

Jochen Werner

Numerische Mathematik 2

Eigenwertaufgaben,
lineare Optimierungsaufgaben,
unrestriktierte Optimierungs-
aufgaben

vieweg studium

Aufbaukurs Mathematik

Jochen Werner

Numerische Mathematik 2

vieweg studium

Aufbaukurs Mathematik

Herausgegeben von Gerd Fischer

Manfredo P. do Carmo

Differentialgeometrie von Kurven und Flächen

Wolfgang Fischer / Ingo Lieb

Funktionentheorie

Wolfgang Fischer / Ingo Lieb

Ausgewählte Kapitel aus der Funktionentheorie

Otto Forster

Analysis 3

Manfred Knebusch / Claus Scheiderer

Einführung in die reelle Algebra

Ulrich Krengel

Einführung in die Wahrscheinlichkeitstheorie und Statistik

Alexander Prestel

Einführung in die mathematische Logik und Modelltheorie

Ernst Kunz

Algebra

Jochen Werner

Numerische Mathematik 1 und 2

Joachim Hilgert und Karl-Hermann Neeb

Lie-Gruppen und Lie-Algebren

Advanced Lectures in Mathematics

Herausgegeben von Gerd Fischer

Johann Baumeister

Stable Solution of Inverse Problems

Manfred Denker

Asymptotic Distribution Theory in Nonparametric Statistics

Alexandru Dimca

Topics on Real and Complex Singularities

An Introduction

Francesco Guardado, Patrizia Macri und Alessandro Tancredi

Topics on Real Analytic Spaces

Heinrich von Weizsäcker und Gerhard Winkler

Stochastic Integrals

An Introduction

Jochen Werner

Optimization

Theory and Applications

Jochen Werner

Numerische Mathematik

Band 2: Eigenwertaufgaben,
lineare Optimierungsaufgaben,
unrestringierte Optimierungsaufgaben

Mit 8 Abbildungen und 122 Aufgaben



Prof. Dr. Jochen Werner
Institut für Numerische und Angewandte Mathematik
Georg-August-Universität Göttingen
Lutzestraße 16-18
D-3400 Göttingen

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Werner, Jochen:

Numerische Mathematik / Jochen Werner. Braunschweig;
Wiesbaden: Vieweg
Bd. 2. Eigenwertaufgaben, lineare Optimierungsaufgaben,
unrestriktive Optimierungsaufgaben: mit 122 Aufgaben. -
1992
(Vieweg-Studium; 33: Aufbaukurs Mathematik)
ISBN 978-3-528-07233-9 ISBN 978-3-663-07714-5 (eBook)
DOI 10.1007/978-3-663-07714-5

NE: GT

Alle Rechte vorbehalten

© Springer Fachmedien Wiesbaden 1992

Ursprünglich erschienen bei Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden 1992



Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Gedruckt auf säurefreiem Papier

ISBN 978-3-528-07233-9

Vorwort

Die Zielsetzung in diesem Buch ist im wesentlichen dieselbe wie die zum Vorgänger Numerische Mathematik I. Und zwar wird eine möglichst gut lesbare Darstellung zur numerischen Behandlung gewisser „Grundaufgaben“ angestrebt, die beiden Worten im Titel „Numerische Mathematik“ gerecht zu werden versucht. Nachdem im ersten Band über

- Lineare Gleichungssysteme,
- Nichtlineare Gleichungssysteme,
- Interpolation,
- Numerische Integration

berichtet wurde, folgen jetzt Kapitel über

- Eigenwertaufgaben,
- Lineare Optimierungsaufgaben,
- Unrestringierte Optimierungsaufgaben.

Wie schon im Vorwort zum ersten Band betont wurde, ist Wert darauf gelegt worden, daß die einzelnen Kapitel weitgehend unabhängig voneinander gelesen werden können. Allerdings werden in diesem Buch Kenntnisse aus Kapitel 1 über lineare Gleichungssysteme vorausgesetzt. Eine andere Anordnung des Stoffes wäre denkbar, vom systematischen Standpunkt aus gesehen vielleicht sogar sinnvoller gewesen. So bilden die Kapitel über lineare Gleichungssysteme, Eigenwertaufgaben und lineare Optimierungsaufgaben einen Block, den man der numerischen, linearen Algebra zuordnen kann. Auch die Kapitel über nichtlineare Gleichungssysteme und unrestringierte Optimierungsaufgaben gehören vom Inhalt her eigentlich zusammen, was für die Kapitel über Interpolation und numerische Integration fast selbstverständlich und bei der hier gewählten Aufteilung auch der Fall ist.

Wieder ist versucht worden, die auftretenden Algorithmen so zu formulieren, daß ihre Umsetzung in ein Computer-Programm i. allg. keine größeren Schwierigkeiten bereiten dürfte. Numerische Beispiele finden sich fast nur in den Aufgaben, die angegebenen Ergebnisse dienen der Kontrolle.

Bei der Lektüre des Inhaltsverzeichnisses wird auffallen, daß dem letzten Kapitel über unrestringierte Optimierungsaufgaben verhältnismäßig viel Platz zugestanden wird. Hierfür waren mehrere Gründe maßgeblich. Der unwichtigste (wenn

auch vielleicht entscheidende) Grund besteht darin, daß die numerische Behandlung unrestringierter Optimierungsaufgaben zu meinen besonderen Interessengebieten gehört. Wichtiger ist, daß der Leser wenigstens in einem Kapitel an aktuelle, neuere Entwicklungen herangeführt werden sollte. Zwar gehört auch eine Darstellung des Karmarkar-Verfahrens im Kapitel über lineare Optimierungsaufgaben nicht zum „Standard“ in Lehrbüchern über numerische Mathematik, was im abgeschwächten Maße auch über die Behandlung der schnellen Fourier-Transformation, B-Splines und die Darstellung von Kurven in Kapitel 3 im ersten Band gilt. Aber nur das letzte Kapitel versucht, etwas mehr als einen Überblick zur numerischen Behandlung einer „Grundaufgabe“ zu bieten. In ihm werden Ergebnisse bewiesen (etwa zur globalen und lokalen Konvergenz des BFGS-Verfahrens oder zur Konvergenz von Trust-Region-Verfahren bei glatten bzw. halbglatten Optimierungsaufgaben), die bisher in der Originalliteratur eher versteckt sind. Im Rahmen eines Kurses über numerische Mathematik eignet sich die numerische Behandlung unrestringierter Optimierungsaufgaben besonders gut als vertiefendes Spezialgebiet, weil hier Kenntnisse aus den Kapiteln über lineare und nichtlineare Gleichungssysteme sowie über lineare Optimierungsaufgaben benutzt werden können.

Wieder danke ich vor allem den Hörern meiner Vorlesungen, die durch konstruktive Kritik die Darstellung beeinflußt haben. Ferner danke ich Horst Karaschewski und Martin Petry, die Teile des Manuskripts gelesen und mich auf einige Fehler hingewiesen haben. Mein besonderer Dank gilt Thomas Bannert, der die letzten beiden Kapitel gelesen und zahlreiche, von mir fast immer aufgegriffene, Verbesserungsvorschläge gemacht hat.

Göttingen, im August 1991

Jochen Werner

Inhaltsverzeichnis

5 Eigenwertaufgaben	1
5.1 Einige theoretische Grundlagen	2
5.1.1 Der Satz von Gershgorin	2
5.1.2 Der Satz von Bauer-Fike	6
5.1.3 Variationsprinzipien für Eigenwerte hermitescher Matrizen	9
5.1.4 Der Satz von Schur	12
Aufgaben	14
5.2 Das QR-Verfahren	18
5.2.1 Die Transformation einer Matrix auf Hessenberg-Form	18
5.2.2 Die QR-Zerlegung einer Hessenberg-Matrix	24
5.2.3 Vektoriteration nach v. Mises	29
5.2.4 Inverse Iteration nach Wielandt	30
5.2.5 Die Konvergenz des einfachen QR-Verfahrens	32
5.2.6 Das QR-Verfahren mit Shifts	40
Aufgaben	46
5.3 Eigenwertaufgaben für symmetrische Matrizen	52
5.3.1 Das Jacobi-Verfahren	52
5.3.2 Das Bisektions-Verfahren	56
5.3.3 Das QR-Verfahren für symmetrische Matrizen	61
5.3.4 Die Berechnung der Singulärwertzerlegung	65
Aufgaben	73
6 Lineare Optimierungsaufgaben	81
6.1 Einführung, Beispiele	81
Aufgaben	86
6.2 Das Simplexverfahren	87
6.2.1 Geometrische Grundlagen des Simplexverfahrens	87
6.2.2 Die Phase II des Simplexverfahrens	93
6.2.3 Die Vermeidung von Zyklen beim Simplexverfahren	98
6.2.4 Die Phase I des Simplexverfahrens	101
Aufgaben	106
6.3 Dualität bei linearen Programmen	110
6.3.1 Schwacher und starker Dualitätssatz	110
6.3.2 Ökonomische Interpretation der Dualität	118
6.3.3 Das duale Simplexverfahren	119

Aufgaben	124
6.4 Das Karmarkar-Verfahren	128
6.4.1 Das Karmarkar-Verfahren und seine Motivation	129
6.4.2 Die Konvergenz des Karmarkar-Verfahrens	135
6.4.3 Zurückführung eines linearen Programms auf Karmarkar-Normalform	138
Aufgaben	141
7 Unrestringierte Optimierungsaufgaben	143
7.1 Grundlagen	144
7.1.1 Einführung	144
7.1.2 Notwendige Optimalitätsbedingungen erster Ordnung	145
7.1.3 Notwendige und hinreichende Optimalitätsbedingungen zweiter Ordnung	152
7.1.4 Glatte konvexe Funktionen	155
Aufgaben	158
7.2 Ein Modellalgorithmus	162
7.2.1 Schrittweitenstrategien bei glatter Zielfunktion	163
7.2.2 Konvergenz des Modellalgorithmus bei glatter Zielfunktion	169
7.2.3 Das gedämpfte Gauß-Newton-Verfahren bei diskreten, nichtlinearen Approximationsaufgaben	173
Aufgaben	184
7.3 Quasi-Newton-Verfahren	192
7.3.1 Das Newton-Verfahren	192
7.3.2 Die Broyden-Klasse und das BFGS-Verfahren	195
7.3.3 Die globale Konvergenz des BFGS-Verfahrens	203
7.3.4 Die superlineare Konvergenz des BFGS-Verfahrens	206
Aufgaben	211
7.4 Verfahren der konjugierten Gradienten	218
7.4.1 Quadratische Zielfunktionen	219
7.4.2 Das Fletcher-Reeves-Verfahren	224
7.4.3 Ein gedächtnisloses BFGS-Verfahren	226
Aufgaben	231
7.5 Trust-Region-Verfahren	236
7.5.1 Einführung	236
7.5.2 Glatte, unrestringierte Optimierungsaufgaben	238
7.5.3 Nichtlineare Ausgleichsprobleme	248
7.5.4 Diskrete, nichtlineare Approximationsaufgaben	249
Aufgaben	258
Literaturverzeichnis	265
Index	273

Kapitel 5

Eigenwertaufgaben

In diesem Kapitel werden wir uns mit Eigenwertaufgaben, genauer mit Matrizeneigenwertaufgaben, und ihrer numerischen Behandlung beschäftigen. Es handelt sich hier darum, einen bestimmten Eigenwert, etwa einen betragsgrößten, gewisse oder alle Eigenwerte einer Matrix $A \in \mathbb{C}^{n \times n}$ und eventuell die zugehörigen Eigenvektoren zu bestimmen. Allerdings werden wir uns bei den Verfahren i. allg. auf den Fall *reeller* Matrizen $A \in \mathbb{R}^{n \times n}$ zurückziehen, auch wenn die notwendigen Modifikationen für komplexe Matrizen meistens nur gering sind. Auch bei Eigenwertaufgaben kann man kaum hoffen, ein „optimales Superverfahren“ angeben zu können, in das man lediglich als Input die gegebene Matrix hineinsteckt, und welches einem als Output die gewünschten Eigenwerte und die zugehörigen Eigenvektoren ausgibt. Neben der Aufgabenstellung (sind nur gewisse oder alle Eigenwerte zu berechnen?, sollen auch zugehörige Eigenvektoren ausgegeben werden?) wird die Struktur der gegebenen Matrix (Symmetrie, Dünnesetztheit) bei der Auswahl eines Verfahrens eine Rolle spielen.

Eigenwertaufgaben treten in den Anwendungen in vielen verschiedenen Bereichen auf. Einen kleinen Eindruck hiervon erhält man z. B. durch A. R. GOURLAY, G. A. WATSON (1973, S. 1–10). In einem ersten Abschnitt werden wir zunächst einige theoretische Aussagen zusammenstellen und danach auf die numerische Behandlung von Eigenwertaufgaben eingehen. Hierbei liegt der Schwerpunkt auf Abschnitt 5.2, in dem das sogenannte *QR*-Verfahren geschildert wird. In Abschnitt 5.3 werden wir uns mit Verfahren zur Bestimmung der Eigenwerte (und Eigenvektoren) symmetrischer Matrizen beschäftigen. Hier soll auch eine Lücke geschlossen werden, die in Abschnitt 1.6 offen blieb. Und zwar soll etwas zur numerischen Berechnung der Singulärwertzerlegung, insbesondere also der singulären Werte, einer gegebenen Matrix $A \in \mathbb{R}^{m \times n}$ ausgesagt werden.

Auch in diesem Kapitel kann nur ein Einblick gegeben werden, der durch die wesentlich ausführlicheren Darstellungen bei J. H. WILKINSON (1965), G. W. STEWART (1973), A. R. GOURLAY, G. A. WATSON (1973), B. N. PARLETT (1980), G. H. GOLUB, C. F. VAN LOAN (1989) beeinflußt wurde. Besonders hingewiesen sei auf J. H. WILKINSON, C. REINSCH (1971), wo man in Einzelbeiträgen verschiedener Autoren ausgefeilte Algol-Programme zu den später zu beschreibenden Algorithmen findet. Diese Programme waren Grundlage für EISPACK, einer Sammlung von in

Fortran geschriebenen Routinen zur numerischen Behandlung von Eigenwertaufgaben (siehe B. T. SMITH ET AL. (1974)).

5.1 Einige theoretische Grundlagen

5.1.1 Der Satz von Gerschgorin

Am Anfang wollen wir einen Satz beweisen, den man sofort „glaubt“, der selten exakt formuliert und noch seltener bewiesen, dafür aber oft angewandt wird:

- Die Eigenwerte einer Matrix hängen stetig von den Koeffizienten ab.

Oder auch:

- Die Wurzeln eines Polynoms hängen stetig von den Koeffizienten ab.

Der Beweis, den wir bringen wollen (siehe J. M. ORTEGA (1972, S. 42 ff.)), benutzt den Satz von Rouché (für einen Beweis kann man auf fast jedes Lehrbuch der Funktionentheorie verweisen, z. B. auf R. REMMERT (1984, S. 278)), den wir für unsere Zwecke wie folgt formulieren:

Satz von Rouché: Sei $D \subset \mathbb{C}$ eine offene Kreisscheibe mit Rand γ . Die komplexwertigen Funktionen f und g seien holomorph in einer Umgebung der abgeschlossenen Kreisscheibe $\text{cl } D = D \cup \gamma$ und es sei $|f(z) - g(z)| < |f(z)|$ für alle $z \in \gamma$. Dann haben f und g genau gleich viele Nullstellen in D .

Die Eigenwerte einer Matrix $A \in \mathbb{C}^{n \times n}$ sind genau die Wurzeln bzw. Nullstellen des zugehörigen charakteristischen Polynoms $p(\lambda) := \det(A - \lambda I)$. Die Koeffizienten von p hängen stetig von den Koeffizienten von A ab. Um die stetige Abhängigkeit der Eigenwerte einer Matrix von ihren Koeffizienten zu beweisen, genügt es daher zu zeigen, daß die Wurzeln eines Polynoms stetig von den Koeffizienten abhängen. Diese Aussage wird im folgenden Satz exakt formuliert und anschließend bewiesen.

Satz 1.1 Sei

$$p(z) := z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0$$

ein Polynom mit den Wurzeln $\lambda_1, \dots, \lambda_n$. Dann gibt es zu jedem $\epsilon > 0$ ein positives $\delta = \delta(\epsilon)$ mit der Eigenschaft: Ist

$$q(z) := z^n + b_{n-1}z^{n-1} + \cdots + b_1z + b_0$$

ein Polynom mit $|b_j - a_j| \leq \delta$, $j = 0, \dots, n-1$, so können die Wurzeln μ_1, \dots, μ_n von q so numeriert werden, daß $|\mu_j - \lambda_j| \leq \epsilon$, $j = 1, \dots, n$.

