmatrix} \tilde{M}_I & m_{s-1} & \\ 0 & 0 & 0 \\ & m_{s+1} & M_3 \end{pmatrix}$$

$$= \begin{pmatrix} \tilde{M}_{I,I}^{-1}\tilde{M}_I & \tilde{M}_{I,I}^{-1}m_{s-1} & \\ 0 & 0 & 0 \\ & M_3^{-1}m_{s+1} & \mathbb{1} \end{pmatrix}.$$

Damit folgt:

$$MM_{I\cup\{s\},I\cup\{s\}}^{-1}M_{I\cup\{s\}} = \begin{pmatrix} \tilde{M} & m_{s-1} & \\ m_{s-1}^T & m_{s,s} & m_{s+1}^T \\ & m_{s+1} & M_3 \end{pmatrix} \begin{pmatrix} \tilde{M}_{I,I}^{-1}\tilde{M}_I & \tilde{M}_{I,I}^{-1}m_{s-1} & \\ 0 & 0 & 0 \\ & M_3^{-1}m_{s+1} & \mathbb{1} \end{pmatrix}$$

$$= \begin{pmatrix} \tilde{M}\tilde{M}_{I,I}^{-1}\tilde{M}_I & \tilde{M}\tilde{M}_{I,I}^{-1}m_{s-1} & \\ m_{s-1}^T\tilde{M}_{I,I}^{-1}\tilde{M}_I & (m_{s-1}^T\tilde{M}_{I,I}^{-1}m_{s-1} + m_{s+1}^TM_3^{-1}m_{s+1}) & m_{s+1}^T \\ & m_{s+1} & M_3 \end{pmatrix}.$$

Unter Verwendung der Induktionsannahme ergibt sich:

$$U = M - MM_{I\cup\{s\},I\cup\{s\}}^{-1}M_{I\cup\{s\}}$$

$$= \begin{pmatrix} \tilde{U} & m_{s-1} - \tilde{M}\tilde{M}_{I,I}^{-1}m_{s-1} & \\ m_{s-1}^T - m_{s-1}^T\tilde{M}_{I,I}^{-1}\tilde{M}_I & (m_{s,s} - m_{s-1}^T\tilde{M}_{I,I}^{-1}m_{s-1} - m_{s+1}^TM_3^{-1}m_{s+1}) & 0 \\ & 0 & 0 \end{pmatrix}.$$

Das Diagonalelement $u_{s,s}$ formen wir mit Hilfe von Lemma 3.24 um:

$$u_{s,s} = m_{s,s} - m_{s-1}^T\tilde{M}_{I,I}^{-1}m_{s-1} - m_{s+1}^TM_3^{-1}m_{s+1}$$

$$= \frac{1}{(M^{-1}_{\{r\},\{r\}})_{s,s}}.$$

Nun zu Untersuchung der Nebendiagonalelemente in der s. Zeile von U. Die Verwendung von $*$ soll andeuten, dass es sich um einen beliebigen Matrixeintrag handelt, dessen genauer Wert aber im weiteren Verlauf der Rechnung keine Rolle spielt.

$$m^T_{s-1} - m^T_{s-1}\tilde{M}^{-1}_{I,I}\tilde{M}_I =$$

$$= m^T_{s-1} - m^T_{s-1}\begin{pmatrix} (M_1)^{-1}_{I\setminus\{r\},I\setminus\{r\}} & 0 & \\ 0 & 1 & 0 \\ & 0 & M_2^{-1} \end{pmatrix}\begin{pmatrix} (M_1)_{I\setminus\{r\}} & m_{r-1} & \\ 0 & 0 & 0 \\ & m_{r+1} & M_2 \end{pmatrix}$$

$$= m^T_{s-1} - (0,...,0,m_{s-1,s})\begin{pmatrix} * & * & \\ * & * & * \\ & M_2^{-1}m_{r+1} & \mathbb{1} \end{pmatrix}$$

$$= (0,...,0,m_{s-1,s}) - (0,...,0,\underbrace{m_{s-1,s}(M_2^{-1})_{s-1,r+1}m_{r+1,r}}_{r.\ Stelle},0,...,0,m_{s-1,s})$$

$$= (0,...,0,\underbrace{m_{s-1,s}(M_2^{-1})_{s-1,r+1}m_{r+1,r}}_{r.\ Stelle},0,...,0).$$

Damit ist $u_{s,r}$ das einzige Nebendiagonalelement in der s. Zeile von U, das nicht Null ist:

$$u_{s,r} = m_{s-1,s}\cdot(M_2^{-1})_{s-1,r+1}\cdot m_{r+1,r}.$$

Nun zur Untersuchung der Nebendiagonalelemente in der s. Spalte von U:

$$m_{s-1} - \tilde{M}\tilde{M}^{-1}_{I,I}m_{s-1}.$$

Hier erhalten wir das analoge transponierte Ergebnis zu dem der s. Zeile, da

$$\tilde{M}\tilde{M}^{-1}_{I,I}$$

die transponierte Matrix von

$$\tilde{M}^{-1}_{I,I}\tilde{M}_I$$

ist, wenn man die $*$-Einträge nicht berücksichtigt:

$$\tilde{M}\tilde{M}^{-1}_{I,I} = \begin{pmatrix} M_1 & m_{r-1} & \\ m^T_{r-1} & m_{r,r} & m^T_{r+1} \\ & m_{r+1} & M_2 \end{pmatrix}\begin{pmatrix} (M_1)^{-1}_{I\setminus\{r\},I\setminus\{r\}} & 0 & \\ 0 & 1 & 0 \\ & 0 & M_2^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} * & * & \\ * & * & m_{r+1}^T M_2^{-1} \\ & * & \mathbb{1} \end{pmatrix}.$$

Mithin ist

$$u_{s,r} = u_{r,s}.$$

2. Fall: $r = s-1$
Die Rechnung in diesem Fall verläuft völlig analog zu der des 1. Falls mit lediglich einer Modifikation: Die Matrix $M_{I\cup\{s\}}$ vereinfacht sich zu:

$$M_{I\cup\{s\}} = \begin{pmatrix} \tilde{M}_I & 0 & \\ 0 & 0 & 0 \\ & m_{s+1} & M_3 \end{pmatrix}.$$

Für die Berechnung der Nebendiagonalelemente von U folgt damit:

$$m_{s-1}^T - m_{s-1}^T \tilde{M}_{I,I}^{-1} \tilde{M}_I = (0,...,0,m_{s-1,s}) - (0,...,0,\underbrace{m_{s-1,s} \cdot (M_2^{-1})_{s-1,r+1} \cdot m_{r+1,r}}_{=0})$$

$$= (0,...,0,m_{s-1,s}).$$

Somit ist:

$$u_{r,s} = u_{s,r} = m_{s-1,s}.$$

Abschließend bleibt noch zu zeigen, dass auch die restlichen Einträge von U die Bedingung des Satzes nach dem Induktionsschritt erfüllen:
Für $u_{r,r}$ gilt nach Induktionsannahme:

$$u_{r,r} = \frac{1}{(\tilde{M}^{-1}_{\{v_I(r)\},\{v_I(r)\}})_{r,r}}.$$

Wegen der blockweisen Invertierbarkeit (Satz 3.23) haben wir:

$$(\tilde{M}^{-1}_{\{v_I(r)\},\{v_I(r)\}})_{r,r} = (M^{-1}_{\{v_I(r),s\},\{v_I(r),s\}})_{r,r}$$

und

$$v_I(r) = v_{I\cup\{s\}}(r)$$

sowie

$$n_{I\cup\{s\}}(r) = s.$$

Gleichermaßen gilt für alle anderen Einträge von $\tilde{U}$:

$$\forall i \in I, i < r : (\tilde{M}^{-1}_{\{v_I(i),n_I(i)\},\{v_I(i),n_I(i)\}})_{i,i} = (M^{-1}_{\{v_I(i),n_I(i)\},\{v_I(i),n_I(i)\}})_{i,i}.$$

Ferner:

$$\forall i,j \in I, j = n_I(i) : (\tilde{M}^{-1}_{\{i,j\},\{i,j\}})_{i+1,j-1} = (M^{-1}_{\{i,j\},\{i,j\}})_{i+1,j-1}$$

und

$$\forall i,j \in I : (M)_{i,j} = (\tilde{M})_{i,j}.$$

Schließlich ist:

$$\forall i \in I, i < r : v_I(i) = v_{I\cup\{s\}}(i) \text{ und } n_I(i) = n_{I\cup\{s\}}(i).$$

□

Dieses Ergebnis erlaubt es uns direkt, eine geschlossene Form von Φ_1 und Φ_2 anzugeben.

Korollar 3.26. *Es gilt:*

$$\Phi_1(I) = \sum_{i\in I} \frac{1}{2} \cdot c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)},$$

$$\Phi_2(J) = \sum_{j\in J} \frac{1}{2} \cdot d_j^2 \cdot \beta_{v_J(j),j,n_J(j)} + d_j \cdot d_{n_J(j)} \cdot \delta_{j,n_J(j)}$$

mit

$$\alpha_{r,s,t} := \begin{cases} 0, & \text{falls } s \notin \Omega_N \\ 1/(A^{-1}_{\{r,t\},\{r,t\}})_{s,s}, & \text{falls } s \in \Omega_N \end{cases}$$

$$\alpha_s := \begin{cases} 0, & \text{falls } s \notin \Omega_N \\ 1/(A^{-1})_{s,s}, & \text{falls } s \in \Omega_N \end{cases}$$

$$\gamma_{r,s} := \begin{cases} 0, & \text{falls } r = s \ \vee \ (r,s) \notin (\Omega_N \times \Omega_N) \\ a_{r,s} & \text{falls } |r-s| = 1 \ \wedge \ (r,s) \in (\Omega_N \times \Omega_N) \\ -a_{r+1,r} \cdot a_{s-1,s} \cdot (A^{-1}_{\{r,s\},\{r,s\}})_{r+1,s-1}, & \text{falls } |r-s| > 1 \ \wedge \ (r,s) \in (\Omega_N \times \Omega_N) \end{cases}$$

$$\beta_{r,s,t} := \begin{cases} 0, & \text{falls } s \notin \Omega_N \\ 1/(B^{-1}_{\{r,t\},\{r,t\}})_{s,s}, & \text{falls } s \in \Omega_N \end{cases}$$

$$\beta_s := \begin{cases} 0, & \text{falls } s \notin \Omega_N \\ 1/(B^{-1})_{s,s}, & \text{falls } s \in \Omega_N \end{cases}$$

$$\delta_{r,s} := \begin{cases} 0, & \textit{falls } r = s \ \vee \ (r,s) \notin (\Omega_N \times \Omega_N) \\ b_{r,s} & \textit{falls } |r-s| = 1 \ \wedge \ (r,s) \in (\Omega_N \times \Omega_N) \\ -b_{r+1,r} \cdot b_{s-1,s} \cdot (B^{-1}_{\{r,s\},\{r,s\}})_{r+1,s-1}, & \textit{falls } |r-s| > 1 \ \wedge \ (r,s) \in (\Omega_N \times \Omega_N) \end{cases}$$

Beweis. *Die Aussage folgt unter Verwendung der vorangegangenen Sätze.*

$$\begin{aligned} \Phi_1(I) &\overset{Satz\ 3.17}{=} \frac{1}{2} \cdot c^T \underbrace{(A - AA^{-1}_{I,I}A_I)}_{:=U} c \\ &\overset{Satz\ 3.25}{=} \frac{1}{2} \cdot \sum_{i \in I} c_i^2 \cdot u_{i,i} + c_i \cdot c_{n_I(i)} \cdot u_{i,n_I(i)} + c_i \cdot c_{n_I(i)} \cdot u_{n_I(i),i} \\ &\overset{Satz\ 3.25}{=} \sum_{i \in I} \frac{1}{2} \cdot c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)}. \end{aligned}$$

Die Aussage für Φ_2 folgt analog. □

Das soeben bewiesene Korollar stellt eine wichtige Zwischenetappe auf dem Weg zu einem hinreichenden Optimalitätskriterium für (AP2-I) dar. Ein solches Optimalitätskriterium sollte nach Möglichkeit mit nur geringem Aufwand überprüfbar sein. Wir benötigen daher noch eine sinnvolle Abschätzung der Summanden $\alpha_{r,s,t}$, $\beta_{r,s,t}$, $\gamma_{r,s}$ und $\delta_{r,s}$, da die Ermittlung aller im Korollar vorkommenden Terme $(A^{-1}_{\{r,t\},\{r,t\}})_{s,s}$ und $(B^{-1}_{\{r,t\},\{r,t\}})_{s,s}$ zu rechenintensiv wäre.

Wir nähern uns diesem Problem, indem wir uns an die Definition der Inversen einer Matrix M durch ihre Adjunkte $\mathrm{adj}(M)$ erinnern:

$$(M^{-1})_{i,j} = \frac{1}{\det(M)} \cdot (\mathrm{adj}(M))_{i,j}$$

mit

$$(\mathrm{adj}(M))_{i,j} = (-1)^{i+j} \cdot \det(M^{\{j\},\{i\}})$$

und $i, j \in \Omega_N$.

Um nun mehr über die Eigenschaften solcher Determinanten zu erfahren, folgen sogleich vier Lemmata eher technischer Natur, die schließlich alle in Satz 3.31 zur Anwendung kommen werden.

Lemma 3.27 (Determinanten von Blockmatrizen)**.** *Es seien $U \in \mathbb{R}^{l_1 \times l_1}, W \in \mathbb{R}^{l_2 \times l_2}$ reguläre Matrizen und $V \in \mathbb{R}^{l_1 \times l_2}$, dann gilt:*

$$\det \begin{pmatrix} U & V \\ 0 & W \end{pmatrix} = \det(U) \cdot \det(W).$$

Lemma 3.28. *Für die Determinante von $\tilde{M}$ gilt:*

$$\begin{aligned}\det(\tilde{M}) &= m_{r,r} \cdot \det(M_1) \cdot \det(M_2) \\ &\quad -m_{r-1,r}^2 \cdot \det(M_1^{\{l_1\},\{l_1\}}) \cdot \det(M_2) \\ &\quad -m_{r,r+1}^2 \cdot \det(M_1) \cdot \det(M_2^{\{1\},\{1\}}).\end{aligned}$$

Beweis. *Wir präzisieren die Form der Matrizen M_1 und M_2 zu:*

$$M_1 = \begin{pmatrix} \hat{M}_1 & m_{r-2} \\ m_{r-2}^T & m_{r-1,r-1} \end{pmatrix},$$

mit

$$\hat{M}_1 \in \mathbb{R}^{(l_1-1)\times(l_1-1)},\ m_{r-2} = (0,\ldots,0,m_{r-2,r-1})^T \in \mathbb{R}^{l_1-1}$$

sowie

$$M_2 = \begin{pmatrix} m_{r+1,r+1} & m_{r+2}^T \\ m_{r+2} & \hat{M}_2 \end{pmatrix}$$

mit

$$\hat{M}_2 \in \mathbb{R}^{(l_2-1)\times(l_2-1)},\ m_{r+2} = (m_{r+1,r+2},0,\ldots,0)^T \in \mathbb{R}^{l_2-1}.$$

Gemäß der in Definition 3.1 eingeführten Notation ist damit

$$M_1^{\{l_1\},\{l_1\}} = \hat{M}_1$$

die Matrix M_1 mit gestrichener letzter Zeile und Spalte. Analog steht

$$M_2^{\{1\},\{1\}} = \hat{M}_2$$

für die Matrix M_2, bei welcher die erste Zeile und Spalte gestrichen wurden.
Der eigentliche Beweis des Lemmas besteht nun lediglich aus einer Determinanten-Entwicklung von $\tilde{M}$ nach der r. Spalte:

$$\det(\tilde{M}) = \det\begin{pmatrix} \hat{M}_1 & m_{r-2} & & & \\ m_{r-2}^T & m_{r-1,r-1} & m_{r-1,r} & & \\ & m_{r-1,r} & m_{r,r} & m_{r,r+1} & \\ & & m_{r,r+1} & m_{r+1,r+1} & m_{r+2}^T \\ & & & m_{r+2} & \hat{M}_2 \end{pmatrix}$$

$$
\begin{aligned}
= \quad & -m_{r-1,r} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} & & \\ 0 & m_{r-1,r} & m_{r,r+1} & \\ & & m_{r+1,r+1} & m_{r+2}^T \\ & & m_{r+2} & \hat{M}_2 \end{pmatrix} \\
& +m_{r,r} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} & & \\ m_{r-2}^T & m_{r-1,r-1} & & \\ & & m_{r+1,r+1} & m_{r+2}^T \\ & & m_{r+2} & \hat{M}_2 \end{pmatrix} \\
& -m_{r,r+1} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} & & \\ m_{r-2}^T & m_{r-1,r-1} & & \\ & m_{r,r-1} & m_{r,r+1} & 0 \\ & & m_{r+2} & \hat{M}_2 \end{pmatrix} \\
\overset{L.\,3.27}{=} \quad & -m_{r-1,r} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} \\ 0 & m_{r-1,r} \end{pmatrix} \cdot \det \begin{pmatrix} m_{r+1,r+1} & m_{r+2}^T \\ m_{r+2} & \hat{M}_2 \end{pmatrix} \\
& +m_{r,r} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} \\ m_{r-2}^T & m_{r-1,r-1} \end{pmatrix} \cdot \det \begin{pmatrix} m_{r+1,r+1} & m_{r+2}^T \\ m_{r+2} & \hat{M}_2 \end{pmatrix} \\
& -m_{r,r+1} \cdot \det \begin{pmatrix} \hat{M}_1 & m_{r-2} \\ m_{r-2}^T & m_{r-1,r-1} \end{pmatrix} \cdot \det \begin{pmatrix} m_{r,r+1} & 0 \\ m_{r+2} & \hat{M}_2 \end{pmatrix} \\
= \quad & m_{r,r} \cdot \det(M_1) \cdot \det(M_2) \\
& -m_{r-1,r}^2 \cdot \det(\hat{M}_1) \cdot \det(M_2) \\
& -m_{r,r+1}^2 \cdot \det(M_1) \cdot \det(\hat{M}_2).
\end{aligned}
$$

□

Lemma 3.29. *Es gilt:*

$$\det(M_2) \cdot \det(\tilde{M}^{\{N\},\{N\}}) - \det(M_2^{\{l_2\},\{l_2\}}) \cdot \det(\tilde{M}) \geq 0.$$

Beweis. *Das Lemma lässt sich durch Induktion über l_2 beweisen. Es sei dazu nochmal an die Form von $\tilde{M}$ erinnert:*

$$\tilde{M} = \begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T \\ & m_{r+1} & M_2 \end{pmatrix}.$$

Induktionsanfang: $l_2 = 1$
In diesem Fall ist $M_2 = m_{r+1,r+1}$ und $M_2^{\{l_2\},\{l_2\}}$ damit die leere Matrix, welche die Determinante Eins hat. Es ist also zu zeigen:

$$m_{r+1,r+1} \cdot \det \begin{pmatrix} M_1 & m_{r-1} \\ m_{r-1}^T & m_{r,r} \end{pmatrix} - 1 \cdot \det \underbrace{\begin{pmatrix} M_1 & m_{r-1} & \\ m_{r-1}^T & m_{r,r} & m_{r,r+1} \\ & m_{r,r+1} & m_{r+1,r+1} \end{pmatrix}}_{=\tilde{M}} \overset{!}{\geq} 0.$$

Wir wenden Lemma 3.28 bzgl. der letzten Spalte von $\det(\tilde{M})$ an und erhalten:

$$m_{r+1,r+1} \cdot \det \begin{pmatrix} M_1 & m_{r-1} \\ m_{r-1}^T & m_{r,r} \end{pmatrix} - \det(\tilde{M}) = m_{r+1,r+1} \cdot \det \begin{pmatrix} M_1 & m_{r-1} \\ m_{r-1}^T & m_{r,r} \end{pmatrix}$$

$$-m_{r+1,r+1} \cdot \det \begin{pmatrix} M_1 & m_{r-1} \\ m_{r-1}^T & m_{r,r} \end{pmatrix} + m_{r,r+1}^2 \cdot \det(M_1) = \underbrace{m_{r,r+1}^2}_{\geq 0} \cdot \underbrace{\det(M_1)}_{>0} \geq 0.$$

$\det(M_1)$ *ist positiv, da $\tilde{M}$ nach Voraussetzung symmetrisch positiv definit ist und nach Satz 3.14 alle Hauptminoren von symmetrisch positiv definiten Matrizen positiv sind.*

Induktionsschritt: $l_2 \rightsquigarrow l_2 + 1$
Die zu beweisende Aussage ist nun die folgende:

$$\det \begin{pmatrix} M_2 & m_{l_2+1} \\ m_{l_2+1}^T & m_{l_2+1,l_2+1} \end{pmatrix} \cdot \det(\tilde{M}) - \det(M_2) \cdot \det \begin{pmatrix} \tilde{M} & m_{l_2+1} \\ m_{l_2+1}^T & m_{l_2+1,l_2+1} \end{pmatrix} \overset{!}{\geq} 0$$

mit m_{l_2+1} wie gehabt:

$$m_{l_2+1} = (0, \ldots, 0, m_{l_2,l_2+1})^T.$$

Wir wenden erneut Lemma 3.28 an:

$$\det \begin{pmatrix} M_2 & m_{l_2+1} \\ m_{l_2+1}^T & m_{l_2+1,l_2+1} \end{pmatrix} \cdot \det(\tilde{M}) - \det(M_2) \cdot \det \begin{pmatrix} \tilde{M} & m_{l_2+1} \\ m_{l_2+1}^T & m_{l_2+1,l_2+1} \end{pmatrix}$$

$$
\begin{aligned}
&= \Big[m_{l_2+1,l_2+1} \cdot \det(M_2) - m_{l_2,l_2+1}^2 \cdot \det(M_2^{\{l_2\},\{l_2\}})\Big] \cdot \det(\tilde{M}) \\
&\quad - \det(M_2) \cdot \Big[m_{l_2+1,l_2+1} \cdot \det(\tilde{M}) - m_{l_2,l_2+1}^2 \cdot \det(\tilde{M}^{\{N\},\{N\}})\Big] \\
&= m_{l_2,l_2+1}^2 \cdot \det(M_2) \cdot \det(\tilde{M}^{\{N\},\{N\}}) - m_{l_2,l_2+1}^2 \cdot \det(M_2^{\{l_2\},\{l_2\}}) \cdot \det(\tilde{M}) \overset{!}{\geq} 0 \\
\iff & \det(M_2) \cdot \det(\tilde{M}^{\{N\},\{N\}}) - \det(M_2^{\{l_2\},\{l_2\}}) \cdot \det(\tilde{M}) \geq 0.
\end{aligned}
$$

Bei der letzten Ungleichung handelt es sich um die Induktionsannahme. □

Lemma 3.30. *Es gilt:*

$$\det(M^{\{r\},\{s\}}) = \det(M_1) \cdot \mu \cdot \det(M_3)$$

mit

$$\mu := \prod_{i=r+1}^{s} m_{i,i-1}$$

Beweis. *Wir präzisieren zunächst wieder die Gestalt von M:*

$$
M = \det \begin{pmatrix}
M_1 & m_{r-1} & & & & & \\
m_{r-1}^T & m_{r,r} & m_{r,r+1} & & & & \\
 & m_{r,r+1} & \ddots & \ddots & & & \\
 & & \ddots & m_{s-1,s-1} & m_{s-1,s} & & \\
 & & & m_{s-1,s} & m_{s,s} & m_{s+1}^T & \\
 & & & & m_{s+1} & M_3 &
\end{pmatrix}.
$$

Mithin erhalten wir für die Determinante der Streichungsmatrix $M^{\{r\},\{s\}}$:

$$
\det(M^{\{r\},\{s\}}) = \det \begin{pmatrix}
M_1 & m_{r-1} & & & & & \\
 & m_{r,r+1} & m_{r+1,r+1} & m_{r+1,r+2} & & & \\
 & & \ddots & \ddots & \ddots & & \\
 & & & \ddots & \ddots & m_{s-2,s-1} & \\
 & & & & \ddots & m_{s-1,s-1} & \\
 & & & & & m_{s-1,s} & m_{s+1}^T \\
 & & & & & & M_3
\end{pmatrix}
$$

$$\overset{L.\,3.27}{=} \det(M_1)\cdot \det\underbrace{\begin{pmatrix} m_{r,r+1} & m_{r+1,r+1} & m_{r+1,r+2} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & m_{s-2,s-1} \\ & & & \ddots & m_{s-1,s-1} \\ & & & & m_{s-1,s} \end{pmatrix}}_{\prod_{i=r+1}^{s} m_{i,i-1}} \cdot \det(M_3).$$

□

Die vorigen Lemmata dienen uns nun dazu, die Elemente der Matrizen $A^{-1}_{\{r,t\},\{r,t\}}$ und $B^{-1}_{\{r,t\},\{r,t\}}$ aus Korollar 3.26 durch die von A^{-1} und B^{-1} abzuschätzen.

Satz 3.31. *Es sei $I \subseteq \Omega_N$ und nach wie vor $M \in \mathbb{R}^{N\times N}$ eine tridiagonale symmetrisch positiv definite Matrix sowie $r, s \in \Omega_N$ mit $r \leq s$.*

Gilt für alle $i \in I$ entweder $i < r$ oder $s < i$, so ist:

$$|(M^{-1}_{I,I})_{r,s}| \leq |(M^{-1})_{r,s}|,$$

$$\operatorname{sign}((M^{-1}_{I,I})_{r,s}) = \operatorname{sign}((M^{-1})_{r,s}).$$

Im Falle, dass ein $i \in I$ existiert mit $r \leq i \leq s$, gilt:

$$(M^{-1}_{I,I})_{r,s} = \begin{cases} 0, & \text{falls } r \neq s \\ 1, & \text{falls } r = s \end{cases}.$$

Beweis. *Es sei $I = \{i\} \subseteq \Omega_N$. M habe für diesen Beweis folgende Form:*

$$M = \begin{pmatrix} H & m_{i-1} & \\ m_{i-1}^T & m_{i,i} & m_{i+1}^T \\ & m_{i+1} & M_4 \end{pmatrix},$$

mit

$$m_{i-1} = (0,\ldots,0,m_{i-1,i})^T,\ m_{i+1} = (m_{i,i+1},0,\ldots,0)^T$$

und

$$H = \begin{pmatrix} M_1 & m_{r-1} & & & \\ m_{r-1}^T & m_{r,r} & m_{r+1}^T & & \\ & m_{r+1} & M_2 & m_{s-1} & \\ & & m_{s-1}^T & m_{s,s} & m_{s+1}^T \\ & & & m_{s+1} & M_3 \end{pmatrix}.$$

1. Fall: $r \leq s < i$:
Gemäß der Definition der Adjunkten einer Matrix lassen sich die Elemente $(M^{-1})_{r,s}$ und $(M_{I,I}^{-1})_{r,s}$ folgendermaßen ausdrücken:

$$(M^{-1})_{r,s} = \frac{(-1)^{r+s}}{\det(M)} \cdot \det(M^{\{r\},\{s\}}), \tag{3.25}$$

$$(M_{I,I}^{-1})_{r,s} = \frac{(-1)^{r+s}}{\det(M_{I,I})} \cdot \det(M_{I,I}^{\{r\},\{s\}}). \tag{3.26}$$

Ziel dieses Beweises ist es, die vier Determinanten

$$\det(M^{\{r\},\{s\}}),\ \det(M),\ \det(M_{I,I}^{\{r\},\{s\}}),\ \det(M_{I,I})$$

so zu entwickeln, dass man sie mit Hilfe der vorhergehenden Lemmata miteinander vergleichen kann:

$$\begin{aligned} \det(M^{\{r\},\{s\}}) &\overset{L.\,3.30}{=} \det(M_1) \cdot \mu \cdot \det \begin{pmatrix} M_3 & m_{i-1} & \\ m_{i-1}^T & m_{i,i} & m_{i+1}^T \\ & m_{i+1} & M_4 \end{pmatrix} \\ &\overset{L.\,3.28}{=} \det(M_1) \cdot \mu \cdot [m_{i,i} \cdot \det(M_3) \cdot \det(M_4) \\ &\quad -m_{i-1,i}^2 \cdot \det(M_3^{\{l_3\},\{l_3\}}) \cdot \det(M_4) \\ &\quad -m_{i,i+1}^2 \cdot \det(M_3) \cdot \det(M_4^{\{1\},\{1\}})], \\ \det(M) &\overset{L.\,3.28}{=} m_{i,i} \cdot \det(H) \cdot \det(M_4) \\ &\quad -m_{i-1,i}^2 \cdot \det(H^{\{i-1\},\{i-1\}}) \cdot \det(M_4) \\ &\quad -m_{i,i+1}^2 \cdot \det(H) \cdot \det(M_4^{\{1\},\{1\}}), \end{aligned}$$

$$\begin{aligned}
\det(M_{I,I}^{\{r\},\{s\}}) &\overset{L.\,3.30}{=} \det(M_1)\cdot\mu\cdot\begin{pmatrix} M_3 & 0 & \\ 0 & 1 & 0 \\ & 0 & M_4 \end{pmatrix} \\
&\overset{L.\,3.27}{=} \det(M_1)\cdot\mu\cdot\det(M_3)\cdot\det(M_4), \\
\det(M_{I,I}) &\overset{L.\,3.27}{=} \det(H)\cdot\det(M_4),
\end{aligned}$$

wobei μ wie in Lemma 3.30 definiert sei. Die so entwickelten Determinanten setzen wir nun in die zu beweisende Ungleichung ein:

$$\begin{aligned}
& |(M^{-1})_{r,s}| - |(M_{I,I}^{-1})_{r,s}| \geq 0 \\
\iff\ & \frac{|\det(M^{\{r\},\{s\}})|}{|\det(M)|} - \frac{|\det(M_{I,I}^{\{r\},\{s\}})|}{|\det(M_{I,I})|} \geq 0 \\
\iff\ & |\det(M_{I,I})|\cdot|\det(M^{\{r\},\{s\}})| - |\det(M_{I,I}^{\{r\},\{s\}})|\cdot|\det(M)| \geq 0 \\
\iff\ & \det(H)\cdot\det(M_1)\cdot|\mu|\cdot\det(M_4)\cdot[m_{i,i}\cdot\det(M_3)\cdot\det(M_4) \\
& -m_{i-1,i}^2\cdot\det(M_3^{\{l_3\},\{l_3\}})\cdot\det(M_4) - m_{i,i+1}^2\cdot\det(M_3)\cdot\det(M_4^{\{1\},\{1\}})] \\
& -\det(M_1)\cdot|\mu|\cdot\det(M_3)\cdot\det(M_4)\cdot[m_{i,i}\cdot\det(H)\cdot\det(M_4) \\
& -m_{i-1,i}^2\cdot\det(H^{\{i-1\},\{i-1\}})\cdot\det(M_4) - m_{i,i+1}^2\cdot\det(H)\cdot\det(M_4^{\{1\},\{1\}})] \geq 0 \\
\iff\ & -m_{i-1,i}^2\cdot\det(H)\cdot\det(M_3^{\{l_3\},\{l_3\}})\cdot\det(M_4) \\
& +m_{i-1,i}^2\cdot\det(H^{\{i-1\},\{i-1\}})\cdot\det(M_3)\cdot\det(M_4) \geq 0 \\
\iff\ & \det(H^{\{i-1\},\{i-1\}})\cdot\det(M_3) - \det(H)\cdot\det(M_3^{\{l_3\},\{l_3\}}) \geq 0
\end{aligned}$$