Beweis: Seien $\hat{\lambda}_1, \dots, \hat{\lambda}_m$ die paarweise verschiedenen Wurzeln von p . Sei $\epsilon > 0$ vorgegeben, o. B. d. A. sei $\epsilon < \frac{1}{2} \min_{1 \leq i < j \leq m} |\hat{\lambda}_i - \hat{\lambda}_j|$. Anschließend definiere man für $i = 1, \dots, m$ die offenen Kreisscheiben $D_i := \{z \in \mathbb{C} : |z - \hat{\lambda}_i| < \epsilon\}$ mit Rand γ_i sowie

$$m_i := \min\{|p(z)| : z \in \gamma_i\}, \quad M_i := \max\left\{\sum_{j=0}^{n-1} |z|^j : z \in \gamma_i\right\}.$$

Nach Konstruktion besitzt p keine Nullstellen auf γ_i , so daß $m_i > 0$. Nun wähle man $\delta > 0$ so klein, daß $M_i \delta < m_i$ für $i = 1, \dots, m$. Ist dann

$$q(z) := z^n + b_{n-1}z^{n-1} + \dots + b_1z + b_0$$

ein Polynom mit $|b_j - a_j| \leq \delta$, $j = 0, \dots, n-1$, so ist

$$|q(z) - p(z)| \leq \sum_{j=0}^{n-1} |b_j - a_j| |z|^j \leq \delta \sum_{j=0}^{n-1} |z|^j \leq \delta M_i < m_i \leq |p(z)| \quad \text{für alle } z \in \gamma_i.$$

Aus dem Satz von Rouché folgt, daß p und q genau dieselbe Anzahl (nämlich die Vielfachheit von λ_i) von Wurzeln in D_i haben. Hieraus folgt die Behauptung. \square

Als unmittelbare Folgerung notieren wir:

Satz 1.2 Die Eigenwerte einer Matrix hängen stetig von den Koeffizienten ab. Genauer: Ist $A \in \mathbb{C}^{n \times n}$ eine Matrix mit Eigenwerten $\lambda_1, \dots, \lambda_n$ und $\|\cdot\|$ eine natürliche Matrixnorm, so gibt es zu jedem $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ mit: Ist $B \in \mathbb{C}^{n \times n}$ eine Matrix mit $\|B - A\| \leq \delta$, so können die Eigenwerte μ_1, \dots, μ_n von B so numeriert werden, daß $|\mu_j - \lambda_j| \leq \epsilon$, $j = 1, \dots, n$.

Nun kommen wir zum *Satz von Gerschgorin*, durch dessen Anwendung eine Lokalisierung aller Eigenwerte einer gegebenen Matrix möglich ist.

Satz 1.3 (Gerschgorin) Sei $A = (a_{ij}) \in \mathbb{C}^{n \times n}$. Für $i = 1, \dots, n$ definiere man die sogenannten Gerschgorin-Kreise

$$G_i := \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\} \quad \text{mit} \quad r_i := \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|.$$

Dann gilt:

1. Ist λ ein Eigenwert von A , so ist $\lambda \in \bigcup_{i=1}^n G_i$. Alle Eigenwerte von A sind also in der Vereinigung aller Gerschgorin-Kreise enthalten.
2. Hat die Vereinigung \hat{G} von $m (< n)$ Kreisen G_i einen leeren Durchschnitt mit den restlichen $n - m$ Kreisen, so enthält \hat{G} genau m Eigenwerte von A (jeden entsprechend seiner algebraischen Vielfachheit gezählt, d. h. seiner Vielfachheit als Nullstelle des charakteristischen Polynoms).

Beweis: Wir beweisen den ersten Teil des Satzes so, daß wir im Anschluß einige Folgerungen ziehen können.

Sei λ ein Eigenwert von A , o. B. d. A. ist $\lambda \neq a_{ii}$, $i = 1, \dots, n$, denn andernfalls wäre λ sogar Mittelpunkt eines Gerschgorin-Kreises. Mit $D := \text{diag}(a_{11}, \dots, a_{nn})$ ist daher $(\lambda I - D)$ nichtsingulär, ferner ist 1 ein Eigenwert von $(\lambda I - D)^{-1}(A - D)$ (denn aus $Ax = \lambda x$ folgt $(\lambda I - D)^{-1}(A - D)x = x$), so daß für eine beliebige natürliche Matrixnorm $\|\cdot\|$ gilt:

$$(*) \quad 1 \leq \rho[(\lambda I - D)^{-1}(A - D)] \leq \|(\lambda I - D)^{-1}(A - D)\|.$$

Hierbei bezeichnet $\rho(\cdot)$ natürlich den Spektralradius, ferner wurde die triviale Abschätzung $\rho(B) \leq \|B\|$ (siehe Satz 2.9 in Abschnitt 1.2) benutzt. Wählt man in (*) als natürliche Matrixnorm $\|\cdot\|$ die Matrixnorm $\|\cdot\|_\infty$ der maximalen Zeilenbetragssumme, berücksichtigt man ferner

$$[(\lambda I - D)^{-1}(A - D)]_{ij} = \frac{a_{ij}}{\lambda - a_{ii}} (1 - \delta_{ij}), \quad 1 \leq i, j \leq n,$$

so erhält man

$$1 \leq \|(\lambda I - D)^{-1}(A - D)\|_\infty = \max_{i=1,\dots,n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|\lambda - a_{ii}|} = \max_{i=1,\dots,n} \frac{r_i}{|\lambda - a_{ii}|}.$$

Daher existiert ein $i \in \{1, \dots, n\}$ mit $|\lambda - a_{ii}| \leq r_i$ bzw. $\lambda \in G_i$. (In einer anschließenden Bemerkung werden wir uns überlegen, daß man durch Wahl einer anderen Norm ähnliche Abschätzungen erhalten kann.)

Für den Beweis des zweiten Teiles nehmen wir o. B. d. A. an, daß $\hat{G} = \bigcup_{i=1}^m G_i$. Mit $\tilde{G} := \bigcup_{i=m+1}^n G_i$ ist dann $\hat{G} \cap \tilde{G} = \emptyset$. Wieder sei $D := \text{diag}(a_{11}, \dots, a_{nn})$. Die Idee des Beweises besteht nun darin, für $t \in [0, 1]$ die Matrix $A(t) := D + t(A - D)$ zu definieren und durch einen Stetigkeitsschluß aus der Gültigkeit der Behauptung für $A(0) = D$ auf die für $A(1) = A$ zu schließen. Hierzu beachten wir zunächst, daß die zu $A(t)$ gehörenden Gerschgorin-Kreise für $t \in [0, 1]$ durch

$$G_i(t) = \{z \in \mathbb{C} : |z - a_{ii}| \leq tr_i\} \subset G_i, \quad i = 1, \dots, n,$$

gegeben sind. Wegen des ersten Teiles des Satzes liegen daher für $t \in [0, 1]$ alle Eigenwerte von $A(t)$ in $\hat{G} \cup \tilde{G} = \bigcup_{i=1}^n G_i$. Wir definieren

$$I := \{t \in [0, 1] : \text{Genau } m \text{ Eigenwerte von } A(t) \text{ liegen in } \hat{G}\}$$

und zeigen $1 \in I$. Hierzu sei

$$t_0 := \sup\{t : t \in I\}, \quad \epsilon := \frac{1}{2} \min \{|\hat{z} - \tilde{z}| : \hat{z} \in \hat{G}, \tilde{z} \in \tilde{G}\}.$$

Zu der gewünschten Aussage „hangeln“ wir uns in drei Schritten. Im ersten Schritt beachten wir, daß $0 \in I$ wegen $A(0) = D$. Im zweiten, entscheidenden Schritt zeigen wir $t_0 \in I$. Denn: Seien $\lambda_1(t_0), \dots, \lambda_n(t_0)$ die Eigenwerte von $A(t_0)$. Wegen $\|A(t) - A(t_0)\| = |t - t_0| \|A - D\|$ und Satz 1.2 existiert zu ϵ ein $\delta = \delta(\epsilon) > 0$ derart, daß es zu jedem $t \in [0, 1]$ mit $|t - t_0| \leq \delta$ eine Numerierung der Eigenwerte $\lambda_1(t), \dots, \lambda_n(t)$ von $A(t)$ mit $|\lambda_i(t) - \lambda_i(t_0)| \leq \epsilon$, $i = 1, \dots, n$, gibt. Nach Definition von t_0 existiert ein $t \in [t_0 - \delta, t_0] \cap I$. D. h. genau m Eigenwerte $\lambda_i(t)$ von $A(t)$ liegen in \hat{G} , die restlichen in \tilde{G} . Wegen $|\lambda_i(t) - \lambda_i(t_0)| \leq \epsilon$ sowie der Definition von ϵ als dem halben Abstand zwischen \hat{G} und \tilde{G} gilt das entsprechende für die Eigenwerte $\lambda_i(t_0)$ von $A(t_0)$. Daher ist $t_0 \in I$. Im dritten Schritt nehmen wir im Widerspruch zur Behauptung $t_0 < 1$ an. Mit $\delta > 0$ wie eben und $t \in (t_0, t_0 + \delta] \cap [0, 1]$ können die Eigenwerte $\lambda_1(t), \dots, \lambda_n(t)$ von $A(t)$ so numeriert werden, daß $|\lambda_i(t) - \lambda_i(t_0)| \leq \epsilon$ für $i = 1, \dots, n$. Wegen $t_0 \in I$ und der Definition von ϵ ist auch $t \in I$, ein Widerspruch zur Definition von t_0 . Insgesamt ist der Satz bewiesen. \square

Bemerkung: Im ersten Teil des Beweises zum Satz von Gershgorin haben wir eigentlich folgendes gezeigt: Sind $A, D \in \mathbb{C}^{n \times n}$ und ist λ ein Eigenwert von A , so ist λ entweder ein Eigenwert von D oder

$$1 \leq \rho[(\lambda I - D)^{-1}(A - D)] \leq \|(\lambda I - D)^{-1}(A - D)\|$$

und

$$1 \leq \rho[(A - D)(\lambda I - D)^{-1}] \leq \|(A - D)(\lambda I - D)^{-1}\|$$

für jede natürliche Matrixnorm $\|\cdot\|$. Zum Nachweis der zweiten Ungleichungskette beachte man, daß 1 auch ein Eigenwert von $(A - D)(\lambda I - D)^{-1}$ ist.

Wählt man daher z. B. $D := \text{diag}(a_{11}, \dots, a_{nn})$ und $\|\cdot\| := \|\cdot\|_1$ (maximale Spaltenbetragssummennorm), so erhält man: Ist $\lambda \neq a_{jj}$, $j = 1, \dots, n$, so ist

$$1 \leq \|(A - D)(\lambda I - D)^{-1}\|_1 = \max_{j=1, \dots, n} \frac{1}{|\lambda - a_{jj}|} \sum_{\substack{i=1 \\ i \neq j}}^n |a_{ij}|.$$

Hieraus schließt man, daß die Aussagen des Satzes von Gershgorin auch mit

$$G_j := \{z \in \mathbb{C} : |z - a_{jj}| \leq q_j\}, \quad q_j := \sum_{\substack{i=1 \\ i \neq j}}^n |a_{ij}|, \quad j = 1, \dots, n,$$

gelten. Ferner bemerken wir noch, daß die Frobenius-Norm $\|\cdot\|_F$ mit der euklidischen Vektornorm verträglich ist. Ist daher λ ein Eigenwert von A und kein Eigenwert von D , so ist

$$1 \leq \rho[(\lambda I - D)^{-1}(A - D)] \leq \|(\lambda I - D)^{-1}(A - D)\|_2 \leq \|(\lambda I - D)^{-1}(A - D)\|_F.$$

Ist daher z. B. $D := \text{diag}(a_{11}, \dots, a_{nn})$ und λ ein Eigenwert von A mit $\lambda \neq a_{ii}$ für $i = 1, \dots, n$, so ist

$$1 \leq \|(\lambda I - D)^{-1}(A - D)\|_F = \left(\sum_{i=1}^n \frac{1}{|\lambda - a_{ii}|^2} \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|^2 \right)^{1/2}.$$

Zu jedem Eigenwert λ von A existiert folglich ein $i \in \{1, \dots, n\}$ mit

$$|\lambda - a_{ii}| \leq \left(\sum_{\substack{i,j=1 \\ i \neq j}}^n |a_{ij}|^2 \right)^{1/2}.$$

Schließlich sei noch angemerkt, daß man die bisher erhaltenen Abschätzungen auch auf Matrizen anwenden kann, die dieselben Eigenwerte wie A besitzen, z. B. A^T oder eine zu A ähnliche Matrix, um hierdurch Aussagen für die Eigenwerte von A zu erhalten. Ist z. B. $P := \text{diag}(p_1, \dots, p_n)$ mit $p_i > 0$, $i = 1, \dots, n$, eine positive Diagonalmatrix, so hat $P^{-1}AP = (a_{ij}p_j/p_i)$ dieselben Eigenwerte wie A . Daher

folgt, daß der Satz von Gerschgorin auch mit

$$G_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \frac{1}{p_i} \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| p_j \right\} \quad (i = 1, \dots, n)$$

bzw.

$$G_j := \left\{ z \in \mathbb{C} : |z - a_{jj}| \leq p_j \sum_{\substack{i=1 \\ i \neq j}}^n \frac{|a_{ij}|}{p_i} \right\} \quad (j = 1, \dots, n)$$

gilt. Diese Bemerkung zeigt, daß es nicht nur *einen* Satz von Gerschgorin gibt, sondern etliche Varianten, die man aus der grundlegenden Beweisidee durch den Übergang zu verschiedenen Normen oder Matrizen mit denselben Eigenwerten wie die zugrunde liegende Matrix erhält. \square

Beispiel: Sind die Außendiagonalelemente einer Matrix sehr klein, so erwartet man, daß die Eigenwerte ungefähr mit den Diagonalelementen übereinstimmen. Durch den Satz von Gerschgorin kann diese qualitative Aussage präzisiert werden. Als Beispiel betrachten wir (siehe J. H. WILKINSON (1965, S. 74) und J. M. ORTEGA (1972, S. 49)) die Matrix

$$A = \begin{pmatrix} 0.9 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.2 \end{pmatrix} + 10^{-5} \begin{pmatrix} 0.1 & 0.4 & -0.2 \\ -0.1 & 0.5 & 0.1 \\ 0.2 & 0.1 & 0.3 \end{pmatrix}.$$

Die drei zugehörigen Gerschgorin-Kreise (im Sinne von Satz 1.3) sind disjunkt, daher besitzt die Matrix A Eigenwerte $\lambda_1, \lambda_2, \lambda_3$ mit

$$\begin{aligned} |\lambda_1 - (0.9 + 0.1 \cdot 10^{-5})| &\leq 0.6 \cdot 10^{-5}, \\ |\lambda_2 - (0.4 + 0.5 \cdot 10^{-5})| &\leq 0.2 \cdot 10^{-5}, \\ |\lambda_3 - (0.2 + 0.3 \cdot 10^{-5})| &\leq 0.3 \cdot 10^{-5}. \end{aligned}$$

Diese Abschätzungen können noch wesentlich verbessert werden. Denn setzt man $P := \text{diag}(10^5, 1, 1)$, so ist

$$P^{-1}AP = \begin{pmatrix} 0.9 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.2 \end{pmatrix} + \begin{pmatrix} 0.1 \cdot 10^{-5} & 0.4 \cdot 10^{-10} & -0.2 \cdot 10^{-10} \\ -0.1 & 0.5 \cdot 10^{-5} & 0.1 \cdot 10^{-5} \\ 0.2 & 0.1 \cdot 10^{-5} & 0.3 \cdot 10^{-5} \end{pmatrix}.$$

Der erste Gerschgorin-Kreis ist nach wie vor disjunkt zu den beiden anderen (die offenbar nicht mehr disjunkt sind), so daß man für den Eigenwert λ_1 die Abschätzung

$$|\lambda_1 - (0.9 + 0.1 \cdot 10^{-5})| \leq 0.6 \cdot 10^{-10}$$

erhält. Entsprechend kann man auch die Abschätzung für die beiden anderen Eigenwerte verbessern. \square

5.1.2 Der Satz von Bauer-Fike

Ziel der weiteren Untersuchungen in diesem Abschnitt wird es u. a. sein, Störungssätze für Eigenwerte zu beweisen, d. h. Aussagen darüber, wie sich Störungen in der Matrix

auf die Eigenwerte auswirken. In diesem Unterabschnitt wird der Satz von Bauer-Fike (siehe F. L. BAUER, C. T. FIKE (1960), aber auch z. B. J. M. ORTEGA (1972, S. 52)) bewiesen. Zunächst aber eine Definition, bei der wir die folgende Bezeichnung benutzen: Ist $x = (x_j) \in \mathbb{C}^n$, so sei $|x| \in \mathbb{R}^n$ definiert durch $|x| := (|x_j|)$. Ferner sei die Ungleichung $|x| \leq |y|$ zwischen den beiden Vektoren $|x|, |y| \in \mathbb{R}^n$ komponentenweise zu verstehen:

$$|x| \leq |y| \stackrel{\text{def}}{\iff} |x_j| \leq |y_j| \quad (j = 1, \dots, n).$$

Definition 1.4 Eine Norm $\|\cdot\|$ auf \mathbb{C}^n heißt *monoton*, wenn $\|x\| \leq \|y\|$ für alle $x, y \in \mathbb{C}^n$ mit $|x| \leq |y|$. Sie heißt *absolut*, wenn $\||x|\| = \|x\|$ für alle $x \in \mathbb{C}^n$.

Einen Beweis des nächsten Lemmas findet man auch bei F. L. BAUER, J. STOER, C. WITZGALL (1961), A. S. HOUSEHOLDER (1964, S. 47) und P. LANCASTER, M. TISMONETSKY (1985, S. 368).

Lemma 1.5 Sei $\|\cdot\|$ eine Norm auf \mathbb{C}^n bzw. die zugeordnete Matrixnorm. Dann sind die folgenden Aussagen äquivalent:

(a) $\|\cdot\|$ ist monoton.

(b) Für jede Diagonalmatrix $D := \text{diag}(d_1, \dots, d_n) \in \mathbb{C}^{n \times n}$ ist

$$\|D\| = \max_{j=1, \dots, n} |d_j|.$$

(c) $\|\cdot\|$ ist absolut.

Beweis: “(a) \Rightarrow (b)” Sei $\|\cdot\|$ monoton und $D := \text{diag}(d_1, \dots, d_n)$. Für ein beliebiges $x \in \mathbb{C}^n$ ist $|Dx| \leq \max_{j=1, \dots, n} |d_j| |x|$, wegen der Monotonie von $\|\cdot\|$ ergibt sich $\|Dx\| \leq \max_{j=1, \dots, n} |d_j| \|x\|$ und damit nach Definition der zugeordneten Matrixnorm

$$\|D\| = \sup_{x \neq 0} \frac{\|Dx\|}{\|x\|} \leq \max_{j=1, \dots, n} |d_j|.$$

Andererseits ist $\rho(D) = \max_{j=1, \dots, n} |d_j| \leq \|D\|$, insgesamt ist $\|D\| = \max_{j=1, \dots, n} |d_j|$ bewiesen.

“(b) \Rightarrow (c)” Sei $x \in \mathbb{C}^n$ beliebig gegeben. Man definiere $D_x := \text{diag}(d_1, \dots, d_n)$ durch

$$d_j := \begin{cases} 1 & \text{für } x_j = 0, \\ \frac{x_j}{|x_j|} & \text{für } x_j \neq 0, \end{cases} \quad (j = 1, \dots, n).$$

Dann ist $(D_x|x|)_j = d_j|x_j| = x_j$, $j = 1, \dots, n$, bzw. $D_x|x| = x$ und daher

$$\begin{aligned} \|x\| &= \|D_x|x|\| \leq \|D_x\| \|x\| = \max_{j=1, \dots, n} |d_j| \|x\| = \|x\|, \\ \|x\| &= \|D_x^{-1}x\| \leq \|D_x^{-1}\| \|x\| = \max_{j=1, \dots, n} |d_j^{-1}| \|x\| = \|x\|, \end{aligned}$$

folglich ist $\|x\| = \|x\|$ für jedes $x \in \mathbb{C}^n$, die Norm $\|\cdot\|$ ist also absolut.

“(c)⇒(a)” Seien $x, y \in \mathbb{C}^n$ mit $|x| \leq |y|$ gegeben, die Norm $\|\cdot\|$ sei absolut. Ist $D = \text{diag}(d_1, \dots, d_n)$ eine Diagonalmatrix mit $d_j \in \{+1, -1\}$ für $j = 1, \dots, n$, so ist $|D|y| = |y|$ und daher $\|D|y|\| = \||y|\| = \|y\|$. Anschaulich gesprochen bedeutet dies, daß die 2^n Eckpunkte des Quaders $Q := \{u \in \mathbb{R}^n : -|y| \leq u \leq |y|\}$ zum Rand der Kugel $K := \{u \in \mathbb{R}^n : \|u\| \leq \|y\|\}$ gehören. Da $|x| \in Q$ sich als Konvexitätskombination der Ecken von Q darstellen läßt, ist $|x| \in K$ und daher $\|x\| = \||x|\| \leq \|y\|$. Damit ist das Lemma bewiesen. \square

Beispiel: Offenbar sind für $p \geq 1$ die durch $\|x\|_p := (\sum_{j=1}^n |x_j|^p)^{1/p}$ definierten p -Normen absolut. Insbesondere erfüllen die p -Normen die Bedingung (b) im vorigen Lemma. Andererseits ist nicht jede Norm absolut, da z. B. transformierte Normen der euklidischen Norm i. allg. nicht absolut sind. \square

Satz 1.6 (Bauer-Fike) Sei $A \in \mathbb{C}^{n \times n}$ eine diagonalisierbare Matrix, d. h. es existiere eine nichtsinguläre Matrix $P \in \mathbb{C}^{n \times n}$ mit $P^{-1}AP = \text{diag}(\lambda_1, \dots, \lambda_n) =: D$. Ferner sei $A + E \in \mathbb{C}^{n \times n}$ eine Störung von A und λ ein Eigenwert von $A + E$. Dann ist

$$\min_{j=1, \dots, n} |\lambda - \lambda_j| \leq \|P^{-1}EP\| \leq \text{cond}(P) \|E\|.$$

Hierbei sei $\|\cdot\|$ die einer absoluten Vektornorm zugeordnete Matrixnorm, ferner bedeute $\text{cond}(P) := \|P\| \|P^{-1}\|$ die Kondition von P bezüglich dieser Matrixnorm.

Beweis: Ist λ auch ein Eigenwert von A , so ist die Behauptung trivialerweise richtig. Sei daher $\lambda \neq \lambda_j$, $j = 1, \dots, n$, und x ein zu λ gehörender Eigenvektor. Wegen $(A + E)x = \lambda x$ bzw.

$$Ex = (\lambda I - A)x = (\lambda I - PDP^{-1})x = P(\lambda I - D)P^{-1}x$$

folgt

$$P^{-1}x = (\lambda I - D)^{-1}(P^{-1}EP)P^{-1}x$$

und daher wegen des vorigen Lemmas mit

$$\|P^{-1}x\| \leq \|(\lambda I - D)^{-1}\| \|P^{-1}EP\| \|P^{-1}x\| = \max_{j=1, \dots, n} \frac{1}{|\lambda - \lambda_j|} \|P^{-1}EP\| \|P^{-1}x\|$$

die Behauptung. \square

Der Satz von Bauer-Fike sagt aus: Ist A durch eine Ähnlichkeitstransformation mit der nichtsingulären Matrix P diagonalisierbar und ist λ ein Eigenwert der gestörten Matrix $A + E$, so existiert ein Eigenwert λ_i von A mit $|\lambda - \lambda_i| \leq \text{cond}(P) \|E\|$. Die Kondition der Matrix P , in deren Spalten die (linear unabhängigen) Eigenvektoren von A stehen, bestimmt also die „Störanfälligkeit“ der Eigenwerte von A .

Das folgende Korollar zeigt, daß das Eigenwertproblem für hermitesche Matrizen gut konditioniert ist.

Korollar 1.7 Ist $A \in \mathbb{C}^{n \times n}$ hermitesch und $A + E$ eine Störung von A , sind $\lambda_1, \dots, \lambda_n$ die Eigenwerte von A und ist λ ein Eigenwert von $A + E$, so ist

$$\min_{j=1, \dots, n} |\lambda - \lambda_j| \leq \|E\|_2.$$

Beweis: Im Satz von Bauer-Fike wähle man als absolute Vektornorm die euklidische Norm $\|\cdot\|_2$ und benutze, daß A als hermitesche Matrix durch eine Ähnlichkeitstransformation mit einer unitären Matrix P auf Diagonalgestalt transformiert werden kann. Wegen $\text{cond}_2(P) = 1$ folgt die Behauptung des Korollars. \square

Bei Eigenwertaufgaben für hermitesche Matrizen kann man aus einer Näherung für einen Eigenwert und einen zugehörigen Eigenvektor eine Fehlerabschätzung erhalten. Das wird im folgenden Korollar präzisiert.

Korollar 1.8 Sei $A \in \mathbb{C}^{n \times n}$ hermitesch mit Eigenwerten $\lambda_1, \dots, \lambda_n$. Sei $\lambda \in \mathbb{R}$ eine Näherung für einen Eigenwert von A und $x \in \mathbb{C}^n \setminus \{0\}$ eine Näherung für einen zugehörigen Eigenvektor. Dann ist

$$\min_{j=1,\dots,n} |\lambda - \lambda_j| \leq \frac{\|\lambda x - Ax\|_2}{\|x\|_2}.$$

Beweis: Man definiere $E \in \mathbb{C}^{n \times n}$ durch $E := (\lambda x - Ax)x^H/\|x\|_2^2$. Dann ist

$$(A + E)x = Ax + (\lambda x - Ax) \frac{x^H x}{\|x\|_2^2} = \lambda x,$$

d. h. λ ist Eigenwert von $A + E$. Ferner ist

$$\|E\|_2 = \rho(E^H E)^{1/2} = \frac{\|\lambda x - Ax\|_2}{\|x\|_2^2} \rho(xx^H)^{1/2} = \frac{\|\lambda x - Ax\|_2}{\|x\|_2},$$

daher folgt aus dem gerade eben bewiesenen Korollar 1.7 die Behauptung. \square

5.1.3 Variationsprinzipien für Eigenwerte hermitescher Matrizen

In diesem Unterabschnitt werden zwei berühmte Variationsprinzipien für die Eigenwerte hermitescher Matrizen bewiesen.

Satz 1.9 (Rayleigh) Die Matrix $A \in \mathbb{K}^{n \times n}$ sei hermitesch (hierbei ist $\mathbb{K} = \mathbb{R}$ und damit A symmetrisch oder $\mathbb{K} = \mathbb{C}$). Die (reellen) Eigenwerte von A seien $\lambda_1 \geq \dots \geq \lambda_n$. Sei $\{u_1, \dots, u_n\} \subset \mathbb{K}^n$ ein zugehöriges Orthonormalsystem von Eigenvektoren, also

$$Au_i = \lambda_i u_i, \quad u_i^H u_j = \delta_{ij}, \quad 1 \leq i \leq j \leq n.$$

Für $j = 1, \dots, n$ definiere man den $(n+1-j)$ -dimensionalen linearen Teilraum

$$M_j := \{x \in \mathbb{K}^n : u_i^H x = 0, \quad i = 1, \dots, j-1\}.$$

Dann ist

$$\lambda_j = \max_{0 \neq x \in M_j} \frac{x^H A x}{x^H x}, \quad j = 1, \dots, n.$$

Beweis: Die unitäre Matrix $U := (u_1 \ \cdots \ u_n)$ transformiert A auf Diagonalschreibweise:

$$U^H A U = \text{diag}(\lambda_1, \dots, \lambda_n) =: \Lambda.$$

Sei $j \in \{1, \dots, n\}$ fest, $x \in M_j \setminus \{0\}$ beliebig und $y := U^H x$. Dann ist $y_i = u_i^H x = 0$ für $i = 1, \dots, j-1$. Wegen $\lambda_i \leq \lambda_j$ für $i = j, \dots, n$ ist

$$\frac{x^H A x}{x^H x} = \frac{y^H \Lambda y}{y^H y} = \sum_{i=j}^n \lambda_i |y_i|^2 \Big/ \sum_{i=j}^n |y_i|^2 \leq \lambda_j$$

und damit

$$\sup_{0 \neq x \in M_j} \frac{x^H A x}{x^H x} \leq \lambda_j.$$

Andererseits ist $u_j \in M_j$ und $u_j^H A u_j / u_j^H u_j = \lambda_j$, so daß insgesamt das Rayleighsche Maximumsprinzip bewiesen ist.

Bemerkung: Durch $R(x) := x^H A x / x^H x$ ist der sogenannte *Rayleigh-Quotient* definiert, wobei $A \in \mathbb{K}^{n \times n}$ als hermitesch vorausgesetzt wird. Definiert man $f: \mathbb{R} \rightarrow \mathbb{R}$ bei festem $x \in \mathbb{K}^n \setminus \{0\}$ durch

$$f(\lambda) := \frac{1}{2} \|Ax - \lambda x\|_2^2 = \frac{\lambda^2}{2} \|x\|_2^2 - \lambda x^H A x + \frac{1}{2} \|Ax\|_2^2,$$

so nimmt f in $\lambda = R(x)$ sein Minimum an. Ist daher x näherungsweise ein Eigenvektor von A , so kann man erwarten, daß $R(x)$ eine gute Näherung für einen zugehörigen Eigenwert ist. \square

Satz 1.10 (Courant) Sei $A \in \mathbb{K}^{n \times n}$ hermitesch mit Eigenwerten $\lambda_1 \geq \dots \geq \lambda_n$. Für $j = 1, \dots, n$ sei

$$\mathcal{N}_j := \{N_j \subset \mathbb{K}^n : N_j \text{ ist linearer Teilraum mit } \dim N_j = n+1-j\}.$$

Dann ist

$$\lambda_j = \min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^H A x}{x^H x}, \quad j = 1, \dots, n.$$

Beweis: Sei $\{u_1, \dots, u_n\}$ ein Orthonormalsystem von Eigenvektoren zu den Eigenwerten $\lambda_1, \dots, \lambda_n$ von A . Man halte $j \in \{1, \dots, n\}$ fest, definiere

$$L_j := \text{span} \{u_1, \dots, u_j\}$$

und wähle $N_j \in \mathcal{N}_j$ beliebig. Wegen

$$\dim(L_j \cap N_j) = \dim L_j + \dim N_j - \dim(L_j + N_j) = n+1 - \dim(L_j + N_j) \geq 1$$

existiert ein $x \in L_j \cap N_j$ mit $x \neq 0$. Als Element von L_j läßt sich x eindeutig in der Form $x = \sum_{i=1}^j \alpha_i u_i$ darstellen. Dann ist

$$\frac{x^H A x}{x^H x} = \sum_{i=1}^j \lambda_i |\alpha_i|^2 \Big/ \sum_{i=1}^j |\alpha_i|^2 \geq \lambda_j$$

und daher

$$\min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^H A x}{x^H x} \geq \lambda_j.$$

Wählt man andererseits

$$N_j = M_j := \{x \in \mathbb{K}^n : u_i^H x = 0, \quad i = 1, \dots, j-1\},$$

so ist nach dem Rayleighschen Maximumprinzip

$$\max_{0 \neq x \in M_j} \frac{x^H A x}{x^H x} = \lambda_j.$$

Insgesamt ist das Courantsche Minimum-Maximum Prinzip bewiesen. \square

Eine wichtige Folgerung aus dem Courantschen Minimum-Maximum Prinzip ist

Satz 1.11 Seien $A, B \in \mathbb{K}^{n \times n}$ hermitesch. Dann genügen die Eigenwerte $\lambda_j(A)$, $\lambda_j(B)$, $j = 1, \dots, n$, von A bzw. B in der Anordnung

$$\lambda_1(A) \geq \dots \geq \lambda_n(A), \quad \lambda_1(B) \geq \dots \geq \lambda_n(B)$$

für jede natürliche Matrixnorm $\|\cdot\|$ der Abschätzung

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|, \quad j = 1, \dots, n.$$

Beweis: Mit A und B ist auch $A - B$ hermitesch. Für ein beliebiges $x \in \mathbb{K}^n \setminus \{0\}$ ist daher

$$\frac{x^H (A - B)x}{x^H x} \leq \|A - B\|_2 = \rho(A - B) \leq \|A - B\|$$

und folglich

$$\frac{x^H A x}{x^H x} \leq \frac{x^H B x}{x^H x} + \|A - B\|.$$

Nun sei $j \in \{1, \dots, n\}$ fest und wie im Courantschen Minimum-Maximum Prinzip

$$\mathcal{N}_j := \{N_j \subset \mathbb{K}^n : N_j \text{ ist linearer Teilraum mit } \dim N_j = n + 1 - j\}.$$