Dass die letzte Ungleichung erfüllt ist, haben wir bereits in Lemma 3.29 gezeigt.
Nun zum Vorzeichen von $(M^{-1})_{r,s}$ und $(M_{I,I}^{-1})_{r,s}$:
Da nach Satz 3.14 alle Hauptminoren von M strikt positiv sind, gilt offensichtlich:

$$\operatorname{sign}((M^{-1})_{r,s}) = (-1)^{r+s}\cdot\operatorname{sign}\left(\frac{\det(M^{\{r\},\{s\}})}{\det(M)}\right) = (-1)^{r+s}\cdot\operatorname{sign}(\mu)$$

und gleichermaßen

$$\operatorname{sign}((M_{I,I}^{-1})_{r,s}) = (-1)^{r+s}\cdot\operatorname{sign}\left(\frac{\det(M_{I,I}^{\{r\},\{s\}})}{\det(M_{I,I})}\right) = (-1)^{r+s}\cdot\operatorname{sign}(\mu).$$

2. Fall: $i < r \leq s$:
Dieser Fall verläuft völlig analog zum 1. Fall.

Die Aussage des Lemmas folgt nun induktiv, da sich jede Erweiterung der Menge I um ein oder mehrere $j \in \Omega_N$ (mit $j < r$ oder $s < j$) auf den hier behandelten Fall zurückführen lässt:

$$|(M^{-1}_{I\cup\{j\},I\cup\{j\}})_{r,s}| \leq |(M^{-1}_{I,I})_{r,s}| \leq |(M^{-1})_{r,s}|.$$

3. Fall: $r \leq i \leq s$:
In diesem Fall ist entweder $r = i = s$ oder $r \neq s$. Liegt ersteres vor, ist $M_{I,I}$ von der Form

$$M_{I,I} = \begin{pmatrix} M_1 & 0 & \\ 0 & m_{i,i} & 0 \\ & 0 & M_4 \end{pmatrix}$$

mit $m_{i,i} = m_{r,s} = 1$. Nach Lemma 3.23 gilt:

$$M^{-1}_{I,I} = \begin{pmatrix} M_1^{-1} & 0 & \\ 0 & m_{i,i} & 0 \\ & 0 & M_4^{-1} \end{pmatrix}$$

und daher

$$(M^{-1}_{I,I})_{r,s} = 1.$$

Alternativ ist $r \neq s$ und damit $M_{I,I}$ der Gestalt

$$M_{I,I} = \begin{pmatrix} M_1 & 0 & \\ 0 & m_{i,i} & 0 \\ & 0 & M^* \end{pmatrix},$$

wobei $M^* \in \mathbb{R}^{(N-l_1-1)\times(N-l_1-1)}$, $m_{i,i} = 1$ und

$$\text{entweder } r < i \leq s \text{ oder } r \leq i < s.$$

Nach Lemma 3.23 ist

$$M^{-1}_{I,I} = \begin{pmatrix} M_1^{-1} & 0 & \\ 0 & m_{i,i} & 0 \\ & 0 & (M^*)^{-1} \end{pmatrix}.$$

Offensichtlich sind aber dann alle Matrixeinträge $(M^{-1}_{I,I})_{p,q}$ mit

$$\min(p,q) < i \leq \max(p,q) \text{ oder } \min(p,q) \leq i < \max(p,q)$$

gleich Null. Es folgt:

$$(M_{I,I}^{-1})_{r,s} = 0.$$

□

Diese Erkenntnis führt uns schließlich zu folgendem Korollar:

Korollar 3.32. *Für eine aufsteigende Folge von Teilmengen $I_1 \subseteq I_2 \subseteq ... \subseteq I_l \subseteq \Omega_N$ und $r \notin (I_1 \cup I_2 \cup ... \cup I_l)$ gilt:*

$$\frac{1}{m_{r,r}} \leq (M_{I_l,I_l}^{-1})_{r,r} \leq ... \leq (M_{I_2,I_2}^{-1})_{r,r} \leq (M_{I_1,I_1}^{-1})_{r,r} \leq (M^{-1})_{r,r}.$$

Beweis. *Die Aussage folgt induktiv aus Satz 3.31. Man beachte dazu, dass*

$$\frac{1}{m_{r,r}} = (M_{\Omega_N\setminus\{r\},\Omega_N\setminus\{r\}}^{-1})_{r,r}$$

gilt, denn

$$M_{\Omega_N\setminus\{r\},\Omega_N\setminus\{r\}} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & m_{r,r} & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}.$$

□

Wir wollen als nächstes die soeben hergeleiteten Abschätzungen auf die Summanden von Φ_1 und Φ_2 anwenden.

Lemma 3.33. *Es seien $I, J \subseteq \Omega_N$, dann gelten für die Summanden von $\Phi_1(I)$ (vgl. Satz 3.26) folgende Abschätzungen:*
$\forall i \in I$:

$$\begin{aligned} & c_i^2(\alpha_i - \varepsilon_{n_I(i)} \cdot \gamma_{max}) - c_{n_I(i)}^2 \cdot \gamma_{max} \\ \leq\ & c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} \\ \leq\ & c_i^2(a_{ii} + \varepsilon_{n_I(i)} \cdot \gamma_{max}) + c_{n_I(i)}^2 \cdot \gamma_{max}, \end{aligned}$$

wobei $\gamma_{max} := \max_{r,s \in \Omega_N} |\gamma_{r,s}|$.
Analoges gilt für die Summanden von $\Phi_2(J)$:
$\forall j \in J$:

$$\begin{aligned} & d_j^2(\beta_j - \varepsilon_{n_J(j)} \cdot \delta_{max}) - d_{n_J(j)}^2 \cdot \delta_{max} \\ \leq\ & d_i^2 \cdot \beta_{v_J(j),j,n_J(j)} + 2 \cdot d_j \cdot d_{n_J(j)} \cdot \delta_{j,n_J(j)} \\ \leq\ & d_j^2(b_{jj} + \varepsilon_{n_J(j)} \cdot \delta_{max}) + d_{n_J(j)}^2 \cdot \delta_{max}, \end{aligned}$$

wobei $\delta_{max} := \max_{r,s \in \Omega_N} |\delta_{r,s}|$.

Beweis. *Zu allererst zu einer fundamentalen Abschätzung, die aus den binomischen Formeln folgt:*

$$0 \leq (p-q)^2 \Longleftrightarrow 0 \leq p^2 - 2pq + q^2 \Longleftrightarrow 2pq \leq p^2 + q^2 \tag{3.27}$$

was weiterhin äquivalent ist zu

$$-2pq \geq -p^2 - q^2. \tag{3.28}$$

Wir beginnen damit, die Summanden von Φ_1 nach oben abzuschätzen und unterscheiden dabei zwei Fälle:

1. Fall: i hat einen Nachfolger in I, d. h. $n_I(i) \neq 0$

$$\begin{aligned} c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} & \overset{K.\,3.32}{\leq} c_i^2 \cdot a_{ii} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} \\ & \leq c_i^2 \cdot a_{ii} + 2 \cdot |c_i| \cdot |c_{n_I(i)}| \cdot \gamma_{max} \\ & \overset{(3.27)}{\leq} c_i^2 \cdot a_{ii} + |c_i|^2 \cdot \gamma_{max} + |c_{n_I(i)}|^2 \cdot \gamma_{max} \\ & = c_i^2 \cdot (a_{ii} + \gamma_{max}) + c_{n_I(i)}^2 \cdot \gamma_{max}. \end{aligned}$$

2. Fall: i hat keinen Nachfolger in I, d. h. $n_I(i) = 0$

$$c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} = c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} \overset{K.\,3.32}{\leq} c_i^2 \cdot a_{ii}.$$

Aufgrund der Festlegung, dass $c_0 = 0$ (siehe Definition 2.1) und $\varepsilon_0 = 0$ (siehe Definition 3.1), können wir beide Fälle auf eine gemeinsame Notation zurückführen und schreiben:

$$c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} \leq c_i^2(a_{ii} + \varepsilon_{n_I(i)} \cdot \gamma_{max}) + c_{n_I(i)}^2 \cdot \gamma_{max}.$$

Die Abschätzung der Summanden von Φ_1 nach unten verläuft analog:

$$\begin{aligned} c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} &\overset{K.\,3.32}{\geq} c_i^2 \cdot \alpha_i + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)} \\ &\geq c_i^2 \cdot \alpha_i - 2 \cdot |c_i| \cdot |c_{n_I(i)}| \cdot \gamma_{max} \\ &\overset{(3.28)}{\geq} c_i^2 \cdot \alpha_i - \varepsilon_{n_I(i)} \cdot |c_i|^2 \cdot \gamma_{max} \\ &\quad -|c_{n_I(i)}|^2 \cdot \gamma_{max} \\ &= c_i^2(\alpha_i - \varepsilon_{n_I(i)} \cdot \gamma_{max}) - c_{n_I(i)}^2 \cdot \gamma_{max}. \end{aligned}$$

□

Anmerkung 3.34. *γ_{max} und δ_{max} lassen sich dabei folgendermaßen abschätzen:*

$$\begin{aligned} 0 \leq \gamma_{max} &\leq \max \left[\max_{\substack{r,s \in \Omega_N \\ |r-s|=1}} (|a_{r,s}|) \,,\, \max_{r \in \Omega_N}(|a_{r+1,r}|) \cdot \max_{s \in \Omega_N}(|a_{s-1,s}|) \cdot \max_{\substack{r,s \in \Omega_N \\ |r-s|>1}} \left|(A^{-1})_{r,s}\right| \right] \\ &= \max \left[\max_{r \in \Omega_N}(|a_{r,r+1}|) \,,\, \left(\max_{r \in \Omega_N}(|a_{r,r+1}|)\right)^2 \cdot \max_{\substack{r,s \in \Omega_N \\ |r-s|>1}} \left|(A^{-1})_{r,s}\right| \right], \\ 0 \leq \delta_{max} &\leq \max \left[\max_{\substack{r,s \in \Omega_N \\ |r-s|=1}} (|b_{r,s}|) \,,\, \max_{r \in \Omega_N}(|b_{r+1,r}|) \cdot \max_{s \in \Omega_N}(|b_{s-1,s}|) \cdot \max_{\substack{r,s \in \Omega_N \\ |r-s|>1}} \left|(B^{-1})_{r,s}\right| \right] \\ &= \max \left[\max_{r \in \Omega_N}(|b_{r,r+1}|) \,,\, \left(\max_{r \in \Omega_N}(|b_{r,r+1}|)\right)^2 \cdot \max_{\substack{r,s \in \Omega_N \\ |r-s|>1}} \left|(B^{-1})_{r,s}\right| \right]. \end{aligned}$$

Nach diesen umfangreichen Vorbereitungen können wir jetzt eine hinreichende Bedingung dafür aufstellen, dass eine Menge $I_0 \in \Omega_N$ eine globale Lösung von (AP2-I) ist.

Satz 3.35 (Hinreichende Optimalitätsbedingung für (AP2-I))**.** *Existiert eine Menge $I_0 \subseteq \Omega_N$, für die gilt:*

$$\begin{aligned}
&\forall i \in I_0 : \\
&c^2_{v_{I_0}(i)}(a_{v_{I_0}(i),v_{I_0}(i)} + \gamma_{max}) + c^2_i(a_{ii} + \varepsilon_{v_{I_0}(i),n_{I_0}(i)} \cdot \gamma_{max}) \\
&+c^2_{n_{I_0}(i)}(a_{n_{I_0}(i),n_{I_0}(i)} + \gamma_{max}) + d^2_{v_{J_0}(i)}(b_{v_{J_0}(i),v_{J_0}(i)} + \delta_{max}) \\
&+d^2_{n_{J_0}(i)}(b_{n_{J_0}(i),n_{J_0}(i)} + \delta_{max}) \\
\leq\ & c^2_{v_{I_0}(i)}(\alpha_{v_{I_0}(i)} - \gamma_{max}) + c^2_{n_{I_0}(i)}(\alpha_{n_{I_0}(i)} - \gamma_{max}) + d^2_{v_{J_0}(i)}(\beta_{v_{J_0}(i)} - \delta_{max}) \\
&+d^2_i(\beta_i - \varepsilon_{v_\Omega(i),n_\Omega(i)} \cdot \delta_{max}) + d^2_{n_{J_0}(i)}(\beta_{n_{J_0}(i)} - \delta_{max}), \qquad (3.29)
\end{aligned}$$

$$\begin{aligned}
&\forall j \in J_0 : \\
&d^2_{v_{J_0}(j)}(b_{v_{J_0}(j),v_{J_0}(j)} + \delta_{max}) + d^2_j(b_{jj} + \varepsilon_{v_{J_0}(j),n_{J_0}(j)} \cdot \delta_{max}) \\
&+d^2_{n_{J_0}(j)}(b_{n_{J_0}(j),n_{J_0}(j)} + \delta_{max}) + c^2_{v_{I_0}(j)}(a_{v_{I_0}(j),v_{I_0}(j)} + \gamma_{max}) \\
&+c^2_{n_{I_0}(j)}(a_{n_{I_0}(j),n_{I_0}(j)} + \gamma_{max}) \\
\leq\ & d^2_{v_{J_0}(j)}(\beta_{v_{J_0}(i)} - \delta_{max}) + d^2_{n_{J_0}(j)}(\beta_{n_{J_0}(j)} - \delta_{max}) + c^2_{v_{I_0}(j)}(\alpha_{v_{I_0}(j)} - \gamma_{max}) \\
&+c^2_j(\alpha_j - \varepsilon_{v_\Omega(j),n_\Omega(j)} \cdot \gamma_{max}) + c^2_{n_{I_0}(j)}(\alpha_{n_{I_0}(j)} - \gamma_{max}), \qquad (3.30)
\end{aligned}$$

wobei

$$J_0 := \bar{I}_0$$

und

$$\forall i \in \Omega_N : v_\Omega(i) := v_{\Omega_N}(i),\ n_\Omega(i) := n_{\Omega_N}(i).$$