Mit beliebigem $N_j \in \mathcal{N}_j$ ist dann

$$\max_{0 \neq x \in N_j} \frac{x^H A x}{x^H x} \leq \max_{0 \neq x \in N_j} \frac{x^H B x}{x^H x} + \|A - B\|$$

und folglich

$$\min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^H A x}{x^H x} \leq \min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^H B x}{x^H x} + \|A - B\|$$

bzw. wegen des Courantschen Minimum-Maximum Prinzips

$$\lambda_j(A) \leq \lambda_j(B) + \|A - B\|, \quad j = 1, \dots, n.$$

Vertauscht man hier die Rollen von A und B , so erhält man auch

$$\lambda_j(B) \leq \lambda_j(A) + \|A - B\|, \quad j = 1, \dots, n,$$

insgesamt ist der Satz bewiesen. \square

Setzt man im vorigen Satz $B := \text{diag}(A)$ und $\|\cdot\| := \|\cdot\|_\infty$ bzw. $\|\cdot\| := \|\cdot\|_F$ (die Frobenius-Norm $\|\cdot\|_F$ ist zwar keine natürliche Matrixnorm, aber es ist $\|A\|_2 \leq \|A\|_F$), so erhält man:

Korollar 1.12 Sei $A = (a_{ij}) \in \mathbb{C}^{n \times n}$ hermitesch und $\{a'_{11}, \dots, a'_{nn}\}$ eine Permutation der (reellen) Diagonalelemente $\{a_{11}, \dots, a_{nn}\}$ mit $a'_{11} \geq \dots \geq a'_{nn}$. Für die Eigenwerte $\lambda_1 \geq \dots \geq \lambda_n$ von A gelten dann die Abschätzungen

$$|\lambda_j - a'_{jj}| \leq \max_{i=1, \dots, n} \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}|, \quad j = 1, \dots, n,$$

und

$$|\lambda_j - a'_{jj}| \leq \left(\sum_{\substack{i, k=1 \\ i \neq k}}^n |a_{ik}|^2 \right)^{1/2}, \quad j = 1, \dots, n.$$

5.1.4 Der Satz von Schur

Eine (komplexe) hermitesche Matrix lässt sich durch eine (komplexe) unitäre Ähnlichkeitstransformation auf (reelle) Diagonalgestalt transformieren. Entsprechend kann eine (reelle) symmetrische Matrix durch eine (reelle) orthogonale Ähnlichkeitstransformation auf (reelle) Diagonalgestalt transformiert werden. Diese Aussagen liefern die theoretische Grundlage für das Jacobi-Verfahren (siehe Unterabschnitt 5.3.1) zur Berechnung der Eigenwerte einer hermiteschen bzw. symmetrischen Matrix, bei dem eine Folge $\{U_k\}$ bzw. $\{Q_k\}$ unitärer bzw. orthogonaler Matrizen konstruiert wird mit der Eigenschaft, daß $\{U_k^H A U_k\}$ bzw. $\{Q_k^T A Q_k\}$ gegen eine Diagonalmatrix konvergiert, die gerade die gesuchten Eigenwerte von A enthält. Entsprechend ist eine Verallgemeinerung, nämlich der gleich zu formulierende Satz von Schur, Grundlage für das QR-Verfahren, das für die Praxis wichtigste Verfahren zur Berechnung der Eigenwerte einer beliebigen Matrix. Wir werden auch hier eine komplexe und eine reelle Version angeben.

Satz 1.13 (Komplexe Schur-Zerlegung) Sei $A \in \mathbb{C}^{n \times n}$. Dann existiert eine unitäre Matrix $U \in \mathbb{C}^{n \times n}$ derart, daß $U^H A U$ eine obere Dreiecksmatrix (mit den Eigenwerten von A als Diagonalelementen) ist.

Beweis: Die Aussage wird durch vollständige Induktion nach n bewiesen. Für $n = 1$ ist die Aussage trivial. Für den Induktionsschluß nehmen wir an, die Aussage sei für $(n-1) \times (n-1)$ -Matrizen richtig. λ_1 sei ein Eigenwert von A mit zugehörigem, durch $\|v_1\|_2 = 1$ normierten Eigenvektor v_1 . Man ergänze v_1 durch $\{v_2, \dots, v_n\}$ zu einer Orthonormalbasis des \mathbb{C}^n und definiere die unitäre Matrix $V := (\begin{array}{cccc} v_1 & v_2 & \cdots & v_n \end{array})$. Dann ist

$$V^H A V = \left(\begin{array}{c|c} \lambda_1 & y^T \\ \hline 0 & A_1 \end{array} \right) \quad \text{mit} \quad A_1 \in \mathbb{C}^{(n-1) \times (n-1)}, \quad y \in \mathbb{C}^{n-1}.$$

Nach Induktionsvoraussetzung existiert eine unitäre Matrix $W \in \mathbb{C}^{(n-1) \times (n-1)}$ derart, daß $W^H A_1 W$ eine obere Dreiecksmatrix ist. Man definiere die unitäre Matrix

$$U := V \begin{pmatrix} 1 & 0^T \\ 0 & W \end{pmatrix} \in \mathbb{C}^{n \times n}.$$

Dann ist

$$U^H A U = \begin{pmatrix} 1 & 0^T \\ 0 & W^H \end{pmatrix} \begin{pmatrix} \lambda_1 & y^T \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0^T \\ 0 & W \end{pmatrix} = \begin{pmatrix} \lambda_1 & y^T W \\ 0 & W^H A_1 W \end{pmatrix}$$

eine obere Dreiecksmatrix. Damit ist der Satz bewiesen. \square

Für reelle Matrizen, die auch komplexe Eigenwerte (dann notwendig in konjugiert komplexen Paaren auftretend) besitzen, kann kein direktes reelles Analogon zum komplexen Schurschen Zerlegungssatz existieren. Da wir uns aber später im wesentlichen auf die Berechnung der Eigenwerte reeller Matrizen zurückziehen werden, ist es wichtig, auch diesen Fall zu klären.

Satz 1.14 (Reelle Schur-Zerlegung) Sei $A \in \mathbb{R}^{n \times n}$. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n \times n}$ mit

$$Q^T A Q = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{mm} \end{pmatrix},$$

wobei die Diagonalblöcke R_{ii} entweder 1×1 -Matrizen oder 2×2 -Matrizen mit einem Paar konjugiert komplexer (nicht reeller) Eigenwerte sind. Insbesondere ist eine reelle Matrix, die nur reelle Eigenwerte besitzt, orthogonal ähnlich zu einer oberen Dreiecksmatrix.

Beweis: Sei k die Anzahl von Paaren konjugiert komplexer Eigenwerte von A . Der Satz wird durch vollständige Induktion nach k bewiesen. Für $k = 0$ besitzt A nur reelle Eigenwerte. Dann ist die Aussage des Satzes offenbar richtig, da man in diesem Falle so wie im komplexen Fall schließen kann. Für $k \geq 1$ sei $\lambda = \alpha + i\beta$ mit $\beta \neq 0$ ein komplexer Eigenwert und $y + iz$ ein zugehöriger Eigenvektor ($y, z \in \mathbb{R}^n$ und $z \neq 0$). Dann ist $A(y + iz) = (\alpha + i\beta)(y + iz)$ und daher

$$\begin{aligned} Ay &= \alpha y - \beta z, \\ Az &= \beta y + \alpha z \end{aligned} \quad \text{bzw.} \quad A \begin{pmatrix} y & z \end{pmatrix} = \begin{pmatrix} y & z \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}.$$

Die (reellen) Vektoren y und z sind linear unabhängig (andernfalls wäre z ein reeller Eigenvektor zum komplexen Eigenwert λ), daher hat die $n \times 2$ -Matrix $\begin{pmatrix} y & z \end{pmatrix}$ den Rang 2. Folglich existiert eine orthogonale Matrix $V \in \mathbb{R}^{n \times n}$ und eine obere 2×2 -Dreiecksmatrix $R_1 \in \mathbb{R}^{2 \times 2}$ mit

$$\begin{pmatrix} y & z \end{pmatrix} = V \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \quad \left\{ \begin{array}{l} 2 \\ n-2 \end{array} \right\}$$

(die ersten beiden Spalten von V erhält man durch Orthonormieren von y und z nach E. Schmidt, die restlichen $n - 2$ Spalten durch Ergänzen der ersten beiden zu einer Orthonormalbasis des \mathbb{R}^n). Einsetzen ergibt

$$V^T A V \begin{pmatrix} R_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{R}_1 \\ 0 \end{pmatrix} \quad \text{mit } \tilde{R}_1 := R_1 \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \in \mathbb{R}^{2 \times 2}.$$

Partitioniert man $V^T A V$ in der Form

$$V^T A V = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right) \quad \left. \begin{array}{l} 2 \\ n-2 \end{array} \right\} \quad \text{mit } B_{11} \in \mathbb{R}^{2 \times 2}, B_{22} \in \mathbb{R}^{(n-2) \times (n-2)},$$

so ist

$$\begin{pmatrix} B_{11} R_1 \\ B_{21} R_1 \end{pmatrix} = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right) \begin{pmatrix} R_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{R}_1 \\ 0 \end{pmatrix}.$$

Da R_1 nichtsingulär ist, folgt

$$R_1^{-1} B_{11} R_1 = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}, \quad B_{21} = 0.$$

Die Eigenwerte von A bzw. der orthogonal ähnlichen Matrix

$$V^T A V = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array} \right)$$

sind daher die Eigenwerte von B_{11} , also das konjugiert komplexe Paar $\alpha \pm i\beta$, und die Eigenwerte von B_{22} . Auf B_{22} kann die Induktionsannahme angewandt werden. Hiernach existiert eine orthogonale Matrix $W \in \mathbb{R}^{(n-2) \times (n-2)}$ derart, daß $W^T B_{22} W$ die behauptete Gestalt hat. Mit $Q := V \operatorname{diag}(I_2, W)$ folgt die Behauptung. \square

Der reelle Schursche Zerlegungssatz läßt uns hoffen, daß man unter geeigneten Voraussetzungen an eine reelle Matrix $A \in \mathbb{R}^{n \times n}$ eine Folge $\{Q_k\}$ orthogonaler Matrizen bestimmen kann mit der Eigenschaft, daß $\{Q_k^T A Q_k\}$ gegen eine obere Block-Dreiecksmatrix konvergiert, deren Diagonalblöcke entweder 1×1 -Matrizen (die reellen Eigenwerte von A) oder 2×2 -Matrizen mit einem Paar konjugiert komplexer Eigenwerte (komplexe Eigenwerte von A) sind.

Aufgaben

- Mit $A = (a_{ij}) \in \mathbb{C}^{n \times n}$, $\alpha > 0$ und $i \in \{1, \dots, n\}$ sei der Kreis

$$R_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \alpha \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \right\}$$

disjunkt zu den Kreisen

$$R_k := \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq \alpha^{-1} |a_{ki}| + \sum_{\substack{j=1 \\ j \neq k,i}}^n |a_{kj}| \right\}, \quad k \in \{1, \dots, n\} \setminus \{i\}.$$

Dann enthält R_i genau einen Eigenwert von A .

Hinweis: Man definiere die Diagonalmatrix $P := \text{diag}(p_1, \dots, p_n)$ mit

$$p_j := \begin{cases} \alpha^{-1} & \text{für } j = i, \\ 1 & \text{für } j \neq i, \end{cases} \quad j = 1, \dots, n,$$

und wende den Satz von Gerschgorin auf $P^{-1}AP$ an.

2. Sei $A = (a_{ij}) \in \mathbb{C}^{n \times n}$. Mit

$$G_i := \{z \in \mathbb{C} : |\lambda - a_{ii}| \leq r_i\}, \quad r_i := \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad i = 1, \dots, n,$$

seien die zugehörigen Gerschgorin-Kreise bezeichnet. Für ein $i \in \{1, \dots, n\}$ sei G_i disjunkt zu allen anderen Kreisen, also $G_i \cap \bigcup_{j \neq i} G_j = \emptyset$. Dann besitzt A genau einen Eigenwert λ in G_i und hierzu existiert ein Eigenvektor x mit $x_i = 1, |x_j| < 1$ für $j \neq i$.

3. Man gebe eine möglichst gute Lokalisierung der Eigenwerte von

$$A := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} + 10^{-5} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

an.

4. Mit Hilfe des Satzes von Gerschgorin zeige man:

- (a) Ist $A \in \mathbb{C}^{n \times n}$ hermitesch und $\sum_{j=1, j \neq i}^n |a_{ij}| < a_{ii}$ für $i = 1, \dots, n$ (also A diagonal dominant mit positiven Diagonalelementen), so ist A positiv definit.
- (b) Die Matrix

$$A := \begin{pmatrix} n & 1 & 1 & \cdots & 1 \\ 1 & 2 & 0 & \cdots & 0 \\ 1 & 0 & 3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \cdots & 0 & n \end{pmatrix} \in \mathbb{R}^{n \times n}$$

ist positiv definit.

5. Eine kleine Störung der Koeffizienten einer *nichtsymmetrischen* Matrix kann eine große Störung der Eigenwerte bewirken. Hierzu betrachte man das folgende Beispiel (siehe K. E. ATKINSON (1978, S. 507)). Sei

$$A := \begin{pmatrix} 101 & -90 \\ 110 & -98 \end{pmatrix}, \quad E(\epsilon) := \epsilon \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix}.$$

Man vergleiche die vom Satz von Bauer-Fike vorhergesagten Störungen der Eigenwerte von A mit den exakten Werten.

6. Sei $M \in \mathbb{R}^{n \times n}$ symmetrisch mit Eigenwerten $\mu_n \leq \dots \leq \mu_1$. Sei $X \in \mathbb{R}^{n \times (n-1)}$ mit $X^T X = I$ gegeben. $X^T M X \in \mathbb{R}^{(n-1) \times (n-1)}$ ist symmetrisch und besitze die Eigenwerte $\nu_{n-1} \leq \dots \leq \nu_1$. Dann ist

$$\mu_{j+1} \leq \nu_j \leq \mu_j, \quad j = 1, \dots, n-1.$$

Hinweis: Sei $j \in \{1, \dots, n-1\}$ fest. Wie im Courantschen Minimum-Maximum Prinzip sei

$$\mathcal{N}_j^{(n)} := \{N_j \subset \mathbb{R}^n : N_j \text{ ist linearer Teilraum mit } \dim N_j = n+1-j\}.$$

Eine Anwendung des Courantschen Minimum-Maximum Prinzips auf $X^T M X$ liefert die Existenz von $L_j \in \mathcal{N}_j^{(n-1)}$ mit

$$\nu_j = \max_{0 \neq y \in L_j} \frac{y^T X^T M X y}{y^T y} = \max_{0 \neq y \in L_j} \frac{(X y)^T M (X y)}{y^T y}.$$

Nun definiere man $N_{j+1} := X(L_j) \in \mathcal{N}_{j+1}^{(n)}$ als Bild von L_j unter der durch X gegebenen linearen Abbildung. Mit dem Courantschen Minimum-Maximum Prinzip folgt

$$\nu_j = \max_{0 \neq x \in N_{j+1}} \frac{x^T M x}{x^T x} \geq \mu_{j+1}, \quad j = 1, \dots, n-1.$$

Nun wende man die gerade eben bewiesene Aussage auf $-M$ statt M an. Dann folgt $\nu_j \leq \mu_j$ für $j = 1, \dots, n-1$.

7. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch, $a \in \mathbb{R}^n$ und $M := A + \alpha a a^T$ mit $\alpha \neq 0$.

$$\begin{aligned} \lambda_n &\leq \lambda_{n-1} \leq \dots \leq \lambda_1 && \text{seien die Eigenwerte von } A, \\ \mu_n &\leq \mu_{n-1} \leq \dots \leq \mu_1 && \text{seien die Eigenwerte von } M. \end{aligned}$$

Dann gilt:

- (a) Ist $\alpha > 0$, so ist $\lambda_n \leq \mu_n \leq \lambda_{n-1} \leq \dots \leq \lambda_1 \leq \mu_1$.
- (b) Ist $\alpha < 0$, so ist $\mu_n \leq \lambda_n \leq \mu_{n-1} \leq \dots \leq \mu_1 \leq \lambda_1$.

Hinweis: Man betrachte den Fall $\alpha > 0$. Da $M - A = \alpha a a^T$ positiv semidefinit ist, folgt aus dem Courantschen Minimum-Maximum Prinzip, daß $\lambda_j \leq \mu_j$ für $j = 1, \dots, n$. O. B. d. A. sei $a \neq 0$. Durch $\{x_1, \dots, x_{n-1}\}$ sei eine Orthonormalbasis von $\{x \in \mathbb{R}^n : a^T x = 0\}$ gegeben. Hiermit definiere man $X := (x_1 \ \dots \ x_{n-1}) \in \mathbb{R}^{n \times (n-1)}$. Dann ist

$$X^T X = I, \quad X^T M X = X^T A X + \alpha \underbrace{(X^T a)(X^T a)^T}_{=0} = X^T A X.$$

Seien $\nu_{n-1} \leq \dots \leq \nu_1$ die Eigenwerte von $X^T M X = X^T A X$. Eine Anwendung der Aussage der vorigen Aufgabe liefert

$$\lambda_{j+1} \leq \nu_j \leq \lambda_j, \quad \mu_{j+1} \leq \nu_j \leq \mu_j, \quad j = 1, \dots, n-1.$$

Für $j = 2, \dots, n$ ist daher $\lambda_j \leq \mu_j \leq \nu_{j-1} \leq \lambda_{j-1} \leq \mu_{j-1}$, womit die Aussage für $\alpha > 0$ bewiesen ist. Für $\alpha < 0$ verläuft der Beweis analog.

8. Eine Matrix $A \in \mathbb{C}^{n \times n}$ heißt *normal*, wenn $A^H A = AA^H$. Mit Hilfe des Schurschen Zerlegungssatzes zeige man: Eine Matrix $A \in \mathbb{C}^{n \times n}$ ist genau dann normal, wenn sich A durch eine unitäre Ähnlichkeitstransformation auf Diagonalgestalt transformieren läßt.
9. Der *Brouwersche Fixpunktsatz* sagt aus:

- Ist $K \subset \mathbb{R}^n$ nichtleer, konvex, kompakt und $f: K \rightarrow \mathbb{R}^n$ stetig mit $f(K) \subset K$, so besitzt f einen Fixpunkt in K , es existiert also ein $x \in K$ mit $f(x) = x$.

Mit dem Brouwerschen Fixpunktsatz beweise man den *Satz von Perron-Frobenius* (bzw. genauer Teile davon):

- Sei $A \in \mathbb{R}^{n \times n}$ nichtnegativ und unzerlegbar. Dann ist der Spektralradius $\rho(A)$ von A ein positiver Eigenwert von A , zu dem es einen positiven Eigenvektor (dessen Komponenten also sämtlich positiv sind) gibt.

Hierbei heißt $A = (a_{ij})$ *nichtnegativ*, wenn $a_{ij} \geq 0$ für $1 \leq i, j \leq n$. Die Matrix A heißt *unzerlegbar* (oder auch *irreduzibel*, siehe Definition 4.2 in Abschnitt 2.4), wenn es keine nichtleeren Teilmengen N_1, N_2 von $\{1, \dots, n\}$ gibt mit

- $N_1 \cap N_2 = \emptyset$, $N_1 \cup N_2 = \{1, \dots, n\}$,
- $a_{ij} = 0$ für $(i, j) \in N_1 \times N_2$.

Hinweis: Definiere die konvexe, kompakte Menge $K \subset \mathbb{R}^n$ und $f: K \rightarrow \mathbb{R}^n$ durch

$$K := \{x \in \mathbb{R}^n : x \geq 0, e^T x = 1\}, \quad f(x) := \frac{Ax}{e^T Ax}$$

mit $e := (1, \dots, 1)^T$. Für $x \in K$ ist $e^T Ax > 0$ (Beweis?). Daher ist f auf K stetig und $f(K) \subset K$. Der Brouwersche Fixpunktsatz liefert die Existenz von $x \in K$ mit $Ax = (e^T Ax)x$. Wegen $x \neq 0$ ist x Eigenvektor zum Eigenwert $r := e^T Ax > 0$. Ferner ist sogar $x > 0$. Denn andernfalls definiere man die Mengen $N_1 := \{i : x_i = 0\}$, $N_2 := \{j : x_j > 0\}$. Dann sind N_1, N_2 nichtleer, $N_1 \cap N_2 = \emptyset$ und $N_1 \cup N_2 = \{1, \dots, n\}$. Für $(i, j) \in N_1 \times N_2$ ist $(Ax)_i = \sum_{k=1}^n a_{ik} x_k = 0$, wegen $x_j > 0$ also $a_{ij} = 0$, ein Widerspruch dazu, daß A irreduzibel. Zu zeigen bleibt $r = \rho(A)$. Der Eigenwert λ von A ist auch Eigenwert von A^T . Daher existiert $y \in \mathbb{C}^n$ mit $A^T y = \lambda y$ und $\|y\|_1 = e^T |y| = 1$. Dann ist

$$|\lambda| |y| = |\lambda y| = |A^T y| \leq A^T |y|.$$

Mit dem Eigenvektor $x > 0$ zu $r > 0$ ist daher

$$|\lambda| |x^T |y| \leq x^T A^T |y| = (Ax)^T |y| = r x^T |y|.$$

Wegen $x > 0$ und $y \neq 0$ ist $x^T |y| > 0$, daher $|\lambda| \leq r$ und folglich $r = \rho(A)$. Die behaupteten Aussagen sind damit bewiesen.

5.2 Das QR -Verfahren

In diesem Abschnitt beginnen wir mit der numerischen Behandlung von Eigenwertaufgaben und werden das wohl wichtigste Verfahren zur Berechnung der Eigenwerte und Eigenvektoren einer Matrix, das QR -Verfahren, angeben, motivieren und analysieren. Dieses Verfahren ist gleichzeitig und unabhängig voneinander von J. G. F. FRANCIS (1961/62) und V. N. KUBLANOVSKAYA (1961) entwickelt worden. Der Einfachheit halber werden wir (meistens) annehmen, daß die (quadratische) Matrix A , deren Eigenwerte und Eigenvektoren zu berechnen sind, *reell* ist. Im Prinzip sieht das QR -Verfahren folgendermaßen aus:

- In einem Reduktionsschritt führe man die gegebene Matrix $A \in \mathbb{R}^{n \times n}$ in eine ähnliche, „einfachere“ Matrix A_1 über.
- Für $k = 1, 2, \dots$

Wähle bzw. bestimme einen „Shift-Parameter“ $\sigma_k \in \mathbb{R}$, bilde eine QR -Zerlegung $A_k - \sigma_k I = Q_k R_k$ (mit einer orthogonalen Matrix Q_k und einer oberen Dreiecksmatrix R_k) und berechne $A_{k+1} := R_k Q_k + \sigma_k I$.

Man kann sich leicht überlegen, daß $A_{k+1} = Q_k^T A_k Q_k$, so daß durch das QR -Verfahren eine Folge $\{A_k\}$ von Matrizen erzeugt wird, die orthogonal ähnlich zu der Matrix A_1 sind (welche wiederum ähnlich zu der Ausgangsmatrix A ist). Von dieser Folge erhofft man sich, daß sie gegen eine obere Block-Dreiecksmatrix konvergiert (siehe reelle Schur-Zerlegung, Satz 1.14), aus der man die gesuchten Eigenwerte ablesen bzw. leicht berechnen kann. Es wird also massiv davon Gebrauch gemacht, daß sich bei einer Ähnlichkeitstransformation die Eigenwerte einer Matrix nicht verändern.

Der Reduktionsschritt am Anfang wird in 5.2.1 beschrieben. Er dient dazu, die Kosten einer QR -Zerlegung in jedem Schritt des Iterationsverfahrens entscheidend zu senken. Die Durchführung eines Schrittes des QR -Verfahrens mit einem (expliziten) Shift $\sigma \in \mathbb{R}$ ist Gegenstand von 5.2.2. Hier wird auch begründet, daß die im Reduktionsschritt gewonnene „einfachere“ Form A_1 bei der gesamten, durch das QR -Verfahren gewonnenen Folge $\{A_k\}$ erhalten bleibt. In 5.2.3 wird die Vektoriteration nach v. Mises beschrieben, welche in einem gewissen Sinne als ein „alter Verwandter“ des QR -Verfahrens angesehen werden kann. Die Einführung von Shift-Parametern dient zur Konvergenzbeschleunigung, ganz ähnlich wie bei der inversen Iteration nach Wielandt in 5.2.4. In 5.2.5 wird ein Konvergenzsatz für das einfache QR -Verfahren ($\sigma_k = 0$ für alle k) bewiesen, außerdem wird hier die Analogie zur Vektoriteration deutlich. Der Unterabschnitt 5.2.6 beschäftigt sich mit der Wahl geeigneter Shift-Parameter und geht insbesondere auf den nicht nur für die Berechnung (konjugiert) komplexer Eigenwerte wichtigen QR -Doppelschritt ein. Erst im nächsten Abschnitt, in 5.3.3, wird das QR -Verfahren für symmetrische Matrizen beschrieben.

5.2.1 Die Transformation einer Matrix auf Hessenberg-Form

Der erste Schritt bei der numerischen Behandlung von Eigenwertaufgaben besteht oft darin, die gegebene Matrix $A \in \mathbb{R}^{n \times n}$ durch eine Ähnlichkeitstransformation auf

„einfachere“ Form zu transformieren. Dies gilt insbesondere für das QR-Verfahren.

Definition 2.1 Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ heißt eine *obere Hessenberg-Matrix* (oder besitzt *obere Hessenberg-Form*), wenn $a_{ij} = 0$ für $1 \leq j \leq i - 2$. Eine obere Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$ heißt *unreduziert*, wenn sämtliche Subdiagonalelemente $a_{i+1,i}$, $i = 1, \dots, n - 1$, von Null verschieden sind.

Eine (obere) Hessenberg-Matrix (in Zukunft werden wir das Wort „obere“ in diesem Zusammenhang weglassen) hat daher unterhalb der Subdiagonalen keine von Null verschiedenen Elemente. Eine *symmetrische* Hessenberg-Matrix ist offenbar notwendig eine (symmetrische) Tridiagonalmatrix.

Wir wollen uns überlegen, daß man eine gegebene Matrix $A \in \mathbb{R}^{n \times n}$ durch $n - 2$ Ähnlichkeitstransformationen mit (symmetrischen und orthogonalen) Householder-Matrizen P_1, \dots, P_{n-2} auf Hessenberg-Form transformieren kann. Die Idee hierzu ist so einfach wie die entsprechende zur Berechnung der QR-Zerlegung nach Householder, siehe Unterabschnitt 1.5.2. Wir erinnern daran, daß eine (reelle) Householder-Matrix $P \in \mathbb{R}^{n \times n}$ die Form

$$P = I - \frac{2uu^T}{u^Tu} \quad \text{mit } u \in \mathbb{R}^n \setminus \{0\}$$

hat. Ferner wissen wir, daß es bei gegebenem $x \in \mathbb{R}^n \setminus \{0\}$ eine Householder-Matrix gibt, die x in ein Vielfaches des ersten Einheitsvektors überführt (siehe Lemma 5.3 in Abschnitt 1.5). Um dem Leser das Blättern zu ersparen, zitieren wir das entsprechende, leicht beweisbare Ergebnis hier noch einmal:

Lemma 2.2 Ist $x \in \mathbb{R}^n \setminus \{0\}$ und definiert man $u := x + \operatorname{sign}(x_1)\|x\|_2 e_1$, so ist durch

$$P := I - \frac{2uu^T}{u^Tu} = I - \beta uu^T \quad \text{mit } \beta := \frac{2}{u^Tu} = \frac{1}{\|x\|_2(\|x\|_2 + |x_1|)}$$

eine Householder-Matrix mit $Px = -\operatorname{sign}(x_1)\|x\|_2 e_1$ gegeben.

Angenommen, es seien schon $k - 1$ Householder-Matrizen P_1, \dots, P_{k-1} mit

$$P_{k-1} \cdots P_1 AP_1 \cdots P_{k-1} = \left(\begin{array}{c|c} H_k & B_k \\ \hline 0 & a_k \\ \hline & C_k \end{array} \right) \quad \begin{array}{l} \} k \\ \} n - k \end{array}$$

bestimmt, wobei $H_k \in \mathbb{R}^{k \times k}$ eine Hessenberg-Matrix und a_k ein Vektor mit $n - k$ Komponenten ist. Für P_k mache man den Ansatz

$$P_k = \operatorname{diag}(I_k, \overline{P}_k) = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \overline{P}_k \end{array} \right)$$

mit der $k \times k$ -Identität I_k und einer $(n - k) \times (n - k)$ -Householder-Matrix \overline{P}_k . Dann ist

$$P_k \cdots P_1 AP_1 \cdots P_k = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \overline{P}_k \end{array} \right) \left(\begin{array}{c|c} H_k & B_k \\ \hline 0 & a_k \\ \hline & C_k \end{array} \right) \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \overline{P}_k \end{array} \right)$$

$$\begin{aligned}
 &= \left(\begin{array}{c|c} H_k & B_k \bar{P}_k \\ \hline 0 & \bar{P}_k a_k \end{array} \right) \\
 &= \left(\begin{array}{c|c} H_{k+1} & B_{k+1} \\ \hline 0 & a_{k+1} \end{array} \right) \} k+1 \\
 &\quad \left(\begin{array}{c|c} & C_{k+1} \end{array} \right) \} n-k-1
 \end{aligned}$$

mit der Hessenberg-Matrix $H_{k+1} \in \mathbb{R}^{(k+1) \times (k+1)}$, wenn die Householder-Matrix \bar{P}_k so gewählt wird, daß $\bar{P}_k a_k$ ein Vielfaches des ersten Einheitsvektors ist. Dies ist nach Lemma 2.2 möglich, falls $a_k \neq 0$.

Im Prinzip erhält man den folgenden Algorithmus zur Transformation einer beliebigen Matrix $A \in \mathbb{R}^{n \times n}$ mit Hilfe von $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen auf Hessenberg-Form.

- Input: Gegeben $A = (a_{ij}) \in \mathbb{R}^{n \times n}$.

- Für $k = 1, \dots, n-2$:

Falls $a_k := (a_{k+1,k}, \dots, a_{n,k})^T \neq 0$, dann:

Berechne die Householder-Matrix $\bar{P}_k \in \mathbb{R}^{(n-k) \times (n-k)}$ durch

$$\begin{aligned}
 u^k &:= (a_{k+1,k} + \text{sign}(a_{k+1,k}) \|a_k\|_2, a_{k+2,k}, \dots, a_{n,k})^T \\
 \beta_k &:= \frac{1}{\|a_k\|_2 (\|a_k\|_2 + |a_{k+1,k}|)} \\
 \bar{P}_k &:= I_{n-k} - \beta_k u^k (u^k)^T
 \end{aligned}$$

Setze $P_k := \text{diag}(I_k, \bar{P}_k)$ und berechne $A := P_k A P_k$.

Andernfalls: Setze $P_k := I$.

- Output: Die Ausgangsmatrix A wird in $n-2$ Schritten mit der orthogonal ähnlichen Hessenberg-Matrix $P^T A P$ überschrieben. Hierbei ist $P := P_1 \cdots P_{n-2}$.

Insgesamt erhalten wir damit

Satz 2.3 Zu einer beliebigen Matrix $A \in \mathbb{R}^{n \times n}$ existieren $n-2$ Householder-Matrizen P_1, \dots, P_{n-2} derart, daß $P_{n-2} \cdots P_1 A P_1 \cdots P_{n-2}$ eine zu A orthogonal ähnliche Hessenberg-Matrix ist.

Nun soll noch etwas genauer auf eine mögliche Implementation des obigen Verfahrens eingegangen werden. Mit $n-2$ Ähnlichkeitstransformationen durch die Householder-Matrizen P_1, \dots, P_{n-2} wird die Ausgangsmatrix A in die Hessenberg-Matrix $H = P_{n-2} \cdots P_1 A P_1 \cdots P_{n-2}$ überführt. Als ähnliche Matrizen haben A und H dieselben Eigenwerte. Ist ferner x ein Eigenvektor von H , so erhält man durch $y := P_1 \cdots P_{n-2} x$ einen Eigenvektor (zum gleichen Eigenwert) von A . Im k -ten Schritt können die relevanten Informationen über $P_k = \text{diag}(I_k, \bar{P}_k)$ mit $\bar{P}_k = I_{n-k} - \beta_k u^k (u^k)^T$, also $u^k = (u_{k+1}^k, u_{k+2}^k, \dots, u_n^k)^T$ und β_k , weitgehend in den gerade frei werdenden bzw. annullierten Stellen von A gespeichert werden. Wie bei der QR -Zerlegung einer Matrix nach Householder fehlt auch hier wieder etwas Platz. Daher liegt es nahe,

$(u_{k+2}^k, \dots, u_n^k)$ in den frei werdenden $n - k - 1$ Stellen der k -ten Spalte von A und u_{k+1}^k und β_k für $k = 1, \dots, n - 2$ jeweils als k -te Komponente in einem gesonderten Feld d bzw. β der Länge $n - 2$ zu speichern. Als Output-Daten nach Abschluß des Verfahrens erhält man daher:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2,n-1} & a_{2n} \\ u_3^1 & a_{32} & \cdots & \cdots & a_{3,n-1} & a_{3n} \\ u_4^1 & u_4^2 & \ddots & \cdots & a_{4,n-1} & a_{4n} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ u_n^1 & u_n^2 & \cdots & u_n^{n-2} & a_{n,n-1} & a_{nn} \end{pmatrix}, \quad d = \begin{pmatrix} u_2^1 \\ u_3^2 \\ \vdots \\ u_{n-1}^{n-2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-2} \end{pmatrix}.$$

Insgesamt resultiert der folgende Algorithmus zur Transformation einer Matrix auf obere Hessenberg-Form mit Hilfe von Householder-Matrizen. Hierbei wird wie bei dem entsprechenden Algorithmus zur QR-Zerlegung nach Householder (siehe Unterabschnitt 1.5.2) ausgenutzt, daß der Vektor u^k mit einem Faktor multipliziert werden kann, ohne daß dabei \bar{P}_k bzw. P_k verändert werden. Ferner ist es praktisch, zunächst u_{k+1}^k an der Stelle $a_{k+1,k}$ zu speichern und dies erst am Schluß des k -ten Schrittes zu korrigieren.