Dann ist I_0 Lösung von (AP2-I)*, d. h.*

$$\forall I \subseteq \Omega_N : \Phi_1(I_0) + \Phi_2(\bar{I}_0) \leq \Phi_1(I) + \Phi_2(\bar{I}).$$

Beweis. *$I_0 \subseteq \Omega_N$ erfülle die Bedingung des Satzes. $I \neq I_0$ sei eine beliebige Teilmenge von Ω_N. Zur besseren Lesbarkeit definieren wir:*

$$J_0 := \bar{I}_0,$$

$$J := \bar{I}.$$

Wir zerlegen als erstes die Mengen I, I_0:

$$\begin{aligned}
\hat{I}_0 &:= I_0 \setminus I, \\
\hat{I} &:= I \setminus I_0, \\
I_s^{(1)} &:= \{i \in I_0 \cap I \mid v_{I_0}(i) \neq v_I(i)\ \wedge\ n_{I_0}(i) \neq n_I(i)\},
\end{aligned}$$

$$I_s^{(2)} := \{i \in I_0 \cap I \mid v_{I_0}(i) = v_I(i) \ \wedge \ n_{I_0}(i) \neq n_I(i)\},$$
$$I_s^{(3)} := \{i \in I_0 \cap I \mid v_{I_0}(i) \neq v_I(i) \ \wedge \ n_{I_0}(i) = n_I(i)\},$$
$$I_s^{(4)} := \{i \in I_0 \cap I \mid v_{I_0}(i) = v_I(i) \ \wedge \ n_{I_0}(i) = n_I(i)\}.$$

Man beachte, dass

$$I_0 = \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)} \cup I_s^{(3)} \cup I_s^{(4)}$$

und $\hat{I}_0, I_s^{(1)}, I_s^{(2)}, I_s^{(3)}, I_s^{(4)}$ *paarweise disjunkt.*
Ebenso ist

$$I = \hat{I} \cup I_s^{(1)} \cup I_s^{(2)} \cup I_s^{(3)} \cup I_s^{(4)}$$

und $\hat{I}, I_s^{(1)}, I_s^{(2)}, I_s^{(3)}, I_s^{(4)}$ *paarweise disjunkt.*
In analoger Weise definieren wir:

$$\hat{J}_0 := J_0 \setminus J,$$
$$\hat{J} := J \setminus J_0,$$
$$J_s^{(1)} := \{j \in J_0 \cap J \mid v_{J_0}(j) \neq v_J(j) \ \wedge \ n_{J_0}(j) \neq n_J(j)\},$$
$$J_s^{(2)} := \{j \in J_0 \cap J \mid v_{J_0}(j) = v_J(j) \ \wedge \ n_{J_0}(j) \neq n_J(j)\},$$
$$J_s^{(3)} := \{j \in J_0 \cap J \mid v_{J_0}(j) \neq v_J(j) \ \wedge \ n_{J_0}(j) = n_J(j)\},$$
$$J_s^{(4)} := \{j \in J_0 \cap J \mid v_{J_0}(j) = v_J(j) \ \wedge \ n_{J_0}(j) = n_J(j)\}.$$

Wir betrachten nun die Ungleichung, die es zu beweisen gilt:

$$\Phi_1(I_0) + \Phi_2(J_0) \leq \Phi_1(I) + \Phi_2(J) \tag{3.31}$$
$$\Longleftrightarrow 2 \cdot \Phi_1(I_0) + 2 \cdot \Phi_2(J_0) \leq 2 \cdot \Phi_1(I) + 2 \cdot \Phi_2(J), \tag{3.32}$$

was weiterhin äquivalent ist zu

$$\sum_{i \in I_0} \underbrace{\left(c_i^2 \cdot \alpha_{v_{I_0}(i),i,n_{I_0}(i)} + 2 \cdot c_i \cdot c_{n_{I_0}(i)} \cdot \gamma_{i,n_{I_0}(i)}\right)}_{=:R_{I_0}(i)} + 2 \cdot \Phi_2(J_0)$$
$$\leq \sum_{i \in I} \underbrace{\left(c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)}\right)}_{=:R_I(i)} + 2 \cdot \Phi_2(J). \tag{3.33}$$

Die Zerlegung von I_0 und I in die oben definierten Teilmengen führt uns von (3.33) zu folgender Ungleichung:

$$\sum_{i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}} R_{I_0}(i) + \sum_{i \in I_s^{(3)}} \left(c_i^2 \cdot \alpha_{v_{I_0}(i),i,n_{I_0}(i)} + 2 \cdot c_i \cdot c_{n_{I_0}(i)} \cdot \gamma_{i,n_{I_0}(i)}\right)$$
$$+ \sum_{i \in I_s^{(4)}} R_{I_0}(i) + 2 \cdot \Phi_2(J_0) \leq \sum_{i \in \hat{I} \cup I_s^{(1)} \cup I_s^{(2)}} R_I(i)$$

$$+ \sum_{i \in I_s^{(3)}} (c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot c_i \cdot c_{n_I(i)} \cdot \gamma_{i,n_I(i)}) + \sum_{i \in I_s^{(4)}} R_I(i) + 2 \cdot \Phi_2(J). \tag{3.34}$$

Offensichtlich ist $R_{I_0}(i) = R_I(i)$ für $i \in I_s^{(4)}$, da für solche i nach Definition

$$v_{I_0}(i) = v_I(i) \text{ und } n_{I_0}(i) = n_I(i)$$

und damit

$$\alpha_{v_{I_0}(i),i,n_{I_0}(i)} = \alpha_{v_I(i),i,n_I(i)}$$

sowie

$$c_{n_{I_0}}(i) = c_{n_I}(i) \text{ und } \gamma_{i,n_{I_0}(i)} = \gamma_{i,n_I(i)}.$$

Für $i \in I_s^{(3)}$ gilt hingegen lediglich:

$$n_{I_0}(i) = n_I(i)$$

und mithin

$$c_{n_{I_0}}(i) = c_{n_I}(i) \text{ und } \gamma_{i,n_{I_0}(i)} = \gamma_{i,n_I(i)}.$$

Damit vereinfacht sich die Ungleichung (3.34) zu:

$$\begin{aligned} &\sum_{i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}} R_{I_0}(i) + \sum_{i \in I_s^{(3)}} c_i^2 \cdot \alpha_{v_{I_0}(i),i,n_{I_0}(i)} + 2 \cdot \Phi_2(J_0) \\ \leq &\sum_{i \in \hat{I} \cup I_s^{(1)} \cup I_s^{(2)}} R_I(i) + \sum_{i \in I_s^{(3)}} c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} + 2 \cdot \Phi_2(J). \end{aligned} \tag{3.35}$$

Untersuchen wir nun die ersten beiden Summen der linken Seite der Ungleichung (3.35) etwas genauer. Mit der Abschätzung aus Lemma 3.33 erhalten wir:

$$\begin{aligned} &\sum_{i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}} R_{I_0}(i) + \sum_{i \in I_s^{(3)}} c_i^2 \cdot \alpha_{v_{I_0}(i),i,n_{I_0}(i)} \leq \\ \leq &\sum_{i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}} (c_i^2(a_{ii} + \varepsilon_{n_{I_0}(i)} \cdot \gamma_{max}) + c_{n_{I_0}(i)}^2 \cdot \gamma_{max}) + \sum_{i \in I_s^{(3)}} c_i^2 \cdot a_{ii} \\ \overset{(*)}{=} &\sum_{i \in \hat{I}_0 \cup I_s^{(1)}} c_i^2(a_{ii} + \varepsilon_{n_{I_0}(i)} \cdot \gamma_{max} + \varepsilon_{v_{I_0}(i)} \cdot \gamma_{max}) \\ &+ \sum_{i \in I_s^{(2)}} c_i^2(a_{ii} + \varepsilon_{n_{I_0}(i)} \cdot \gamma_{max}) + \sum_{i \in I_s^{(3)}} c_i^2(a_{ii} + \varepsilon_{v_{I_0}(i)} \cdot \gamma_{max}) \\ \leq &\sum_{i \in \hat{I}_0 \cup I_s^{(1)}} c_i^2(a_{ii} + \varepsilon_{v_{I_0}(i),n_{I_0}(i)} \cdot \gamma_{max}) + \sum_{i \in I_s^{(2)}} c_i^2(a_{ii} + \gamma_{max}) + \sum_{i \in I_s^{(3)}} c_i^2(a_{ii} + \gamma_{max}). \end{aligned}$$

Zu $(*)$:
Jedes $i \in I_s^{(2)}$ hat einen Nachfolger in $\hat{I}_0 \cup I_s^{(1)} \cup I_s^{(3)}$ oder aber der Nachfolger ist Null, d. h. [2]

$$\exists k \in (\hat{I}_0 \cup I_s^{(1)} \cup I_s^{(3)}) : n_{I_0}(i) = k \text{ oder } n_{I_0}(i) = 0. \tag{3.36}$$

[2] Zur Wiederholung: $n_{I_0}(i) = 0$ bedeutet, dass i bereits das größte Element in I_0 ist.

Denn andernfalls gälte

$$\exists k \in (I_s^{(2)} \cup I_s^{(4)}) : n_{I_0}(i) = k.$$

Für ein solches k wäre dann aber gemäß der Definition von $I_s^{(2)}$ und $I_s^{(4)}$

$$v_{I_0}(k) = v_I(k).$$

Mit $k = n_I(v_I(k)) = n_I(v_{I_0}(k)) = n_I(i)$ folgte daraus jedoch

$$n_{I_0}(i) = k = n_I(i),$$

was ein Widerspruch zur Definition von $I_s^{(2)}$ wäre.
Analog lässt sich für $i \in I_s^{(3)}$ zeigen, dass

$$\exists k \in (\hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}) : v_{I_0}(i) = k \text{ oder } v_{I_0}(i) = 0 \tag{3.37}$$

und für $i \in I_s^{(1)}$, dass

$$\exists k \in (\hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}) : v_{I_0}(i) = k \text{ oder } v_{I_0}(i) = 0 \tag{3.38}$$

und

$$\exists k \in (\hat{I}_0 \cup I_s^{(1)} \cup I_s^{(3)}) : n_{I_0}(i) = k \text{ oder } n_{I_0}(i) = 0. \tag{3.39}$$

Für $i \in \hat{I}_0$ gelten ebenfalls die Aussagen (3.38) und (3.39), denn andernfalls träfe wieder folgende Aussage zu:

$$\exists k \in (I_s^{(2)} \cup I_s^{(4)}) : n_{I_0}(i) = k$$

und damit wäre insbesondere $i = v_{I_0}(k)$. Da allerdings $i \notin I$, aber $k \in I_0 \cap I$ folgte

$$i = v_{I_0}(k) \neq v_I(k),$$

was ein Widerspruch zur Definition von $I_s^{(2)}$ und $I_s^{(4)}$ wäre. Aussage (3.38) zeigt man auf gleiche Weise.
Insgesamt ist also jedes $i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}$ Vorgänger eines $k \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(3)}$ oder aber größtes Element in I_0. Genauso ist jedes $i \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(3)}$ Nachfolger eines $k \in \hat{I}_0 \cup I_s^{(1)} \cup I_s^{(2)}$ oder aber kleinstes Element in I_0, womit sich die Identität $(*)$ erklärt.

Zurück zum eigentlichen Beweis: So wie wir eben die linke Seite der Ungleichung (3.35) umformten, schätzen wir nun mit Hilfe von Lemma 3.33 die ersten beiden Summen der rechten Seite von (3.35) nach unten ab:

$$\sum_{i \in \hat{I} \cup I_s^{(1)} \cup I_s^{(2)}} R_I(i) + \sum_{i \in I_s^{(3)}} c_i^2 \cdot \alpha_{v_I(i),i,n_I(i)} \geq \sum_{i \in \hat{I} \cup I_s^{(1)}} c_i^2 (\alpha_i - \varepsilon_{v_I(i),n_I(i)} \cdot \gamma_{max})$$

$$+ \sum_{i \in I_s^{(2)}} c_i^2 (\alpha_i - \varepsilon_{n_I(i)} \cdot \gamma_{max}) + \sum_{i \in I_s^{(3)}} c_i^2 (\alpha_i - \varepsilon_{v_I(i)} \cdot \gamma_{max})$$

$$\geq \sum_{i\in\hat{I}\cup I_s^{(1)}} c_i^2(\alpha_i - \varepsilon_{v_\Omega(i),n_\Omega(i)}\cdot\gamma_{max}) + \sum_{i\in I_s^{(2)}} c_i^2(\alpha_i-\gamma_{max}) + \sum_{i\in I_s^{(3)}} c_i^2(\alpha_i-\gamma_{max}).$$

Alle bisherigen Umformungen und Abschätzungen verlaufen analog für die Terme $\Phi_2(J_0)$ und $\Phi_2(J)$. Führt man die bisherigen Ergebnisse zusammen, entsteht folgende verschärfte Variante der Ausgangsungleichung (3.31):

$$\begin{aligned}
&\sum_{i\in\hat{I}_0\cup I_s^{(1)}} c_i^2(a_{ii} + \varepsilon_{v_{I_0}(i),n_{I_0}(i)}\cdot\gamma_{max}) + \sum_{i\in I_s^{(2)}} c_i^2(a_{ii}+\gamma_{max}) + \sum_{i\in I_s^{(3)}} c_i^2(a_{ii}+\gamma_{max})\\
&+ \sum_{j\in\hat{J}_0\cup J_s^{(1)}} d_j^2(b_{jj} + \varepsilon_{v_{J_0}(j),n_{J_0}(j)}\cdot\delta_{max}) + \sum_{j\in J_s^{(2)}} d_j^2(b_{jj}+\delta_{max}) + \sum_{j\in J_s^{(3)}} d_j^2(b_{jj}+\delta_{max})\\
&\leq \sum_{i\in\hat{I}\cup I_s^{(1)}} c_i^2(\alpha_i - \varepsilon_{v_\Omega(i),n_\Omega(i)}\cdot\gamma_{max}) + \sum_{i\in I_s^{(2)}} c_i^2(\alpha_i-\gamma_{max}) + \sum_{i\in I_s^{(3)}} c_i^2(\alpha_i-\gamma_{max})\\
&+ \sum_{j\in\hat{J}\cup J_s^{(1)}} d_j^2(\beta_j - \varepsilon_{v_\Omega(j),n_\Omega(j)}\cdot\delta_{max}) + \sum_{j\in J_s^{(2)}} d_j^2(\beta_j-\delta_{max}) + \sum_{j\in J_s^{(3)}} d_j^2(\beta_j-\delta_{max}). \quad (3.40)
\end{aligned}$$

Abschließend müssen die Summen auf beiden Seiten der Ungleichung nur noch umsortiert werden. Dazu stellen wir zunächst fest, dass

$$\hat{I}_0 = \hat{J} \text{ und } \hat{J}_0 = \hat{I} \quad (3.41)$$

gilt, denn:

$$\begin{aligned}
\hat{J} &= J\setminus J_0 = (\Omega_N\setminus I)\setminus J_0 = ((I_0\cup J_0)\setminus I)\setminus J_0\\
&\overset{(**)}{=} ((I_0\setminus I)\cup(J_0\setminus I))\setminus J_0) \overset{(**)}{=} \underbrace{[(I_0\setminus I)\setminus J_0]}_{=I_0\setminus I} \cup \underbrace{[(J_0\setminus I)\setminus J_0]}_{=\emptyset} = I_0\setminus I.
\end{aligned}$$