- Input: Gegeben ist die Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$.
- Für $k = 1, \dots, n - 2$:
 - $\|a_k\|_\infty := \max_{i=k+1, \dots, n} |a_{ik}|$ {Beginn der Berechnung von \bar{P}_k }
 - Falls $\|a_k\|_\infty = 0$, dann: $d_k := 0, \beta_k := 0$
 - Andernfalls:
 - $\alpha := 0$
 - Für $i = k + 1, \dots, n$:
 - $a_{ik} := a_{ik} / \|a_k\|_\infty, \alpha := \alpha + a_{ik}^2$
 - $\alpha := \sqrt{\alpha}, \beta_k := 1 / [\alpha(\alpha + |a_{k+1,k}|)]$
 - $d_k := -\text{sign}(a_{k+1,k})\alpha \|a_k\|_\infty, a_{k+1,k} := a_{k+1,k} + \text{sign}(a_{k+1,k})\alpha$ {Ende}
 - Für $j = k + 1, \dots, n$: {Beginn der Berechnung von $\bar{P}_k C_k$ }
 - $s := \beta_k \sum_{i=k+1}^n a_{ik} a_{ij}$
 - Für $i = k + 1, \dots, n$:
 - $a_{ij} := a_{ij} - s a_{ik}$ {Ende}
 - Für $i = 1, \dots, n$: {Beginn der Berechnung von $B_k \bar{P}_k$ und $\bar{P}_k C_k \bar{P}_k$ }
 - $s := \beta_k \sum_{j=k+1}^n a_{ij} a_{jk}$
 - Für $j = k + 1, \dots, n$:
 - $a_{ij} := a_{ij} - s a_{jk}$ {Ende}
- Vertausche $a_{k+1,k}$ und d_k .

- Output: Die Matrix A wird überschrieben, ferner werden zwei Felder d und β der Länge $n - 2$ ausgegeben. Setzt man $H := (h_{ij})$ mit

$$h_{ij} := \begin{cases} a_{ij} & \text{für } j \geq i - 1, \\ 0 & \text{für } j \leq i - 2, \end{cases}$$

so ist H eine zu der Ausgangsmatrix ähnliche Hessenberg-Matrix. Die zu der Transformation benötigten $n - 2$ Householder-Matrizen P_k erhält man aus

$$u^k := (d_k, a_{k+2,k}, \dots, a_{nk})^T, \quad \bar{P}_k := I_{n-k} - \beta_k u^k (u^k)^T, \quad P_k = \text{diag}(I_k, \bar{P}_k).$$

Nun wollen wir noch auf den Fall eingehen, daß die Ausgangsmatrix A symmetrisch ist, und den obigen Algorithmus hierauf spezialisieren. Bei *orthogonalen Ähnlichkeitstransformationen*, etwa mit den oben benutzten Householder-Matrizen, bleibt die Symmetrie erhalten. Daher läßt sich eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ durch $n - 2$ Ähnlichkeitstransformationen mit Householder-Matrizen P_1, \dots, P_{n-2} auf die ähnliche (symmetrische) Tridiagonalmatrix $P_{n-2} \cdots P_1 A P_1 \cdots P_{n-2}$ transformieren.

Angenommen, es seien schon $k - 1$ Householder-Matrizen P_1, \dots, P_{k-1} bestimmt. Es sei

$$P_{k-1} \cdots P_1 A P_1 \cdots P_{k-1} = \left(\begin{array}{c|c} H_k & \begin{matrix} 0^T \\ a_k^T \end{matrix} \\ \hline 0 & C_k \end{array} \right) \begin{matrix} \} k-1 \\ \} 1 \\ \} n-k \end{matrix}$$

mit einer (symmetrischen) $k \times k$ -Tridiagonalmatrix H_k , einem $(n - k)$ -Vektor a_k und einer $(n - k) \times (n - k)$ -Matrix C_k . Wie im allgemeinen Fall mache man für P_k den Ansatz

$$P_k = \text{diag}(I_k, \bar{P}_k) = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \bar{P}_k \end{array} \right)$$

mit der $k \times k$ -Identität I_k und einer $(n - k) \times (n - k)$ -Householder-Matrix \bar{P}_k . Dann ist

$$P_k \cdots P_1 A P_1 \cdots P_k = \left(\begin{array}{c|c} H_k & \begin{matrix} 0^T \\ a_k^T \bar{P}_k \end{matrix} \\ \hline 0 & \begin{matrix} \bar{P}_k a_k \\ \bar{P}_k C_k \bar{P}_k \end{matrix} \end{array} \right).$$

Wieder wird man also \bar{P}_k so wählen, daß $\bar{P}_k a_k$ ein Vielfaches des ersten Einheitsvektors ist. Die Symmetrie der Ausgangsmatrix A wird bei der anschließenden Berechnung der transformierten Matrix $P_k \cdots P_1 A P_1 \cdots P_k$ zweimal ausgenutzt. Zum einen weiß man, daß der rechte obere $(k - 1) \times (n - k)$ -Block eine Nullmatrix ist. Zum anderen kann die Symmetrie von A bzw. von C_k auch zur effizienten Berechnung von $\bar{P}_k C_k \bar{P}_k$ verwandt werden. Ist nämlich $\bar{P}_k = I_{n-k} - \beta_k u^k (u^k)^T$ mit $\beta_k = 2/(u^k)^T u^k$, definiert man ferner $p^k := \beta_k C_k u^k$, so ist

$$\begin{aligned} \bar{P}_k C_k \bar{P}_k &= [I - \beta_k u^k (u^k)^T] C_k [I - \beta_k u^k (u^k)^T] \\ &= [I - \beta_k u^k (u^k)^T] [C_k - p^k (u^k)^T] \\ &= C_k - u^k (p^k)^T - p^k (u^k)^T + \beta_k (u^k)^T p^k u^k (u^k)^T \\ &= C_k - u^k (w^k)^T - w^k (u^k)^T \end{aligned}$$

mit

$$w^k := p^k - \frac{\beta_k (u^k)^T p^k}{2} u^k.$$

Nun sollen noch einige Bemerkungen zu einer möglichen Implementation des Verfahrens gemacht werden. Hier liegt es nahe, die resultierende (symmetrische) Tridiagonalmatrix mit den Hauptdiagonalelementen $\delta_1, \dots, \delta_n$ und den Nebendiagonalelementen $\gamma_1, \dots, \gamma_{n-1}$ in zwei Feldern δ und γ der Länge n bzw. $n-1$ zu speichern. Die zur Transformation benutzten Householder-Matrizen $\bar{P}_k = I_{n-k} - \beta_k u^k (u^k)^T$ können in der k -ten Spalte von A gespeichert werden, indem man β_k in das k -te Diagonalelement von A und den $(n-k)$ -Vektor u^k darunter speichert. Ferner werden die für die Transformation $\bar{P}_k C_k \bar{P}_k$ benötigten $(n-k)$ -Vektoren p^k bzw. w^k auf den letzten $(n-k)$ (im k -ten Schritt noch nicht fest gelegten) Komponenten des Feldes γ gespeichert. Insgesamt erhält man den folgenden Algorithmus.

- Input: Gegeben ist eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$.

- Für $k = 1, \dots, n-2$:

$$\delta_k := a_{kk} \quad \{ \text{Bestimme Hauptdiagonalelement} \}$$

$$\|a_k\|_\infty := \max_{i=k+1, \dots, n} |a_{ik}| \quad \{ \text{Beginn der Berechnung von } \bar{P}_k \}$$

$$\text{Falls } \|a_k\|_\infty = 0, \text{ dann: } a_{kk} := 0, \quad \gamma_k := 0$$

Andernfalls:

$$\alpha := 0$$

Für $i = k+1, \dots, n$:

$$a_{ik} := a_{ik} / \|a_k\|_\infty, \quad \alpha := \alpha + a_{ik}^2$$

$$\alpha := \sqrt{\alpha}, \quad a_{kk} := 1 / [\alpha(\alpha + |a_{k+1,k}|)]$$

$$\gamma_k := -\text{sign}(a_{k+1,k})\alpha \|a_k\|_\infty, \quad a_{k+1,k} := a_{k+1,k} + \text{sign}(a_{k+1,k})\alpha \quad \{ \text{Ende} \}$$

$$s := 0 \quad \{ \text{Beginn der Berechnung der unteren Hälfte von } \bar{P}_k C_k \bar{P}_k \}$$

Für $i = k+1, \dots, n$:

$$\gamma_i := a_{kk} \left(\sum_{j=k+1}^i a_{ij} a_{jk} + \sum_{j=i+1}^n a_{ji} a_{jk} \right)$$

$$s := s + \gamma_i a_{ik}$$

$$s := a_{kk} s / 2$$

Für $i = k+1, \dots, n$:

$$\gamma_i := \gamma_i - s a_{ik}$$

Für $i = k+1, \dots, n$:

Für $j = k+1, \dots, i$:

$$a_{ij} := a_{ij} - a_{ik} \gamma_j - a_{jk} \gamma_i \quad \{ \text{Ende} \}$$

$$\delta_{n-1} := a_{n-1,n-1}, \quad \delta_n := a_{nn}, \quad \gamma_{n-1} := a_{n,n-1}$$

- Output: In $\delta = (\delta_1, \dots, \delta_n)^T$ bzw. in $\gamma = (\gamma_1, \dots, \gamma_{n-1})^T$ wird die Hauptdiagonale bzw. die Nebendiagonale der resultierenden, zur Ausgangsmatrix orthogonal ähnlichen, symmetrischen Tridiagonalmatrix ausgegeben. Die linke untere Hälfte von A (einschließlich der Diagonalen) wird in den ersten $n - 2$ Spalten mit den relevanten Informationen über die benutzten Householder-Matrizen $\bar{P}_k = I_{n-k} - \beta_k u^k (u^k)^T$ überschrieben. Die (strikte) obere Hälfte von A wird nicht benutzt und nicht verändert.

5.2.2 Die QR-Zerlegung einer Hessenberg-Matrix

Zu Beginn dieses Abschnittes hatten wir angegeben, daß das QR-Verfahren zur Berechnung der Eigenwerte einer Matrix $A \in \mathbb{R}^{n \times n}$ im Prinzip folgendermaßen aussieht:

- Überführe A durch $n - 2$ Ähnlichkeitstransformationen (z. B. mit Householder-Matrizen) in die Hessenberg-Matrix A_1 .
- Für $k = 1, 2, \dots$

Wähle bzw. bestimme einen „Shift-Parameter“ $\sigma_k \in \mathbb{R}$, bilde eine QR-Zerlegung $A_k - \sigma_k I = Q_k R_k$ (mit einer orthogonalen Matrix Q_k und einer oberen Dreiecksmatrix R_k) und berechne $A_{k+1} := R_k Q_k + \sigma_k I$.

Bei einer voll besetzten $n \times n$ -Matrix benötigt man zur Berechnung einer zugehörigen QR-Zerlegung im wesentlichen $\frac{2}{3} n^3$ (QR-Zerlegung nach Householder) bzw. $\frac{4}{3} n^3$ (QR-Zerlegung nach Givens) Multiplikationen oder „flops“ (eine Multiplikation, eine Addition und eine Zuweisung). Wir wollen uns überlegen, daß man bei einer $n \times n$ -Hessenberg-Matrix diese Anzahl auf ein Vielfaches von n^2 drücken kann, und einen Algorithmus angeben, der die folgende Aufgabenstellung effizient löst:

- Input: Gegeben sei die Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$ und ein Shift-Parameter $\sigma \in \mathbb{R}$.
- Bestimme eine orthogonale Matrix Q derart, daß $A - \sigma I = QR$ mit einer oberen Dreiecksmatrix R . Anschließend berechne man

$$A_+ := RQ + \sigma I = Q^T(A - \sigma I)Q + \sigma I = Q^T A Q.$$

(Hierbei sollte A_+ ebenfalls eine Hessenberg-Matrix sein, damit man das entsprechende Verfahren im nächsten Schritt auch auf A_+ anwenden kann.)

- Output: Die Matrix A wird überschrieben durch $A_+ := Q^T A Q$, also eine zur Ausgangsmatrix orthogonal ähnliche (Hessenberg-) Matrix. Ist man bei der gegebenen Eigenwertaufgabe auch an den Eigenvektoren interessiert, so sollte man sich die Matrix Q merken.

Zur Lösung dieser Aufgabenstellung liegt es nahe, eine geeignete Modifikation der QR-Zerlegung nach Givens (siehe Unterabschnitt 1.5.3) anzuwenden. Zu Recht wird man vermuten, daß man durch Multiplikation der Hessenberg-Matrix $A - \sigma I$ von links

mit $n-1$ Givens-Rotationen der Reihe nach die $n-1$ Subdiagonalelemente annullieren kann. Zur Erinnerung: Die sogenannten *Givens-Rotationen* sind orthogonale $n \times n$ -Matrizen der Form

$$G_{ik} = \begin{pmatrix} 1 & & & & & \\ & \vdots & & & & \\ & \cdots & c & \cdots & s & \cdots \\ & & \vdots & & \vdots & \\ & \cdots & -s & \cdots & c & \cdots \\ & & \vdots & & \vdots & 1 \end{pmatrix} \quad \begin{matrix} i \\ k \\ i \\ k \end{matrix}$$

mit $c^2 + s^2 = 1$ und $1 \leq i < k \leq n$. Ist $x \in \mathbb{R}^n$ und $y := G_{ik}x$, so ist

$$y_j = \begin{cases} cx_i + sx_k & \text{für } j = i, \\ -sx_i + cx_k & \text{für } j = k, \\ x_j & \text{für } j \neq i, k. \end{cases}$$

Multipliziert man eine Matrix $A \in \mathbb{R}^{n \times n}$ von links mit der Givens-Rotation G_{ik} , so bewirkt dies lediglich eine Veränderung der i -ten und der k -ten Zeile. Die neuen Zeilen sind eine Linearkombination der alten und gegeben durch

$$(G_{ik}A)_{ij} = ca_{ij} + sa_{kj}, \quad (G_{ik}A)_{kj} = -sa_{ij} + ca_{kj}, \quad j = 1, \dots, n.$$

Insbesondere gilt: Ist $a_{ij} = a_{kj} = 0$, so ist auch $(G_{ik}A)_{ij} = (G_{ik}A)_{kj} = 0$.

Grundlegend für die QR-Zerlegung nach Givens ist, daß man bei vorgegebenen $1 \leq i < k \leq n$ und $x \in \mathbb{R}^n$ eine Givens-Rotation G_{ik} mit $(G_{ik}x)_k = 0$ bestimmen kann. Durch eine Rotation in der (i, k) -Ebene kann also die k -te Komponente von $y := G_{ik}x$ zu Null gemacht werden, wobei außer der i -ten Komponente alle anderen unverändert bleiben. Hierzu ist es praktisch, sich eine Funktion "rot" zu definieren, die zu vorgegebenen $(\alpha, \beta) \in \mathbb{R}^2$ Konstanten c und s mit $c^2 + s^2 = 1$ sowie γ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

bestimmt. Es ist leicht zu sehen, daß eine Lösung dieser Aufgabe durch

$$c := \pm \frac{\alpha}{(\alpha^2 + \beta^2)^{1/2}}, \quad s := \pm \frac{\beta}{(\alpha^2 + \beta^2)^{1/2}}, \quad \gamma := \pm(\alpha^2 + \beta^2)^{1/2}$$

gegeben ist. Dies kann durch den folgenden Algorithmus realisiert werden.

- Input: Gegeben $\alpha, \beta \in \mathbb{R}$.
- Falls $\beta = 0$, dann: $c := 1, s := 0, \gamma := \alpha$

Andernfalls:

Falls $|\beta| \geq |\alpha|$, dann:

$$t := \alpha/\beta, \quad s := 1/(1+t^2)^{1/2}, \quad c := st, \quad \gamma := \beta(1+t^2)^{1/2}$$

Andernfalls:

$$t := \beta/\alpha, \quad c := 1/(1+t^2)^{1/2}, \quad s := ct, \quad \gamma := \alpha(1+t^2)^{1/2}$$

- Output: Für $\text{rot}(\alpha, \beta) := (c, s, \gamma)$ gilt $c^2 + s^2 = 1$ und

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}.$$

Es ist einfach einzusehen, daß man eine Hessenberg-Matrix A (der Übersichtlichkeit halber sehen wir zunächst von dem Shift-Parameter σ ab bzw. nehmen $\sigma = 0$ an) durch sukzessive Multiplikation von links mit Givens-Rotationen $G_{12}, G_{23}, \dots, G_{n-1,n}$ auf obere Dreiecksgestalt transformieren kann. Denn angenommen, $G_{12}, \dots, G_{k-1,k}$ seien schon so bestimmt, daß

$$A^{(k-1)} := G_{k-1,k} \cdots G_{12} A = \left(\begin{array}{c|c} R_{k-1} & * \\ \hline 0 & H_{k-1} \end{array} \right) \quad \begin{array}{l} \} k-1 \\ \} n-k+1 \end{array}$$

mit einer oberen Dreiecksmatrix $R_{k-1} \in \mathbb{R}^{(k-1) \times (k-1)}$ und einer Hessenberg-Matrix $H_{k-1} \in \mathbb{R}^{(n-k+1) \times (n-k+1)}$. Für $k=1$ ist das offenbar mit $H_0 = A$ der Fall. Sei $(a_{kk}, a_{k+1,k}, 0, \dots, 0)^T$ die erste Spalte von H_{k-1} . Man bestimme eine Givens-Rotation $G_{k,k+1} = G_{k,k+1}(c_k, s_k)$, deren Anwendung auf die k -te Spalte von $A^{(k-1)}$ das Element $a_{k+1,k}$ annulliert, d. h. man berechne $(c_k, s_k, a_{kk}) := \text{rot}(a_{kk}, a_{k+1,k})$. Gegenüber $A^{(k-1)}$ verändern sich in $A^{(k)} := G_{k,k+1}A^{(k-1)}$ nur die k -te und die $(k+1)$ -te Zeile. Schon erzeugte Nullen in den Spalten $1, \dots, k-1$ bleiben erhalten. Daher ist

$$A^{(k)} = G_{k,k+1}G_{k-1,k} \cdots G_{12} A = \left(\begin{array}{c|c} R_k & * \\ \hline 0 & H_k \end{array} \right) \quad \begin{array}{l} \} k \\ \} n-k \end{array}$$

mit einer oberen Dreiecksmatrix $R_k \in \mathbb{R}^{k \times k}$ und einer $(n-k) \times (n-k)$ -Hessenberg-Matrix H_k . Der Aufwand zur Berechnung von $A^{(k)}$ aus $A^{(k-1)}$ (wenn wir einmal von der Auswertung von $\text{rot}(a_{kk}, a_{k+1,k})$ absehen) beträgt $4(n-k)$ Multiplikationen bzw. „flops“. Zur Berechnung von

$$\begin{aligned} A^{(n-1)} &= \underbrace{G_{n-1,n} \cdots G_{12}}_{=: Q^T} A = R \\ &= Q^T \end{aligned}$$

aus A werden daher im wesentlichen $2n^2$ Multiplikationen benötigt (sowie $n-1$ Quadratwurzeln).

Nun ist man aber in einem Schritt des QR -Verfahrens eigentlich nicht an der oberen Dreiecksmatrix R , sondern an

$$A_+ = Q^T A Q = R Q = R G_{12}^T \cdots G_{n-1,n}^T$$

(wir nehmen immer noch an, es sei $\sigma = 0$) und $Q = G_{12}^T \cdots G_{n-1,n}^T$ (wenn auch die Eigenvektoren berechnet werden sollen) interessiert. Außerdem bleibt noch zu zeigen, daß A_+ wieder eine Hessenberg-Matrix ist. Letzteres ist leicht einzusehen. Angenommen, $RG_{12}^T \cdots G_{k-1,k}^T$ habe die Form

$$RG_{12}^T \cdots G_{k-1,k}^T = \left(\begin{array}{c|c} H_k & * \\ \hline 0 & R_k \end{array} \right) \quad \} k \quad } n-k$$

mit einer Hessenberg-Matrix $H_k \in \mathbb{R}^{k \times k}$ und einer oberen $(n-k) \times (n-k)$ -Dreiecksmatrix R_k . Für $k=1$ ist dies offenbar der Fall. Eine Multiplikation mit $G_{k,k+1}^T$ von rechts verändert nur die k -te und die $(k+1)$ -te Spalte, die neuen Spalten sind Linearkombinationen der alten. Daher ist

$$RG_{12}^T \cdots G_{k-1,k}^T G_{k,k+1}^T = \left(\begin{array}{c|c} H_{k+1} & * \\ \hline 0 & R_{k+1} \end{array} \right) \quad } k+1 \quad } n-k-1$$

mit einer Hessenberg-Matrix $H_{k+1} \in \mathbb{R}^{(k+1) \times (k+1)}$ und einer oberen Dreiecksmatrix $R_{k+1} \in \mathbb{R}^{(n-k-1) \times (n-k-1)}$. Daher ist $A_+ = RG_{12}^T \cdots G_{n-1,n}^T = H_n$ eine Hessenberg-Matrix. Für die Berechnung von A_+ aus R benötigt man im wesentlichen $2n^2$ Multiplikationen bzw. „flops“, so daß man insgesamt für einen Schritt im QR-Verfahren (angewandt auf eine Hessenberg-Matrix) $4n^2$ Multiplikationen benötigt. Wir fassen das erhaltene Ergebnis in einem Satz zusammen.

Satz 2.4 Sei $A \in \mathbb{R}^{n \times n}$ eine Hessenberg-Matrix. Dann können $n-1$ Givens-Rotationen $G_{k,k+1} = G_{k,k+1}(c_k, s_k)$ so bestimmt werden, daß mit der orthogonalen Matrix $Q := G_{12}^T \cdots G_{n-1,n}^T$ gilt:

1. $Q^T A = R$ ist eine obere Dreiecksmatrix.
2. $A_+ = RQ = Q^T AQ$ ist eine Hessenberg-Matrix.

Zur Berechnung von A_+ aus A benötigt man im wesentlichen $4n^2$ Multiplikationen.

Bemerkungen: Die oben beschriebene Vorgehensweise bei der Berechnung von A_+ aus A hat den Nachteil, daß man sich die Givens-Rotationen $G_{k,k+1}$ bzw. (c_k, s_k) , $k = 1, \dots, n-1$, merken muß, um nach der Transformation von A auf die obere Dreiecksgestalt $R = G_{n-1,n} \cdots G_{12}A$ auch von rechts mit $G_{12}^T, \dots, G_{n-1,n}^T$ multiplizieren zu können. Dies kann weitgehend vermieden werden, indem man nach der Berechnung von $G_{23}G_{12}A$ schon mit G_{12}^T von rechts multipliziert (weitere Multiplikationen von links verändern nämlich die ersten beiden Spalten nicht mehr) und danach abwechselnd von links und rechts multipliziert:

$$G_{23}G_{12}AG_{12}^T \rightarrow G_{34}G_{23}G_{12}AG_{12}^T \rightarrow G_{34}G_{23}G_{12}AG_{12}^TG_{23}^T \rightarrow \dots$$

Nach der Berechnung von $G_{n-1,n} \cdots G_{12}AG_{12}^T \cdots G_{n-2,n-1}^T$ muß am Schluß noch mit $G_{n-1,n}^T$ von rechts multipliziert werden.

Der Einbau eines Shift-Parameters σ ist nicht schwierig. Natürlich kann man zunächst $A - \sigma I$ bilden, dann einen QR-Schritt machen und schließlich σ zu den neuen

Diagonalelementen wieder hinzu addieren. Ein Programm wird noch etwas einfacher, wenn σ von $a_{k+1,k+1}$ gerade vor der Linksmultiplikation mit $G_{k,k+1}$ subtrahiert wird und nach der Multiplikation von rechts mit $G_{k-1,k}^T$ zu $a_{k-1,k-1}$ wieder hinzu addiert wird. \square

Ein Programm, das die eben gemachten Bemerkungen berücksichtigt, könnte folgendermaßen aussehen (siehe auch G. W. STEWART (1973, S. 360) und H. R. SCHWARZ (1988, S. 269)):

- Input: Gegeben sei die Hessenberg-Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ und ein Shift-Parameter $\sigma \in \mathbb{R}$. Benutzt und verändert werden nur die Elemente a_{ij} von A mit $j \geq i - 1$.
- $a_{11} := a_{11} - \sigma$

Für $k = 1, \dots, n$:

Falls $k < n$, dann:

{Bestimme $G_{k,k+1}$, subtrahiere Shift-Parameter und multipliziere von links mit $G_{k,k+1}$.}

$$(c_{\text{neu}}, s_{\text{neu}}, a_{kk}) := \text{rot}(a_{kk}, a_{k+1,k}), \quad a_{k+1,k} := 0$$

$$a_{k+1,k+1} := a_{k+1,k+1} - \sigma$$

Für $j := k + 1, \dots, n$:

$$\begin{pmatrix} a_{kj} \\ a_{k+1,j} \end{pmatrix} := \begin{pmatrix} c_{\text{neu}} & s_{\text{neu}} \\ -s_{\text{neu}} & c_{\text{neu}} \end{pmatrix} \begin{pmatrix} a_{kj} \\ a_{k+1,j} \end{pmatrix}$$

Falls $k > 1$, dann:

{Multipliziere mit $G_{k-1,k}^T$ von rechts und addiere Shift-Parameter.}

Für $i = 1, \dots, k$:

$$(a_{i,k-1}, a_{i,k}) := (a_{i,k-1}, a_{i,k}) \begin{pmatrix} c_{\text{alt}} & -s_{\text{alt}} \\ s_{\text{alt}} & c_{\text{alt}} \end{pmatrix}$$

$$a_{k-1,k-1} := a_{k-1,k-1} + \sigma$$

$$c_{\text{alt}} := c_{\text{neu}}, \quad s_{\text{alt}} := s_{\text{neu}}$$

$$a_{nn} := a_{nn} + \sigma$$

- Output: Die Ausgangsmatrix A wird mit der orthogonal ähnlichen Hessenberg-Matrix $Q^T A Q$ überschrieben. Hierbei ist $A - \sigma I = QR$ mit einer oberen Dreiecksmatrix R und daher $Q^T A Q = RQ + \sigma I$. Ferner ist $Q = G_{12}^T \cdots G_{n-1,n}^T$ mit Givens-Rotationen $G_{k,k+1}$, $k = 1, \dots, n - 1$.

Ist A eine symmetrische Hessenberg-Matrix und damit eine (symmetrische) Tridiagonalmatrix, so kann der Aufwand für einen QR -Schritt noch wesentlich verringert werden. Hierauf werden wir in 5.3.3 eingehen.

5.2.3 Vektoriteration nach v. Mises

Zu einem vom Konzept her außerordentlich einfachen Verfahren, der Vektoriteration nach v. Mises, zur Bestimmung des dominanten Eigenwertes einer Matrix $A \in \mathbb{C}^{n \times n}$ und eines zugehörigen Eigenvektors, sollen nun einige Bemerkungen gemacht werden.

Sei $A \in \mathbb{C}^{n \times n}$ eine *diagonalähnliche* bzw. *diagonalsierbare* Matrix, d. h. es existiere ein System linear unabhängiger Eigenvektoren u_1, \dots, u_n zu den Eigenwerten $\lambda_1, \dots, \lambda_n$ von A . Es sei $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$. Man sagt dann, λ_1 sei ein *dominanter* Eigenwert von A . Sei $\|\cdot\|$ eine vorgegebene Vektornorm im \mathbb{C}^n , z. B. $\|\cdot\| = \|\cdot\|_\infty$, und¹ $y^{(0)} = \sum_{j=1}^n \alpha_j u_j$ mit $\alpha_1 \neq 0$ eine Näherung für einen zu λ_1 gehörenden Eigenvektor. Der Ausgangsvektor $y^{(0)}$ habe also eine Basisdarstellung durch die Eigenvektoren von A mit einer von Null verschiedenen Komponente bezüglich u_1 . Dann ist

$$A^k y^{(0)} = \sum_{j=1}^n \alpha_j \lambda_j^k u_j = \lambda_1^k \left[\alpha_1 u_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k u_2 + \dots + \alpha_n \left(\frac{\lambda_n}{\lambda_1} \right)^k u_n \right].$$

Ist $(u_1)_l \neq 0$, die l -te Komponente von u_1 also von Null verschieden, so ist

$$\begin{aligned} \frac{(A^{k+1} y^{(0)})_l}{(A^k y^{(0)})_l} &= \frac{\lambda_1^{k+1} [\alpha_1 (u_1)_l + \alpha_2 (\lambda_2/\lambda_1)^{k+1} (u_2)_l + \dots + \alpha_n (\lambda_n/\lambda_1)^{k+1} (u_n)_l]}{\lambda_1^k [\alpha_1 (u_1)_l + \alpha_2 (\lambda_2/\lambda_1)^k (u_2)_l + \dots + \alpha_n (\lambda_n/\lambda_1)^k (u_n)_l]} \\ &= \lambda_1 \frac{[\alpha_1 (u_1)_l + \alpha_2 (\lambda_2/\lambda_1)^{k+1} (u_2)_l + \dots + \alpha_n (\lambda_n/\lambda_1)^{k+1} (u_n)_l]}{[\alpha_1 (u_1)_l + \alpha_2 (\lambda_2/\lambda_1)^k (u_2)_l + \dots + \alpha_n (\lambda_n/\lambda_1)^k (u_n)_l]} \\ &\rightarrow \lambda_1 \quad \text{mit } k \rightarrow \infty. \end{aligned}$$

Ist $\lambda_1 = |\lambda_1| e^{i\phi}$ mit $\phi \in [0, 2\pi)$ (man beachte: Ist $A \in \mathbb{R}^{n \times n}$ reell und λ_1 ein dominanter Eigenwert, so ist λ_1 notwendig reell und daher $\phi = 0$ oder $\phi = \pi$), so ist

$$e^{-ik\phi} \frac{A^k y^{(0)}}{\|A^k y^{(0)}\|} = \frac{|\lambda_1|^k [\alpha_1 u_1 + \alpha_2 (\lambda_2/\lambda_1)^k u_2 + \dots + \alpha_n (\lambda_n/\lambda_1)^k u_n]}{|\lambda_1|^k \|\alpha_1 u_1 + \alpha_2 (\lambda_2/\lambda_1)^k u_2 + \dots + \alpha_n (\lambda_n/\lambda_1)^k u_n\|} \rightarrow \frac{\alpha_1 u_1}{\|\alpha_1 u_1\|}$$

mit $k \rightarrow \infty$. Definiert man also die Folge $\{x^{(k)}\}$ durch $x^{(k)} := A^k y^{(0)} / \|A^k y^{(0)}\|$, so konvergiert $\{x^{(k)}\}$ der „Richtung nach“ gegen einen zu λ_1 gehörenden normierten Eigenvektor, wobei die Konvergenzgeschwindigkeit durch den Quotienten $q := |\lambda_2| / |\lambda_1|$ bestimmt wird. Je dominanter der Eigenwert λ_1 ist, desto besser ist die Konvergenz.

Das Verfahren sollte nun nicht genau in der angegebenen Weise angewandt werden. Besser ist es, folgende Iterationsvorschrift anzuwenden, bei der nach jedem Schritt bezüglich einer vorgegebenen Vektornorm $\|\cdot\|$ normiert wird.

- Sei $y^{(0)} = \sum_{j=1}^n \alpha_j u_j$ mit $\alpha_1 \neq 0$ gegeben. Ferner sei $(u_1)_l \neq 0$.
 - $x^{(0)} := y^{(0)} / \|y^{(0)}\|$.
 - Für $k = 1, 2, \dots$:
- $$y^{(k)} := Ax^{(k-1)}, \quad \lambda_1^{(k)} := y_l^{(k)} / x_l^{(k-1)}, \quad x^{(k)} := y^{(k)} / \|y^{(k)}\|$$

¹Iterationsindizes bei Näherungen für Eigenvektoren und Eigenwerte schreiben wir in diesem und dem nächsten Unterabschnitt ausnahmsweise in *Klammern* nach oben, um keine Gelegenheit für Verwechslungen zu geben.

Dann ist $x^{(k)} = A^k y^{(0)} / \|A^k y^{(0)}\|$, $k = 0, 1, \dots$, wie man sofort durch vollständige Induktion beweist. Daher ist $y^{(k)} = A^k y^{(0)} / \|A^{k-1} y^{(0)}\|$, $k = 1, \dots$, und folglich

$$\lambda_1^{(k)} = \frac{y_i^{(k)}}{x_i^{(k-1)}} = \frac{(A^k y^{(0)})_i}{(A^{k-1} y^{(0)})_i}, \quad k = 1, \dots$$

Dies bedeutet, daß das scheinbar modifizierte Verfahren mit dem oben angegebenen übereinstimmt, so daß die dort gemachten Konvergenzaussagen gelten.