Zu (**):
Nach den Distributivgesetzen der Mengenlehre (vgl. [5]) gilt für Teilmengen X, Y, Z derselben Menge:

$$(X\cup Y)\setminus Z = (X\setminus Z)\cup(Y\setminus Z).$$

Die zweite Identität von (3.41) gilt entsprechend.
Betrachten wir (3.40) noch etwas genauer: Jedes $i\in I_s^{(2)}$ hat einen Nachfolger in $\hat{I}_0$ oder in $\hat{J}_0$, d. h. es gilt

$$(\exists k\in\hat{I}_0 : n_{I_0}(i)=k) \ \vee \ (\exists k\in\hat{J}_0 : n_{J_0}(i)=k). \quad (3.42)$$

Würde keine der beiden Bedingungen zutreffen, so hieße dies, dass sowohl

$$n_{I_0}(i)\in(I_0\cap I)$$

als auch

$$n_{J_0}(i)\in(J_0\cap J).$$

Somit wäre aber

$$n_{I_0}(i) = n_I(i),$$

was der Definition von $I_s^{(2)}$ widerspräche. Mit der gleichen Argumentation erhält man für alle $i \in I_s^{(3)}$:

$$(\exists k \in \hat{I}_0 : v_{I_0}(i) = k) \ \vee (\exists k \in \hat{J}_0 : v_{J_0}(i) = k). \tag{3.43}$$

Mithin muss für alle $i \in I_s^{(1)}$ sowohl

$$(\exists k \in \hat{I}_0 : v_{I_0}(i) = k) \ \vee (\exists k \in \hat{J}_0 : v_{J_0}(i) = k) \tag{3.44}$$

als auch

$$(\exists k \in \hat{I}_0 : n_{I_0}(i) = k) \ \vee (\exists k \in \hat{J}_0 : n_{J_0}(i) = k) \tag{3.45}$$

gelten. Analoges gilt natürlich auch für die Elemente von $J_s^{(1)}, J_s^{(2)}, J_s^{(3)}$.
Mit diesem Wissen und (3.41) können wir (3.40) in eine noch striktere Form überführen:

$$\begin{aligned}
&\textstyle\sum_{i\in\hat{I}_0}[c^2_{v_{I_0}(i)}(a_{v_{I_0}(i),v_{I_0}(i)} + \gamma_{max}) + c_i^2(a_{ii} + \varepsilon_{v_{I_0}(i),n_{I_0}(i)} \cdot \gamma_{max}) \\
&+c^2_{n_{I_0}(i)}(a_{n_{I_0}(i),n_{I_0}(i)} + \gamma_{max}) + d^2_{v_{J_0}(i)}(b_{v_{J_0}(i),v_{J_0}(i)} + \delta_{max}) \\
&+d^2_{n_{J_0}(i)}(b_{n_{J_0}(i),n_{J_0}(i)} + \delta_{max})] \\
&+\textstyle\sum_{j\in\hat{J}_0}[d^2_{v_{J_0}(j)}(b_{v_{J_0}(j),v_{J_0}(j)} + \delta_{max}) + d_j^2(b_{jj} + \varepsilon_{v_{J_0}(j),n_{J_0}(j)} \cdot \delta_{max}) \\
&+d^2_{n_{J_0}(j)}(b_{n_{J_0}(j),n_{J_0}(j)} + \delta_{max}) + c^2_{v_{I_0}(j)}(a_{v_{I_0}(j),v_{I_0}(j)} + \gamma_{max}) \\
&+c^2_{n_{I_0}(j)}(a_{n_{I_0}(j),n_{I_0}(j)} + \gamma_{max})] \\
\leq\ &\textstyle\sum_{i\in\hat{J}_0}[c^2_{v_{I_0}(i)}(\alpha_{v_{I_0}(i)} - \gamma_{max}) + c^2_{n_{I_0}(i)}(\alpha_{n_{I_0}(i)} - \gamma_{max}) + d^2_{v_{J_0}(i)}(\beta_{v_{J_0}(i)} - \delta_{max}) \\
&+d_i^2(\beta_i - \varepsilon_{v_\Omega(i),n_\Omega(i)} \cdot \delta_{max}) + d^2_{n_{J_0}(i)}(\beta_{n_{J_0}(i)} - \delta_{max})] \\
&+\textstyle\sum_{j\in\hat{I}_0}[d^2_{v_{J_0}(j)}(\beta_{v_{J_0}(i)} - \delta_{max}) + d^2_{n_{J_0}(j)}(\beta_{n_{J_0}(j)} - \delta_{max}) + c^2_{v_{I_0}(j)}(\alpha_{v_{I_0}(j)} - \gamma_{max}) \\
&+c_j^2(\alpha_j - \varepsilon_{v_\Omega(j),n_\Omega(j)} \cdot \gamma_{max}) + c^2_{n_{I_0}(j)}(\alpha_{n_{I_0}(j)} - \gamma_{max})].
\end{aligned} \tag{3.46}$$

Man beachte, dass nach dieser Umformung von (3.40) zu (3.46) die Terme auf der linken Seite der Ungleichung von der Form

$$c_i^2 \cdot (a_{ii} + \gamma_{max}) \ \textit{bzw.}\ d_i^2 \cdot (b_{ii} + \delta_{max})$$

doppelt summiert werden, wenn für i eine der Bedingungen (3.42), (3.43), (3.44) oder (3.45) mit '$\wedge$' und nicht nur mit '$\vee$' erfüllt ist.
Dieser scheinbare Fehler wird aber dadurch geheilt, dass er auf der rechten Seiten des Ungleichheitszeichens von (3.46) gleichermaßen begangen wird, allerdings mit Termen der Form:

$$c_i^2 \cdot (\alpha_i - \gamma_{max}) \ \textit{bzw.}\ d_i^2 \cdot (\beta_i - \delta_{max}).$$

Es gilt jedoch:

$$\forall i \in \Omega_N : c_i^2 \cdot (a_{ii} + \gamma_{max}) \geq c_i^2 \cdot (\alpha_i - \gamma_{max})$$

sowie

$$\forall j \in \Omega_N : d_j^2 \cdot (b_{jj} + \delta_{max}) \geq d_j^2 \cdot (\beta_j - \delta_{max}),$$

womit (3.46) eine striktere Form von (d. h. hinreichend für) (3.40) ist. Die Ungleichung (3.46) ist wiederum sicher für beliebige $I \subseteq \Omega_N$ erfüllt, wenn sie „summandenweise“ erfüllt ist, d. h.:

$$\begin{aligned}
& \forall i \in I_0 \supseteq \hat{I}_0 : \\
& c^2_{v_{I_0}(i)}(a_{v_{I_0}(i),v_{I_0}(i)} + \gamma_{max}) + c_i^2(a_{ii} + \varepsilon_{v_{I_0}(i),n_{I_0}(i)} \cdot \gamma_{max}) \\
& + c^2_{n_{I_0}(i)}(a_{n_{I_0}(i),n_{I_0}(i)} + \gamma_{max}) + d^2_{v_{J_0}(i)}(b_{v_{J_0}(i),v_{J_0}(i)} + \delta_{max}) \\
& + d^2_{n_{J_0}(i)}(b_{n_{J_0}(i),n_{J_0}(i)} + \delta_{max}) \\
\leq\ & c^2_{v_{I_0}(i)}(\alpha_{v_{I_0}(i)} - \gamma_{max}) + c^2_{n_{I_0}(i)}(\alpha_{n_{I_0}(i)} - \gamma_{max}) + d^2_{v_{J_0}(i)}(\beta_{v_{J_0}(i)} - \delta_{max}) \\
& + d_i^2(\beta_i - \varepsilon_{v_\Omega(i),n_\Omega(i)} \cdot \delta_{max}) + d^2_{n_{J_0}(i)}(\beta_{n_{J_0}(i)} - \delta_{max}),
\end{aligned}$$

$$\begin{aligned}
& \forall j \in J_0 \supseteq \hat{J}_0 : \\
& d^2_{v_{J_0}(j)}(b_{v_{J_0}(j),v_{J_0}(j)} + \delta_{max}) + d_j^2(b_{jj} + \varepsilon_{v_{J_0}(j),n_{J_0}(j)} \cdot \delta_{max}) \\
& + d^2_{n_{J_0}(j)}(b_{n_{J_0}(j),n_{J_0}(j)} + \delta_{max}) + c^2_{v_{I_0}(j)}(a_{v_{I_0}(j),v_{I_0}(j)} + \gamma_{max}) \\
& + c^2_{n_{I_0}(j)}(a_{n_{I_0}(j),n_{I_0}(j)} + \gamma_{max}) \\
\leq\ & d^2_{v_{J_0}(j)}(\beta_{v_{J_0}(i)} - \delta_{max}) + d^2_{n_{J_0}(j)}(\beta_{n_{J_0}(j)} - \delta_{max}) + c^2_{v_{I_0}(j)}(\alpha_{v_{I_0}(j)} - \gamma_{max}) \\
& + c_j^2(\alpha_j - \varepsilon_{v_\Omega(j),n_\Omega(j)} \cdot \gamma_{max}) + c^2_{n_{I_0}(j)}(\alpha_{n_{I_0}(j)} - \gamma_{max}).
\end{aligned}$$

Ist diese Bedingung, welche zugleich die Bedingung des Satzes ist, erfüllt, so folgt:

$$(3.46) \Longrightarrow (3.40) \Longrightarrow (3.35) \Longleftrightarrow (3.34) \Longleftrightarrow (3.33) \Longleftrightarrow (3.32) \Longleftrightarrow (3.31).$$

□

Anmerkung 3.36. *Sind A, B Diagonalmatrizen, vereinfacht sich die Aussage zu der des Diagonalfalls aus Satz 3.20, denn dann erhalten wir für alle $i \in \Omega_N$:*

$$a_{ii} = \alpha_i,\ b_{ii} = \beta_i$$

sowie

$$\gamma_{max} = 0,\ \delta_{max} = 0.$$

Die Bedingung des Satzes 3.35 ist eine Verschärfung der Bedingung aus Satz 3.20 derart, dass jede Menge $I_0 \subseteq \Omega_N$, die Satz 3.35 genügt, auch Satz 3.20 genügt. Diese

Tatsache hat für uns sehr praktischen Nutzen, da wir im Rahmen eines späteren Lösungsalgorithmus nicht alle denkbaren Mengen $I_0 \subseteq \Omega_N$ auf die Bedingung von Satz 3.35 hin testen müssen, sondern nur solche, die auch die (deutlich schneller zu überprüfende) Bedingung des Diagonalfalls aus Satz 3.20 erfüllen.

Anmerkung 3.37. *Eine Menge I_0 wie in Satz 3.35 existiert offensichtlich sicher, wenn alle c_i, d_j für $i \in I_0, j \in \bar{I}_0$ hinreichend klein (bzw. alle c_i, d_j für $i \in \bar{I}_0, j \in I_0$ hinreichend groß) und zudem alle vorkommenden Summanden positiv sind. Letzteres ist bei hinreichend großen Diagonalelementen bzw. bei hinreichend kleinen Nebendiagonalelementen von A und B gewährt.*

Satz 3.38 (Eindeutigkeit von I_0)**.** *Ist zumindest eine der Matrizen A und B echt tridiagonal (d. h. $\gamma_{max} > 0$ oder $\delta_{max} > 0$), dann existiert höchstens eine Menge $I_0 \subseteq \Omega_N$, die der Bedingung in Satz 3.35 genügt.*

Beweis. *Es sei $\gamma_{max} > 0$ und $I_0 \subseteq \Omega_N$ eine Menge, welche die Bedingung in Satz 3.35 erfülle. Dann ist die linke Seite von (3.29) echt größer als die linke Seite von (3.20) und gleichermaßen die rechte Seite von (3.29) echt kleiner als die rechte Seite von (3.20). Mithin ist (3.20) für alle $i \in I_0$ strikt erfüllt, d. h.*

$$\Phi_1(\{i\}) < \Phi_2(\{i\}). \tag{3.47}$$

Analoges Räsonnement führt zur Erkenntnis, dass für alle $j \in \bar{I}_0$

$$\Phi_1(\{j\}) > \Phi_2(\{j\}). \tag{3.48}$$

gilt. Da es neben I_0 keine weitere Teilmenge von Ω_N geben kann, die (3.47) und (3.48) erfüllt, ist I_0 auch die einzige Teilmenge von Ω_N, die Satz 3.35 genügen kann. □

Anmerkung 3.39. *Selbst im Falle der eindeutigen Existenz einer solchen Menge I_0 muss allerdings nicht gelten, dass* (AP2-I) *genau I_0 als Lösung besitzt. Denn Satz 3.35 gibt lediglich ein hinreichendes und kein notwendiges Kriterium für die Globalität einer Lösung von* (AP2-I) *an. Es kann also insbesondere eine globale Lösung vorliegen, selbst wenn Satz 3.35 verletzt ist.*

3.5.2 Konstruktion eines Lösungsalgorithmus

Wir kommen nun zum eigentlichen Ziel des Kapitels: der Konstruktion eines Lösungsalgorithmus für (AP1). Dazu erinnern wir uns noch einmal an Anmerkung 3.11, in welcher die grundsätzliche Struktur unserer Lösungsalgorithmen beschrieben wurde:

1. Finde eine globale Lösung von (AP2-I). Ist sie zulässig für (AP1), so ist sie auch eine globale Lösung von (AP1).

2. Wenn nicht, dann suche unter allen lokalen Lösungen von (AP2-II), die zulässig für (AP1) sind, eine mit kleinstem Zielfunktionswert. Eine solche lokale Lösung existiert sicher und ist auch eine globale Lösung von (AP1).

Für Punkt 1. haben wir soeben eine sinnvolle Strategie entwickelt: Wir testen dazu die Mengen $I \subseteq \Omega_N$, die Satz 3.20 genügen, auf die Bedingung aus Satz 3.35. Ist die Bedingung erfüllt und die entsprechende Lösung zulässig für (AP1), d. h. sind alle Komponenten der entsprechenden Lösung (x, y) negativ oder Null, so haben wir eine Lösung für (AP1) gefunden.

Ist dies nicht der Fall, d. h. ist unsere Lösung von (AP2-I) entweder unzulässig für (AP1) oder konnten wir erst gar keine Lösung für (AP2-I) angeben, so verfahren wir weiter mit Punkt 2.

Dazu ist also unter allen lokalen Lösungen von

$$\textbf{(AP2-II)} \quad \begin{cases} \min\limits_{\substack{I,J\subseteq\Omega_N \\ I\cup J=\Omega_N}} \Phi_1(I) + \Phi_2(J) \,, \end{cases}$$

die zulässig für (AP1) sind, eine mit kleinstem Zielfunktionswert zu finden.

Wir werden im Folgenden zeigen, dass sich ein solches Optimum durch Lösen von nur $\mathcal{O}(N^2)$ Gleichungssystemen ermitteln lässt.

Hierbei werden wir die nachfolgende Definition benötigen.

Definition 3.40. *Für $M \in \mathbb{R}^{N\times N}$, $v \in \mathbb{R}^N$ und $r < s \in \Omega_N$ definieren wir folgende Schreibweisen:*

$$M|_{r,\dots,s} := (m_{i,j})_{i,j=r,\dots,s} \in \mathbb{R}^{(r-s+1)\times(r-s+1)},$$
$$v|_{r,\dots,s} := (v_i)_{i,j=r,\dots,s} \in \mathbb{R}^{r-s+1}.$$

Beispiel 3.41. *Für*

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix}, \; v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

ist

$$M|_{2,\dots,4} = \begin{pmatrix} m_{22} & m_{23} & m_{24} \\ m_{32} & m_{33} & m_{34} \\ m_{42} & m_{43} & m_{44} \end{pmatrix}, \; v|_{2,\dots,4} = \begin{pmatrix} v_2 \\ v_3 \\ v_4 \end{pmatrix}.$$

Definition 3.42. *In Anlehnung an Definition 3.40 schreiben wir für $r < s \in \Omega_N$ und $I \subseteq \{r, ..., s\}$:*

$$\Phi_1(I)|_{r,...,s} = \min_{x \in \mathbb{R}^{s-r+1}} \frac{1}{2}(x - c|_{r,...,s})^T A|_{r,...,s}(x - c|_{r,...,s})$$

und

$$\Phi_2(I)|_{r,...,s} = \min_{y \in \mathbb{R}^{s-r+1}} \frac{1}{2}(y - d|_{r,...,s})^T B|_{r,...,s}(y - d|_{r,...,s}).$$

Die entscheidende Eigenschaft, die es uns ermöglicht $\mathcal{O}(2^N)$ Optimalwerte durch das Lösen von nur $\mathcal{O}(N^2)$ Gleichungssystemen zu ermitteln, soll in den beiden folgenden Sätzen verdeutlicht werden.

Satz 3.43 (Blockweise Funktionsauswertung)**.** *Es seien $r \in \Omega_N$ und $I \subseteq \{1, ..., r-1\}$, $J \subseteq \{r+1, ..., N\}$, dann gilt:*

$$\Phi_1(I \cup \{r\} \cup J) = \Phi_1(I \cup \{r\})|_{1,...,r} + \Phi_1(\{r\} \cup J)|_{r,...,N} - \frac{1}{2}c_r^2 a_{r,r}$$

sowie

$$\Phi_2(I \cup \{r\} \cup J) = \Phi_2(I \cup \{r\})|_{1,...,r} + \Phi_2(\{r\} \cup J)|_{r,...,N} - \frac{1}{2}d_r^2 b_{r,r}.$$

Beweis. *Es seien r, I, J wie in der Angabe und A der Form*

$$A = \begin{pmatrix} \hat{A} & a_{r-1} & \\ a_{r-1}^T & a_{r,r} & a_{r+1}^T \\ & a_{r+1} & \tilde{A} \end{pmatrix}$$

mit

$$\hat{A} = A|_{1,...,r-1},\ \tilde{A} = A|_{r+1,...,N}$$

und

$$a_{r-1} = (0, \ldots, 0, a_{r-1,r})^T,\ a_{r+1} = (a_{r,r+1}, 0, \ldots, 0)^T.$$

Analog sei

$$c = \begin{pmatrix} \hat{c} \\ c_r \\ \tilde{c} \end{pmatrix}$$

mit

$$\hat{c} = c|_{1,...,r-1} \text{ und } \tilde{c} = c|_{r+1,...,N}.$$

Ferner sei

$$x = \begin{pmatrix} \hat{x} \\ x_r \\ \tilde{x} \end{pmatrix} \in \mathbb{R}^N$$

mit

$$\hat{x} \in \mathbb{R}^{r-1} \text{ und } \tilde{x} \in \mathbb{R}^{N-r}$$

die (eindeutige) Lösung von

$$\Phi_1(I \cup \{r\} \cup J) = \min_{x \in \mathbb{R}^N} \frac{1}{2}(x-c)^T A(x-c)$$
$$s.\ t.\ \forall i \in (I \cup \{r\} \cup J) : x_i = 0.$$

Um nun die Aussage des Satzes zu beweisen, zeigen wir zuerst, dass man die Lösung x „blockweise“ berechnen kann. Nach Satz 3.15 ist

$$\begin{aligned}
x &= A^{-1}_{I\cup\{r\}\cup J,\, I\cup\{r\}\cup J} A_{I\cup\{r\}\cup J} c \\
&= \begin{pmatrix} \hat{A}_{I,I} & 0 & \\ 0 & 1 & 0 \\ & 0 & \tilde{A}_{J,J} \end{pmatrix}^{-1} \begin{pmatrix} \hat{A}_I & a_{r-1} & \\ 0 & 0 & 0 \\ & a_{r+1} & \tilde{A}_J \end{pmatrix} \begin{pmatrix} \hat{c} \\ c_r \\ \tilde{c} \end{pmatrix} \\
&\overset{L.\,3.23}{=} \begin{pmatrix} \hat{A}^{-1}_{I,I} & 0 & \\ 0 & 1 & 0 \\ & 0 & \tilde{A}^{-1}_{J,J} \end{pmatrix} \begin{pmatrix} \hat{A}_I & a_{r-1} & \\ 0 & 0 & 0 \\ & a_{r+1} & \tilde{A}_J \end{pmatrix} \begin{pmatrix} \hat{c} \\ c_r \\ \tilde{c} \end{pmatrix} \\
&= \begin{pmatrix} \hat{A}^{-1}_{I,I}\hat{A}_I & \hat{A}^{-1}_{I,I}a_{r-1} & \\ 0 & 0 & 0 \\ & \tilde{A}^{-1}_{J,J}a_{r+1} & \tilde{A}^{-1}_{J,J}\tilde{A}_J \end{pmatrix} \begin{pmatrix} \hat{c} \\ c_r \\ \tilde{c} \end{pmatrix} \\
&= \begin{pmatrix} \hat{A}^{-1}_{I,I}\hat{A}_I\hat{c} + \hat{A}^{-1}_{I,I}a_{r-1}c_r \\ 0 \\ \tilde{A}^{-1}_{J,J}a_{r+1}c_r + \tilde{A}^{-1}_{J,J}\tilde{A}_J\tilde{c} \end{pmatrix} = \begin{pmatrix} \hat{x} \\ 0 \\ \tilde{x} \end{pmatrix}.
\end{aligned}$$

Offensichtlich ist aber $(\hat{x}^T, 0)$ die Lösung von

$$\Phi_1(I \cup \{r\})|_{1,\ldots r} = \min_{x \in \mathbb{R}^r} \frac{1}{2}(x - c|_{1,\ldots,r})^T A|_{1,\ldots,r}(x - c|_{1,\ldots,r})$$

$$s.\ t.\ \forall i \in (I \cup \{r\}) : x_i = 0,$$

denn selbige berechnet sich zu

$$\begin{pmatrix} \hat{A}_{I,I} & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{A}_I & a_{r-1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c} \\ c_r \end{pmatrix} = \begin{pmatrix} \hat{A}_{I,I}^{-1}\hat{A}_I\hat{c} + \hat{A}_{I,I}^{-1}a_{r-1}c_r \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix}.$$

Gleichermaßen ist $(0, \tilde{x}^T)$ *die Lösung von*

$$\Phi_1(\{r\} \cup J)|_{r,\dots,N}.$$

Abschließend müssen wir noch untersuchen, wie sich der Wert der Zielfunktion zusammensetzt:

$$\begin{aligned}
\Phi_1(I \cup \{r\} \cup J) &= \frac{1}{2}(x-c)^T A(x-c) \\
&= \frac{1}{2}(\hat{x}^T - \hat{c}^T,\ 0 - c_r,\ \tilde{x}^T - \tilde{c}^T) \begin{pmatrix} \hat{A} & a_{r-1} & \\ a_{r-1}^T & a_{r,r} & a_{r+1}^T \\ & a_{r+1} & \tilde{A} \end{pmatrix} \begin{pmatrix} \hat{x} - \hat{c} \\ 0 - c_r \\ \tilde{x} - \tilde{c} \end{pmatrix} \\
&= \frac{1}{2}(\hat{x}^T - \hat{c}^T,\ -c_r) \begin{pmatrix} \hat{A} & a_{r-1} \\ a_{r-1}^T & a_{r,r} \end{pmatrix} \begin{pmatrix} \hat{x} - \hat{c} \\ -c_r \end{pmatrix} \\
&\quad + \frac{1}{2}(-c_r,\ \tilde{x}^T - \tilde{c}^T) \begin{pmatrix} a_{r,r} & a_{r+1}^T \\ a_{r+1} & \tilde{A} \end{pmatrix} \begin{pmatrix} -c_r \\ \tilde{x} - \tilde{c} \end{pmatrix} - \frac{1}{2}c_r a_{r,r} c_r \\
&= \Phi_1(I \cup \{r\})|_{1,\dots,r} + \Phi_1(\{r\} \cup J)|_{r,\dots,N} - \frac{1}{2}c_r^2 a_{r,r}.
\end{aligned}$$

□

Satz 3.44. *Um alle (lokalen) Lösungen von* (AP2-II) *zu bestimmen, sind* $\mathcal{O}(N^2)$ *(genauer:* $2N^2 - 2N$*) lineare Gleichungssysteme zu lösen.*

Beweis. *Wir betrachten dazu nochmals* (AP2-II)*:*

$$\min_{\substack{I,J \subseteq \Omega_N \\ I \cup J = \Omega_N}} \Phi_1(I) + \Phi_2(J)$$

Um alle lokalen Lösungen zu bestimmen, muss also $\Phi_1(I)$ *und auch* $\Phi_2(I)$ *für alle* $I \subseteq \Omega_N$ *berechnet werden. Wir beweisen daher mittels vollständiger Induktion über* N*, dass es* $N^2 - N$ *lineare Gleichungssysteme für* Φ_1 *zu lösen gibt.*

Induktionsanfang: $N = 1$
In diesem Fall ist $\Omega_N = \{1\}$ und damit $\mathcal{P}(\Omega_N) = \{\emptyset, \Omega_N\}$. Es gibt also kein lineares Gleichungssystem zu lösen, da die Lösung von $\Phi_1(\emptyset)$ trivialerweise der Vektor $x = c$ und die Lösung von $\Phi_1(\Omega_N)$ der Nullvektor ist.

Induktionsschritt: $N \rightsquigarrow N + 1$
Wir erweitern nun die Indexmenge auf

$$\Omega_N \cup \{N + 1\}.$$

Offensichtlich gilt für alle $J \subseteq (\Omega_N \cup \{N + 1\})$ entweder

$$J \subseteq \Omega_N \tag{3.49}$$

oder aber

$$\exists! I \subseteq \Omega_N : J = I \cup \{N + 1\}. \tag{3.50}$$

Zu (3.49):
In diesem Fall sei $r \in \Omega_N$ das größte Element in J und $I = J \setminus \{r\}$. Damit können wir Satz 3.43 anwenden und erhalten:

$$\Phi_1(J) = \Phi_1(I \cup \{r\}) = \Phi_1(I \cup \{r\})|_{1,\dots,r} + \Phi_1(\{r\})|_{r,\dots,N+1} - \frac{1}{2} c_r^2 a_{r,r}.$$

Der erste Summand (und damit der erste Teil des Lösungsvektors) $\Phi_1(I \cup \{r\})|_{1,\dots,r}$ wurde bereits im Rahmen der Induktionsannahme berechnet. Es bleibt also für $r = 1, \dots, N$ die Lösung von $\Phi_1(\{r\})|_{r,\dots,N+1}$ zu ermitteln.

Zu (3.50):
In diesem Fall sei $r \in \Omega_N$ das größte Element in I und $\tilde{I} = I \setminus \{r\}$. Wir wenden wieder Satz 3.43 zur blockweisen Funktionsauswertung an:

$$\Phi_1(J) = \Phi_1(\tilde{I} \cup \{r\} \cup \{N+1\}) = \Phi_1(\tilde{I} \cup \{r\})|_{1,\dots,r} + \Phi_1(\{r\} \cup \{N+1\})|_{r,\dots,N+1} - \frac{1}{2} c_r^2 a_{r,r}.$$

Wie schon im Falle von (3.49) ist lediglich der zweite Summand für $r = 1, \dots, N$ zu berechnen.

Insgesamt gibt es also im Induktionsschritt $2N$ lineare Gleichungssysteme zu lösen, woraus folgt:

$$N^2 - N + 2N = N^2 + 2N + 1 - N - 1 = (N + 1)^2 - (N + 1).$$

□

Definition 3.45. *Hat der zu $\Phi_1(I)$, $I \subseteq \Omega_N$, gehörende Lösungsvektor (x, y) mit $x, y \in \mathbb{R}^N$ ausschließlich nicht-positive Komponenten, so bezeichnen wir $\Phi_1(I)$ als „zulässig für (AP1)“ oder schlicht „zulässig“, andernfalls als „unzulässig für (AP1)“ oder schlicht „unzulässig“. Analoges gilt für Φ_2.*

Für den letzten Satz, der uns zeigen wird, wie eine globale Lösung von (AP1) berechnet werden kann, zerlegen wir im Vorfeld die Potenzmenge $\mathcal{P}(\Omega_N)$ in paarweise disjunkte Teilmengen:

$$\Lambda^{(1)} := \{I \subseteq \Omega_N \mid \Phi_1(I) \text{ zulässig } \wedge\ \Phi_2(\bar{I}) \text{ zulässig}\},$$
$$\Lambda^{(2)} := \{I \subseteq \Omega_N \mid \Phi_1(I) \text{ zulässig } \wedge\ \Phi_2(\bar{I}) \text{ unzulässig}\},$$
$$\Lambda^{(3)} := \{I \subseteq \Omega_N \mid \Phi_1(I) \text{ unzulässig } \wedge\ \Phi_2(\bar{I}) \text{ zulässig}\},$$
$$\Lambda^{(4)} := \{I \subseteq \Omega_N \mid \Phi_1(I) \text{ unzulässig } \wedge\ \Phi_2(\bar{I}) \text{ unzulässig}\}.$$

Offensichtlich ist

$$\mathcal{P}(\Omega_N) = \Lambda^{(1)} \cup \Lambda^{(2)} \cup \Lambda^{(3)} \cup \Lambda^{(4)}.$$

Satz 3.46. *Der globale Optimalwert von* (AP1) *ist gleich*

$$\min \left[\min_{I \in \Lambda^{(1)}} \Phi_1(I) + \Phi_2(\bar{I}),\ \min_{\substack{I \in \Lambda^{(2)} \\ J \in \Lambda^{(3)} \\ I \cup J = \Omega_N}} \Phi_1(I) + \Phi_2(J) \right].$$

Beweis. *Nach Satz 3.10 ist jede globale Lösung von* (AP1) *auch eine lokale Lösung von* (AP2-II)*. Unter den lokalen Lösungen von* (AP2-II) *kommen dabei jedoch nur solche Mengen $I \subseteq \Omega_N$ bzw. $J \subseteq \Omega_N$ in Frage, für welche $\Phi_1(I)$ bzw. $\Phi_2(J)$ zulässig für* (AP1) *ist. Sind für $I \subseteq \Omega_N$ sowohl $\Phi_1(I)$ als auch $\Phi_2(\bar{I})$ zulässig für* (AP1)*, so gilt nach Lemma 3.9*

$$\forall J \subseteq \Omega_N, I \cup J = \Omega_N : \Phi_2(J) \geq \Phi_2(\bar{I}),$$

da für solche J offensichtlich

$$\bar{I} \subseteq J$$

ist. □

Insgesamt ergibt sich der nachfolgende Lösungsalgorithmus für (AP1) im Tridiagonalfall.

BEGIN AP1TRIDIAG

for $i = 1$ to N **do**
if $\Phi_1(\{i\}) \leq \Phi_2(\{i\})$ **then**
$I_0 \leftarrow i$;
else
$\bar{I}_0 \leftarrow i$;
end if
end for
if $\Phi_1(I_0)$ zulässig und $\Phi_2(\bar{I}_0)$ zulässig und I_0 erfüllt Satz 3.35 **then**
STOP;
else
Berechne alle lokalen Minimalwerte von Φ_1 und Φ_2 nach Satz 3.44 unter Erstellung von $\Lambda^{(1)}, \Lambda^{(2)}, \Lambda^{(3)}$;
Ermittle

$$I^{(1)} = \arg \min_{I \in \Lambda^{(1)}} \Phi_1(I) + \Phi_2(\bar{I});$$

Ermittle

$$(I^{(2)}, J^{(3)}) = \arg \min_{\substack{I \in \Lambda^{(2)} \\ J \in \Lambda^{(3)} \\ I \cup J = \Omega_N}} \Phi_1(I) + \Phi_2(J);$$

Die Lösung von (AP1) ergibt sich über

$$\min \left[\Phi_1(I^{(1)}) + \Phi_2(\overline{I^{(1)}}),\ \Phi_1(I^{(2)}) + \Phi_2(J^{(3)}) \right];$$

end if

END AP1TRIDIAG

Kapitel 4

Numerische Ergebnisse

Um ein Gefühl für die Güte der im Rahmen dieser Arbeit entwickelten Algorithmen zu erhalten, wollen wir selbige in diesem Kapitel auf reelle Testdaten anwenden und auch mit herkömmlichen Optimierungsalgorithmen vergleichen. Als Testdaten verwenden wir die aus der Einleitung bereits bekannte verhallte Äußerung „one, one, eight“ (vgl. Abbildung 4.1).

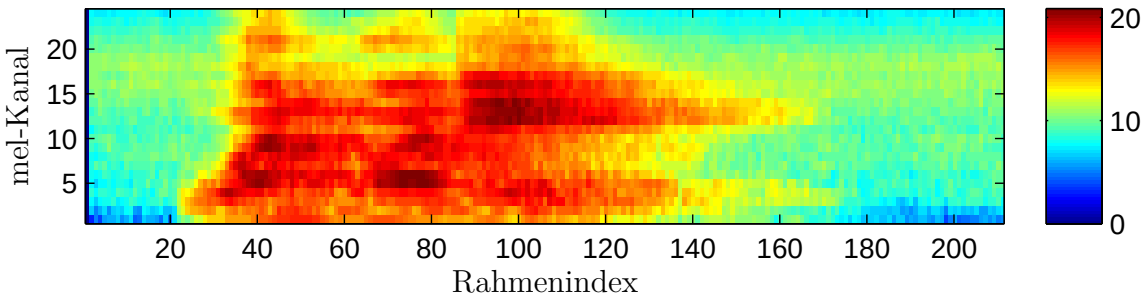


Abbildung 4.1: Logmelspec-Darstellung der (verhallten) Äußerung „one, one, eight“.

4.1 Vergleich von AP1DIAG und AP1TRIDIAG

Wegen der hohen Anzahl lokaler Lösungen (2^N) konzentrieren wir uns hier auf die berechneten Optimalwerte. Ein Vergleich der Optimalstellen ist nur wenig aussagekräftig, da zwei stark unterschiedliche Lösungen gleichermaßen globale Lösungen sein können.

Verglichen werden die Algorithmen AP1DIAG und AP1TRIDIAG für verschiedene Parameter. Als Referenzalgorithmus verwenden wir ein Brute-Force-Verfahren, welches für jede Art von Matrix immer eine globale Lösung (durch „Ausprobieren“ aller lokalen) von (AP1) ermittelt.

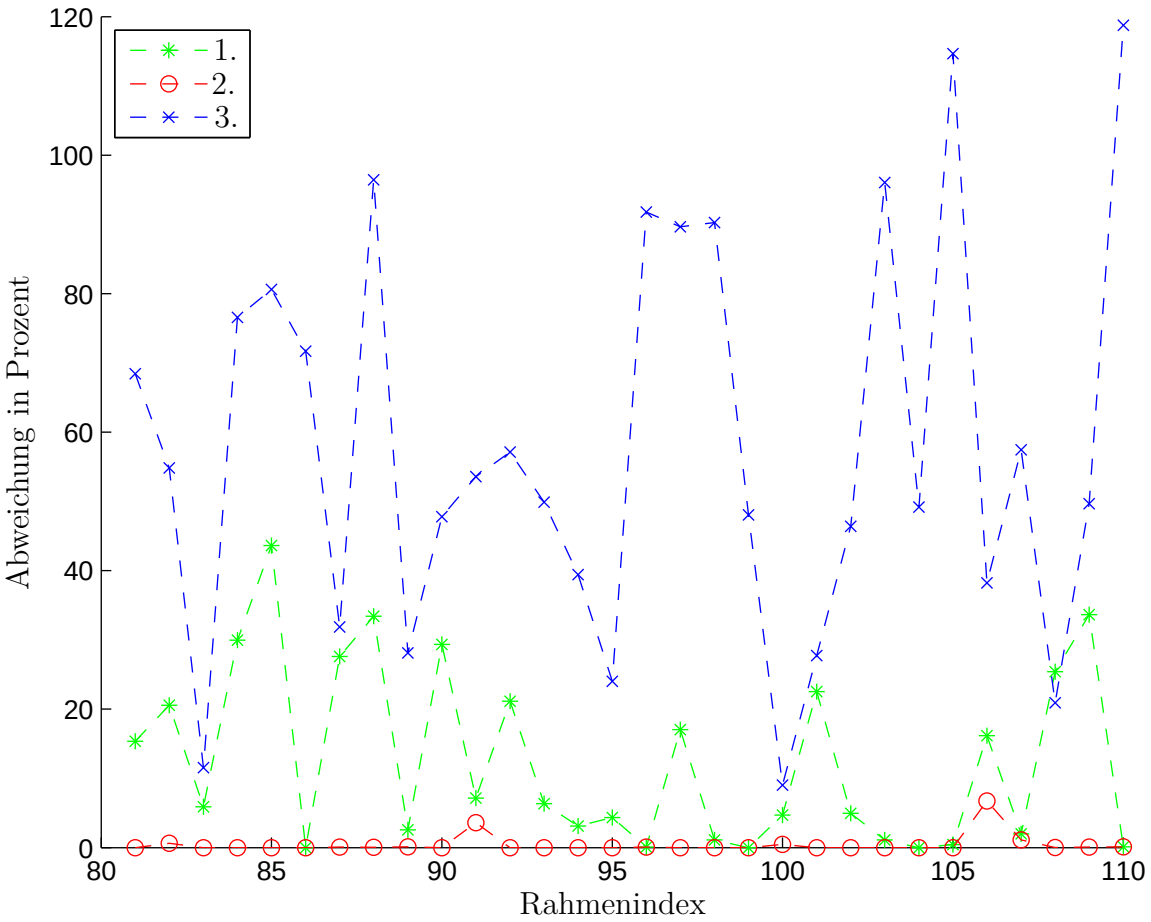
Als Testdaten wurden Submatrizen A und B der Rahmen 81 bis 110 der Testäußerung „one, one, eight“ verwendet. Die Scheitelvektoren c und d wurden mit der random-Funktion von Matlab zufällig gleichverteilt in einem Intervall der Breite 10 erstellt. Die Qualität der gefundenen Lösungen von AP1DIAG hängt stark von der Wahl der Mittelwerte der Gleichverteilung ab.

In Abbildung 4.2 wurde als Mittelwert der Gleichverteilung der Scheitelvektoren jeweils 3 gewählt. AP1DIAG ermittelte in diesem Fall (und auch in vergleichbaren Fällen) zumeist eine gute lokale Lösung, aber selten eine globale. Demgegenüber erzielte AP1DIAG sehr gute Ergebnisse bei negativen oder aber bei Mittelwerten unterschiedlichen Vorzeichens, wie in Abbildung 4.3 zu sehen ist. Dies lässt sich dadurch begründen, dass AP1DIAG besonders gute Lösungen für (AP2-I) berechnet, welche oft auch gültig für (AP1) sind, wenn nicht beide Scheitelkoordinaten positiv sind. Sind jedoch die Scheitelkoordinaten überwiegend beide positiv (im Sinne der Gleichverteilung), so treten vermehrt Fälle auf, in denen die globale Lösung von (AP2-I) unzulässig für (AP1) ist. In diesen Fällen setzt AP1DIAG die ungültigen Komponenten zu Null, was mit einem zusätzlichen Fehler verbunden ist. Dieser Fehler wächst allerdings mit steigenden Mittelwerten der Gleichverteilung der Scheitelvektoren nicht beliebig, da bei hinreichend großen (positiven) Mittelwerten der Nullvektor die einzige Lösung von (AP1) ist.

In beiden Abbildungen 4.2 und 4.3 ist als zusätzliche Kurve der Verlauf des arithmetischen Mittels über alle lokalen Minimalwerte von (AP1) abgedruckt. Damit soll aufgezeigt werden, dass AP1DIAG nicht etwa deshalb „gute“ lokale Lösungen berechnet, weil es keine „schlechten“ lokalen Lösungen gibt.

In allen Fällen ermittelte AP1TRIDIAG eine sehr gute Approximation der globalen Lösung von (AP1). Dies ist darauf zurückzuführen, dass AP1TRIDIAG im schlimmsten Fall, ähnlich wie das Brute-Force-Verfahren, alle lokalen Lösungen von (AP2-II) betrachtet, jedoch selbige dabei nur approximativ unter der Annahme berechnet, die Matrizen A, B seien tridiagonal. Da die tatsächliche Form von A und B der Tridiagonalgestalt sehr nahe kommt, erzielt AP1TRIDIAG unabhängig von der Scheitellage sehr gute Ergebnisse.

Wirft man einen Blick auf die Laufzeiten, so wird deutlich, dass sich AP1TRIDIAG trotz des eventuell notwendigen „Durchprobierens“ aller lokalen Lösungen stark vom Brute-Force-Verfahren absetzt (vgl. Tabelle 4.1).

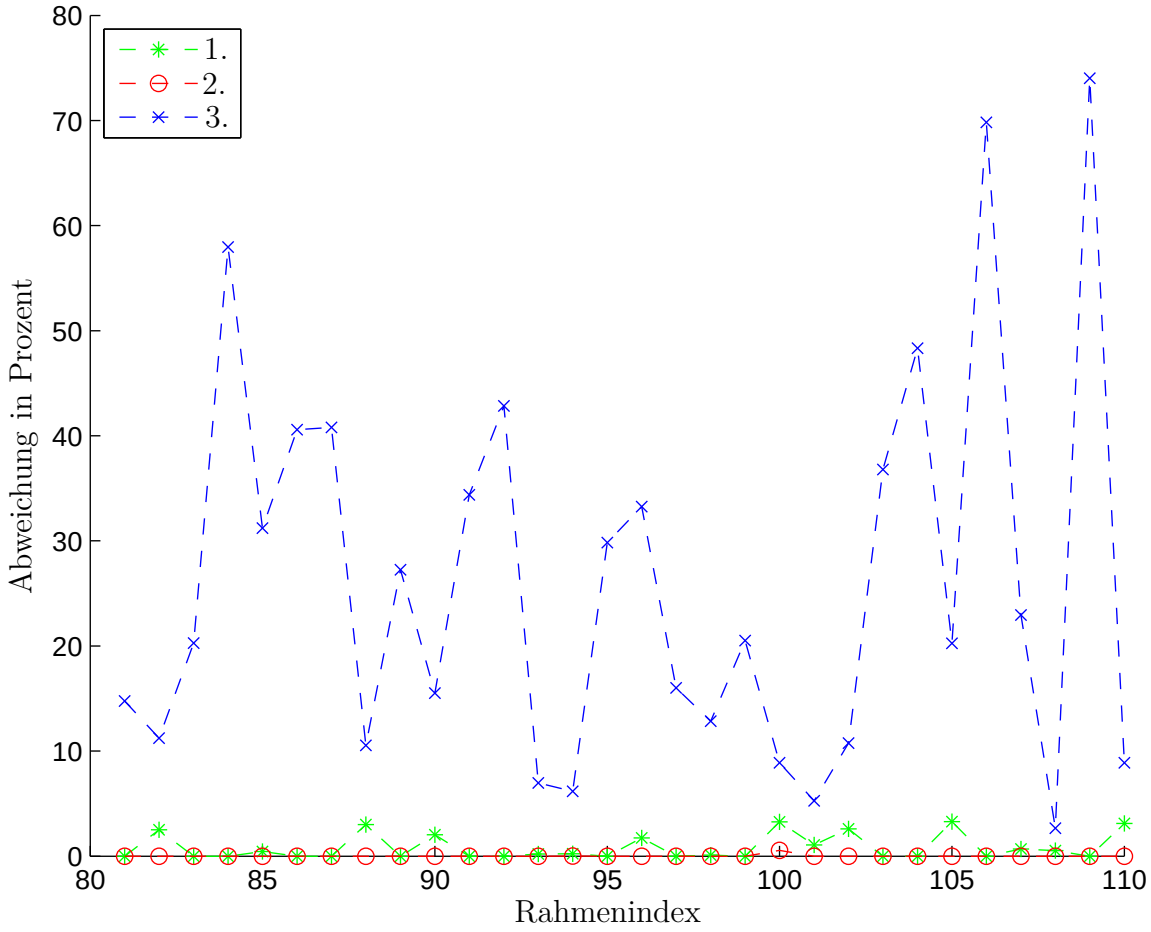


Der 1. Graph stellt den von AP1DIAG und der 2. den von AP1TRIDIAG ermittelten Minimalwert dar. Der 3. Graph beschreibt den Verlauf des arithmetischen Mittels über alle lokalen Minimalwerte von (AP1).

Abbildung 4.2: Abweichung vom globalen Minimalwert der Zielfunktion.

Algorithmus	**Laufzeit**	**Programmiersprache**
AP1DIAG	0.56 msec	Matlab
AP1TRIDIAG	73 msec	Matlab
Brute-Force	90 sec	Matlab

Tabelle 4.1: Laufzeiten für Rahmen 81 und $N = 12$



Der 1. Graph stellt den von AP1DIAG und der 2. den von AP1TRIDIAG ermittelten Minimalwert dar. Der 3. Graph beschreibt den Verlauf des arithmetischen Mittels über alle lokalen Minimalwerte von (AP1).

Abbildung 4.3: Abweichung vom globalen Minimalwert der Zielfunktion.

4.2 Vergleich enthallter Sprachsignale

In diesem Kapitel wurden verschiedene Verfahren auf das gesamte Beispielsignal „one, one, eight“ zur merkmalbasierten Enthallung bei bekannten Mittelwerten angewendet. In Abbildung 4.4 sind die Ergebnisse für IPOPT, AP1DIAG, AP1TRIDIAG und FMINCON zu sehen. Zum Vergleich sind auch die Mittelwerte der Äußerung „one, one, eight“ abgedruckt, die dem Spracherkenner statisch aufgrund des Trainings vorliegen.

Man erkennt deutlich, dass alle vier Algorithmen ähnlich gute Ergebnisse erzielen, sich jedoch signifikant in den Laufzeiten unterscheiden (vgl. Tabelle 4.2). Insbesondere ist festzustellen, dass trotz des hohen Aufwands des Algorithmus AP1TRIDIAG, dieser kein augenscheinlich (im Sinne der logmelspec-Darstellung) besseres Endergebnis als z. B. AP1DIAG erzielt. Aufgrund dieses Aspektes und der hervorragenden Geschwindigkeit des Algorithmus AP1DIAG scheint selbiger besonders gut für den praktischen Einsatz geeignet zu sein. Es ist darüber hinaus sogar davon auszugehen, dass sich die Laufzeit von AP1DIAG weiter verbessern ließe, realisierte man seine Implementierung beispielsweise in der Programmiersprache C.

Algorithmus	**Laufzeit**	**Programmiersprache**
IPOPT	6.53 sec	C
AP1DIAG	0.167 sec	Matlab
AP1TRIDIAG	22.4 h	Matlab
FMINCON	90.0 sec	Matlab

Tabelle 4.2: Laufzeiten zu Abbildung 4.4 (Rahmen 1 bis 211, $N = 24$)

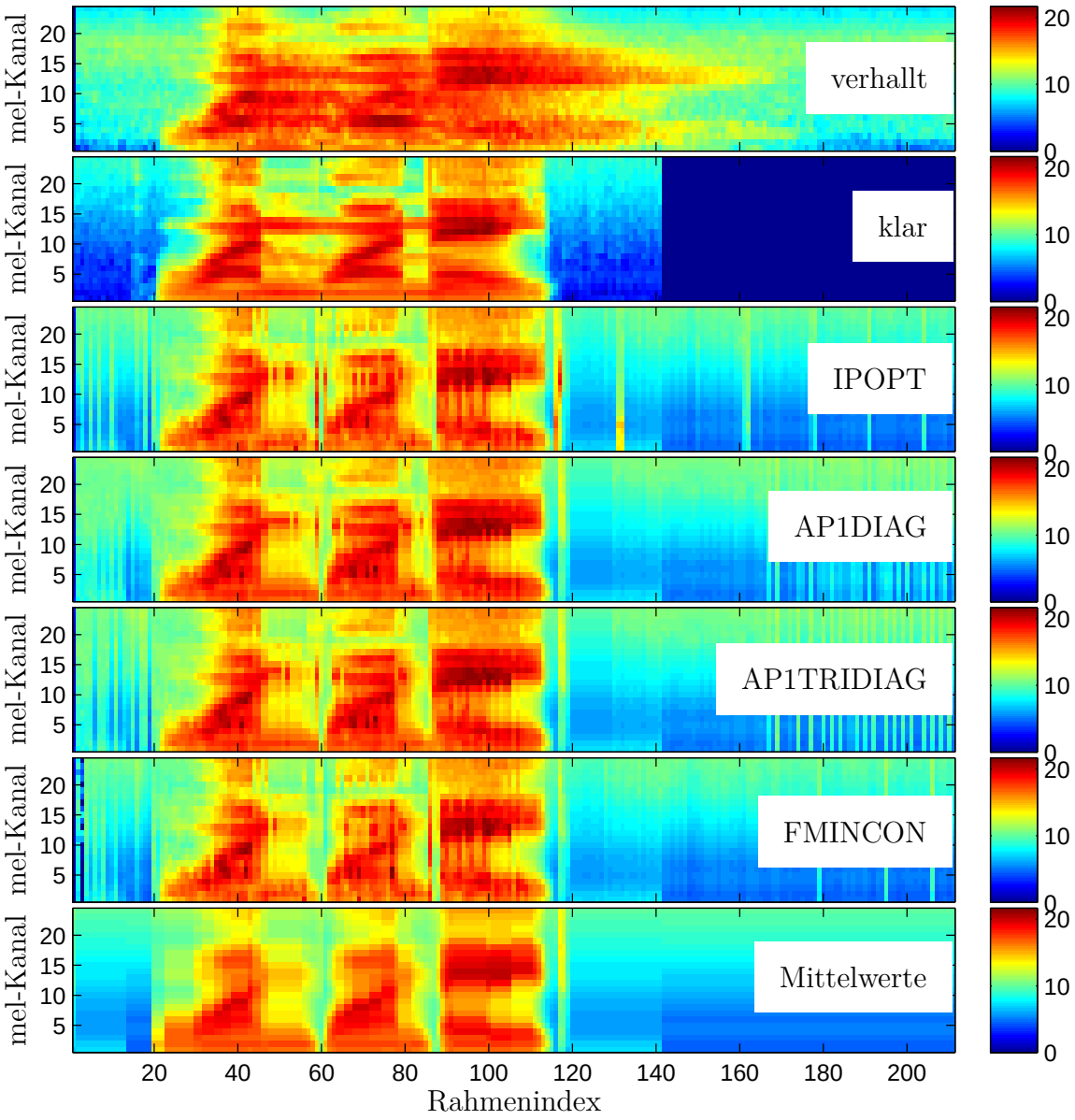


Abbildung 4.4: Vergleich der enthallten Äußerungen im logmelspec-Bereich.

4.3 Korrelationsbetrachtungen

Zur Rechtfertigung des REMOS-Konzeptes unter Verwendung des Algorithmus AP1DIAG ist es entscheidend, dass die enthallten Ergebnisse das klare Sprachsignal besser annähern als die in Abbildung 4.4 zu sehenden Mittelwerte der Äußerung.

Zum Abschluss dieser Arbeit zeigen wir daher noch die Korrelation zwischen den klaren und den durch AP1DIAG enthallten Merkmalvektoren.

Dazu sind in Abbildung 4.5 die logmelspec-Darstellungen der unverhallten Testäußerung und der mittels AP1DIAG enthallten Äußerung mittelwertbereinigt zu sehen. Man erkennt deutlich, dass die Lösung von AP1DIAG Gemeinsamkeiten mit der klaren Äußerung aufweist. Relevant sind dabei vor allem die Rahmen 20 bis 114, da es sich bei diesen nicht um Sprechpausen handelt.

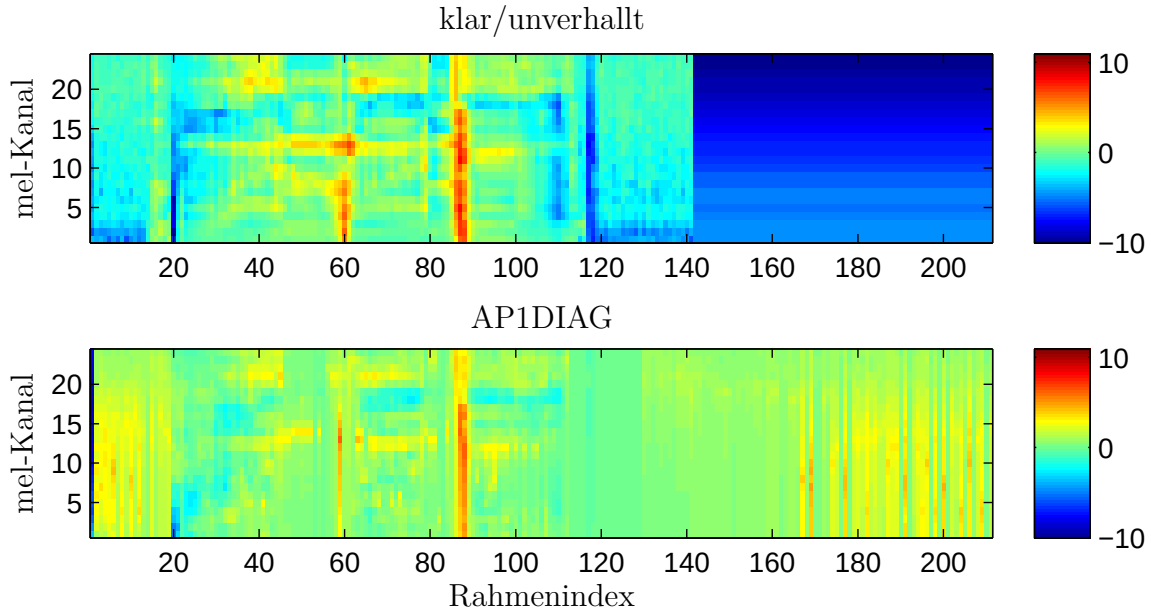
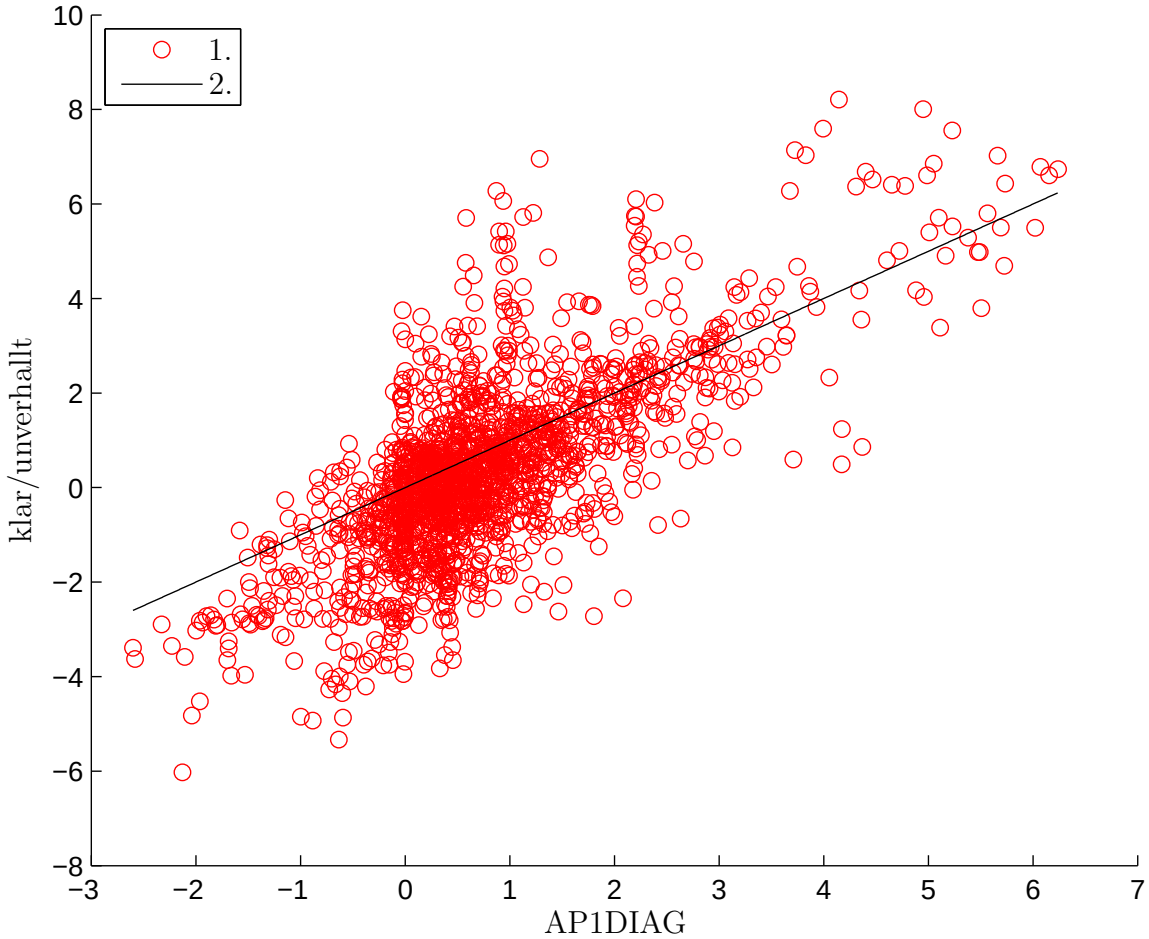


Abbildung 4.5: Mittelwertfreie logmelspec-Darstellungen der Äußerung „one, one, eight“.

Abbildung 4.6 zeigt ein Streudiagramm der Rahmen 20 bis 114 über alle mel-Kanäle: Auf der x-Achse sind die mittelwertbereinigten Ergebnisse von AP1DIAG gegen die mittelwertfreie klare Merkmalvektorfolge auf der y-Achse angetragen. Wie man sieht, ist die Punktwolke tendenziell entlang der Ursprungsgerade ausgerichtet, was ebenfalls auf eine Korrelation beider Größen hindeutet.



Bei der roten Punktmenge (1.) handelt es sich um den Scatterplot. Zum Vergleich ist zusätzlich die winkelhalbierende Ursprungsgerade (2.) zu sehen.

Abbildung 4.6: Streudiagramm der logmelspec-Darstellung der Äußerung „one, one, eight“.

Literaturverzeichnis

[1] G. Fischer. *Lineare Algebra*. Vieweg, 2000.

[2] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, 1991.

[3] C. Geiger und C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, 1999.

[4] C. Geiger und C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, 2002.

[5] W. Hackbusch, H. R. Schwarz und E. Zeidler. *Teubner - Taschenbuch der Mathematik: Teil I*. Teubner, 2003.

[6] X. Huang, A. Acero und H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[7] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[8] E. Polak. *Optimization: Algorithms and Consistent Approximations*. Springer, 1997.

[9] A. Sehr und W. Kellermann. Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments. In E. Hänsler und G. Schmidt, Hrsg., *Topics in Speech and Audio Processing in Adverse Environments*, Seiten 679–728. Springer, Berlin, 2008.

Danksagung

Ich bedanke mich an dieser Stelle in aller Herzlichkeit bei Herrn Dr. Gugat, Herrn Prof. Kellermann und Herrn Sehr für die ausgezeichnete Betreuung, ihren Einsatz und die vielen Stunden, die sie mir widmeten.

Ein besonderer Dank geht auch an meine Familie und Freunde, die der Entwicklung dieser Arbeit beiwohnten.

Schließlich möchte ich diese Gelegenheit nutzen, um Herrn Prof. Keller und Herrn Prof. Ulmer (Univ. Rennes) für ihr offenes Ohr und ihre Unterstützung bei der Gestaltung meines deutsch-französischen Doppelmasterstudiums zu danken.

Erklärung des Verfassers

Hiermit erkläre ich, dass ich diese Arbeit selbstständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe.

Ferner erkläre ich, dass diese Arbeit öffentlich zugänglich gemacht werden darf.

Erlangen, den 17. September 2009 Roland Maas

Lebenslauf des Verfassers

Roland Maas, geboren 1984, absolvierte 2004 sein Abitur am Christoph-Jacob-Treu-Gymnasium in Lauf an der Pegnitz. Im Anschluss studierte er Technomathematik in den Jahren bis 2009 an der Universität Erlangen und der Université de Rennes im Rahmen eines Doppelmasterprogramms der Deutsch-Französischen Hochschule. Im März 2005 wurde er in die Studienstiftung des deutschen Volkes und im Mai 2007 in das Hölderlin-Programm der Siemens Management Consulting aufgenommen.