Die Vor- und Nachteile der Vektoriteration liegen auf der Hand. Ein Vorteil ist die Einfachheit. Ferner wird nicht auf die Elemente von A zugegriffen, sondern es genügt, die Wirkung von A auf einen Vektor zu kennen, was bei großen, schwach besetzten Matrizen von Vorteil sein kann. Schließlich kann die Kenntnis einer guten Näherung für den ersten Eigenvektor ausgenutzt werden. Ein Nachteil ist die schlechte Konvergenz bei nicht oder nur schwach dominierendem ersten Eigenwert, außerdem erhält man natürlich nur eine Näherung für den ersten Eigenwert und einen zugehörigen Eigenvektor. Es gibt Methoden zur Konvergenzverbesserung (z. B. Konvergenzbeschleunigung nach Aitken oder eine Spektralverschiebung, bei der die Vektoriteration mit einem geeigneten p auf $A - pI$ angewandt wird, siehe z. B. J. H. WILKINSON (1965, S. 570 ff.) und K. E. ATKINSON (1978, S. 518 ff.)).

5.2.4 Inverse Iteration nach Wielandt

Nachdem die (einfache) Vektoriteration nach v. Mises zur Bestimmung eines dominanten Eigenwertes und eines zugehörigen Eigenvektors beschrieben wurde, soll nun die Idee der für die Praxis wesentlich wichtigeren inversen Iteration nach Wielandt geschildert werden. Hier geht man davon aus, daß ein „guter“ Näherungswert λ für einen Eigenwert λ_j der diagonalisierbaren Matrix $A \in \mathbb{C}^{n \times n}$ (mit dem System $\{u_1, \dots, u_n\}$ linear unabhängiger Eigenvektoren) bekannt sei, d. h. λ liege wesentlich näher bei λ_j als bei den übrigen λ_i bzw. es gelte

$$|\lambda_j - \lambda| \ll |\lambda_i - \lambda| \quad \text{für } i \neq j.$$

Wir nehmen an, λ sei zwar eine gute Näherung für einen Eigenwert, sei selbst aber keiner. Dann ist $A - \lambda I$ nichtsingulär und $(A - \lambda I)^{-1}$ besitzt die Eigenwerte $1/(\lambda_i - \lambda)$, $i = 1, \dots, n$. Wegen

$$\frac{1}{|\lambda_i - \lambda|} \ll \frac{1}{|\lambda_j - \lambda|} \quad \text{für } i \neq j$$

ist $1/(\lambda_j - \lambda)$ ein dominanter Eigenwert von $(A - \lambda I)^{-1}$. Die jetzt naheliegende Idee der inversen Iteration besteht darin, das Verfahren der Vektoriteration auf $(A - \lambda I)^{-1}$ anzuwenden. Hiermit erhält man das folgende Verfahren:

- Sei λ eine (gute) Näherung für einen Eigenwert λ_j von $A \in \mathbb{C}^{n \times n}$, ferner sei eine Näherung $y^{(0)} \in \mathbb{C}^n \setminus \{0\}$ für einen zugehörigen Eigenvektor u_j gegeben. Es sei $(u_j)_i \neq 0$.
- $x^{(0)} := y^{(0)} / \|y^{(0)}\|$.

- Für $k = 1, 2, \dots$:

$$y^{(k)} := (A - \lambda I)^{-1} x^{(k-1)}, \quad \lambda^{(k)} := \lambda + x_l^{(k-1)} / y_l^{(k)}, \quad x^{(k)} := y^{(k)} / \|y^{(k)}\|$$

Man beachte, daß man zur Berechnung von $y^{(k)} := (A - \lambda I)^{-1} x^{(k-1)}$ ein lineares Gleichungssystem mit der Koeffizientenmatrix $A - \lambda I$ zu lösen hat. Daher wird man zu Beginn zunächst mit Hilfe des Gaußschen Eliminationsverfahrens mit Spaltenpivot-suche eine LR -Zerlegung von $A - \lambda I$ berechnen, also eine Permutationsmatrix P , eine untere Dreiecksmatrix L mit Einsen in der Diagonale sowie eine obere Dreiecksmatrix R mit $P(A - \lambda I) = LR$. Dann erhält man $y^{(k)}$ durch Vorwärts- und Rückwärtsein-setzen aus $LRy^{(k)} = Px^{(k-1)}$. Besonders einfach erhält man $y^{(k)}$, wenn A und damit auch $A - \lambda I$ eine Tridiagonalmatrix ist.

Wir werden zwar keine Konvergenzaussage für das Verfahren der inversen Iteration beweisen, wollen aber doch motivieren, weshalb es sich hier um ein gutes Verfahren zur Berechnung bestimmter Eigenwerte und zugehöriger Eigenvektoren handelt. Hat der normierte Startvektor $x^{(0)}$ bezüglich der (linear unabhängigen) Eigenvektoren u_1, \dots, u_n der diagonalisierbaren Matrix $A \in \mathbb{C}^{n \times n}$ die Darstellung $x^{(0)} = \sum_{i=1}^n \alpha_i u_i$, so ist

$$y^{(1)} = (A - \lambda I)^{-1} x^{(0)} = \sum_{i=1}^n \frac{\alpha_i}{\lambda_i - \lambda} u_i.$$

Nach unserer Annahme liegt λ sehr viel näher bei λ_j als bei den übrigen Eigenwerten λ_i , $i \neq j$. Falls also nicht α_j verglichen mit den α_i , $i \neq j$, sehr klein ist (oder gar verschwindet), wird der j -te Summand die übrigen Terme in der Darstellung von $y^{(1)}$ dominieren, so daß $y^{(1)}$ (bzw. nach der Normierung $x^{(1)}$) eine gute Approximation an einen zu λ_j gehörenden Eigenvektor sein wird. Eine weitere Motivation erhält man, wenn man die „Verwandtschaft“ der inversen Iteration mit dem Newton-Verfahren zur Lösung nichtlinearer Gleichungssysteme (siehe Abschnitt 2.3) beachtet. Hierzu betrachte man bei gegebenem Vektor $c \in \mathbb{C}^n \setminus \{0\}$ (z. B. sei $c = e_l$ der l -te Einheitsvektor) die Nullstellenaufgabe

$$(*) \quad f(u, \lambda) := \begin{pmatrix} Au - \lambda u \\ c^T u - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Ist (u, λ) eine Lösung von $(*)$, so ist u ein Eigenvektor zum Eigenwert λ . Als Funktionalmatrix von f an der Stelle (u, λ) erhält man

$$f'(u, \lambda) = \left(\begin{array}{c|c} A - \lambda I & -u \\ \hline c^T & 0 \end{array} \right).$$

Das Newton-Verfahren zur Lösung von $(*)$ lautet

$$f'(u^{(k)}, \lambda^{(k)}) \begin{pmatrix} u^{(k+1)} - u^{(k)} \\ \lambda^{(k+1)} - \lambda^{(k)} \end{pmatrix} = -f(u^{(k)}, \lambda^{(k)})$$

bzw. (nach Einsetzen und leichter Rechnung)

$$u^{(k+1)} = \frac{1}{c^T (A - \lambda^{(k)} I)^{-1} u^{(k)}} (A - \lambda^{(k)} I)^{-1} u^{(k)}, \quad \lambda^{(k+1)} = \lambda^{(k)} + \frac{1}{c^T (A - \lambda^{(k)} I)^{-1} u^{(k)}},$$

wobei angenommen wird, daß $A - \lambda^{(k)}I$ nichtsingulär und $c^T(A - \lambda^{(k)})^{-1}u^{(k)} \neq 0$ ist. Ein Nachteil dieses Verfahrens besteht u. a. darin, daß für $k = 1, 2, \dots$ ein lineares Gleichungssystem mit der sich in jedem Schritt verändernden Koeffizientenmatrix $A - \lambda^{(k)}I$ zu lösen ist. Betrachtet man statt des Newton-Verfahrens zur Lösung von (*) mit einer Näherung λ für den zu berechnenden Eigenwert das Verfahren

$$f'(u^{(k)}, \lambda) \begin{pmatrix} u^{(k+1)} - u^{(k)} \\ \lambda^{(k+1)} - \lambda^{(k)} \end{pmatrix} = -f(u^{(k)}, \lambda^{(k)})$$

(sozusagen ein Mittelweg zwischen dem eigentlichen und dem vereinfachten Newton-Verfahren) bzw.

$$u^{(k+1)} = \frac{1}{c^T(A - \lambda I)^{-1}u^{(k)}}(A - \lambda I)^{-1}u^{(k)}, \quad \lambda^{(k+1)} = \lambda + \frac{1}{c^T(A - \lambda I)^{-1}u^{(k)}},$$

so erkennt man, daß hier bis auf die Normierung dieselbe Folge von Vektoren wie bei der inversen Iteration erzeugt wird.

5.2.5 Die Konvergenz des einfachen QR-Verfahrens

Das *einfache QR-Verfahren* zur Bestimmung der Eigenwerte einer Matrix $A \in \mathbb{C}^{n \times n}$ (in diesem theoretischen Unterabschnitt, dessen Darstellung entscheidend von D. S. WATKINS (1982) beeinflußt wurde, gehen wir von einer *komplexen* Matrix aus) lautet folgendermaßen:

- Setze $A_1 := A$.
- Für $k = 1, 2, \dots$:

Bestimme eine QR-Zerlegung von A_k , also eine unitäre Matrix $Q_k \in \mathbb{C}^{n \times n}$ und eine obere Dreiecksmatrix $R_k \in \mathbb{C}^{n \times n}$ mit $A_k = Q_k R_k$.

Berechne $A_{k+1} := R_k Q_k$.

Es wird also nicht notwendig ein Reduktionsschritt zu Beginn des Verfahrens gemacht (dieser dient hauptsächlich dazu, die Kosten eines Schrittes des QR-Verfahrens zu senken), ferner wird von der Einführung von Shift-Parametern (diese sind für eine Konvergenzbeschleunigung wesentlich) (noch) abgesehen. Ziel dieses Unterabschnittes ist es, hinreichende Bedingungen dafür anzugeben, daß die durch das einfache QR-Verfahren erzeugte Folge $\{A_k\}$ „im wesentlichen“ gegen eine zur Ausgangsmatrix A unitär ähnliche obere Dreiecksmatrix konvergiert. In den Diagonalelementen dieser oberen Dreiecksmatrix findet man dann die gesuchten Eigenwerte.

Bevor wir auf die Konvergenzanalyse für das einfache QR-Verfahren eingehen, sollten wir die Existenz und „Eindeutigkeit“ einer QR-Zerlegung auch für komplexe Matrizen $A \in \mathbb{C}^{n \times n}$ klären.

Satz 2.5 Sei $A \in \mathbb{C}^{n \times n}$ nichtsingulär. Dann besitzt A eine QR-Zerlegung, d. h. es existieren eine unitäre Matrix $Q \in \mathbb{C}^{n \times n}$ und eine obere Dreiecksmatrix $R \in \mathbb{C}^{n \times n}$

mit $A = QR$. Ist $A = \tilde{Q}\tilde{R}$ eine weitere QR-Zerlegung von A , so existiert eine unitäre Diagonalmatrix $D \in \mathbb{C}^{n \times n}$ mit $\tilde{Q} = QD$ und $\tilde{R} = D^H R$, so daß $\tilde{R}\tilde{Q} = D^H R Q D$ durch eine Ähnlichkeitstransformation mit der unitären Diagonalmatrix D aus RQ hervorgeht.

Beweis: Da man die linear unabhängigen Spalten von A sukzessive nach E. Schmidt orthonormieren kann, existiert eine QR-Zerlegung von A . Da A nichtsingulär ist, folgt aus $A = QR = \tilde{Q}\tilde{R}$, daß $Q^H\tilde{Q} = R\tilde{R}^{-1}$. Daher ist $Q^H\tilde{Q}$ eine unitäre obere Dreiecksmatrix und damit, wie man leicht einsieht, sogar eine unitäre Diagonalmatrix. Hieraus folgen die restlichen Behauptungen. \square

Wegen Satz 2.5 ist A_{k+1} „im wesentlichen“, d. h. bis auf eine Ähnlichkeitstransformation mit einer unitären Diagonalmatrix $D = \text{diag}(\exp(i\phi_1), \dots, \exp(i\phi_n))$, eindeutig durch A_k bestimmt. Das bedeutet insbesondere, daß die Diagonalelemente von A_{k+1} eindeutig durch A_k festgelegt sind.

In dem folgenden Lemma werden einige Eigenschaften der durch das einfache QR-Verfahren erzeugten Folgen von Matrizen gesammelt.

Lemma 2.6 Sei $A \in \mathbb{C}^{n \times n}$ nichtsingulär. Durch das obige einfache QR-Verfahren seien Folgen $\{Q_k\}$ (unitärer Matrizen), $\{R_k\}$ (oberer Dreiecksmatrizen) und $\{A_k\}$ erzeugt worden. Man definiere

$$\hat{Q}_0 := I, \quad \hat{Q}_k := Q_1 \cdots Q_k = (\hat{q}_1^{(k)} \ \cdots \ \hat{q}_n^{(k)}), \quad \hat{R}_k := R_k \cdots R_1.$$

Dann gilt für $k = 0, 1, \dots$:

1. $\hat{Q}_k^H A \hat{Q}_k = A_{k+1}$, d. h. A_{k+1} ist unitär ähnlich zu A .
2. $A \hat{Q}_k = \hat{Q}_{k+1} \hat{R}_{k+1}$.
3. Mit $S_m := \text{span}\{e_1, \dots, e_m\}$ (hierbei bezeichne e_j den j -ten Einheitsvektor im \mathbb{C}^n) ist

$$A^{k+1}(S_m) = \text{span}\{\hat{q}_1^{(k+1)}, \dots, \hat{q}_m^{(k+1)}\}, \quad m = 1, \dots, n,$$

d. h. die ersten m Spalten von \hat{Q}_{k+1} bilden eine Orthonormalbasis von $A^{k+1}(S_m)$.

4. $A^{k+1} = \hat{Q}_{k+1} \hat{R}_{k+1}$, d. h. durch \hat{Q}_{k+1} und \hat{R}_{k+1} ist eine QR-Zerlegung von A^{k+1} gegeben.

Beweis: Der Beweis erfolgt durch vollständige Induktion nach k . Für $k = 0$ sind die Aussagen offenbar richtig. Dies gelte auch für $k - 1$.

Unter Benutzung der Induktionsannahme ist

$$\hat{Q}_k^H A \hat{Q}_k = Q_k^H \hat{Q}_{k-1}^H A \hat{Q}_{k-1} Q_k = Q_k^H A_k Q_k = Q_k^H Q_k R_k Q_k = R_k Q_k = A_{k+1}.$$

Hieraus folgt

$$A \hat{Q}_k = \hat{Q}_k A_{k+1} = \hat{Q}_k Q_{k+1} R_{k+1} = \hat{Q}_{k+1} R_{k+1}.$$

Nach Induktionsannahme ist $A^k(S_m) = \text{span}\{\hat{q}_1^{(k)}, \dots, \hat{q}_m^{(k)}\}$, $m = 1, \dots, n$. Da A als nichtsingulär vorausgesetzt wurde, ist

$$A^{k+1}(S_m) = \text{span}\{A \hat{q}_1^{(k)}, \dots, A \hat{q}_m^{(k)}\}, \quad m = 1, \dots, n.$$

Wegen der gerade eben bewiesenen Beziehung $A\hat{Q}_k = \hat{Q}_{k+1}R_{k+1}$ entsteht die unitäre Matrix \hat{Q}_{k+1} durch sukzessives Orthonormieren der Spalten $A\hat{q}_1^{(k)}, \dots, A\hat{q}_n^{(k)}$ von $A\hat{Q}_k$:

$$A\hat{q}_j^{(k)} = \sum_{i=1}^j r_{ij}^{(k+1)} \hat{q}_i^{(k+1)}, \quad j = 1, \dots, m \leq n.$$

Daher ist $A^{k+1}(\mathcal{S}_m) = \text{span}\{\hat{q}_1^{(k+1)}, \dots, \hat{q}_m^{(k+1)}\}$. Schließlich ist

$$A^{k+1} = AA^k = A\hat{Q}_k\hat{R}_k = \hat{Q}_{k+1}R_{k+1}\hat{R}_k = \hat{Q}_{k+1}\hat{R}_{k+1}.$$

Damit ist das Lemma bewiesen. \square

Nun erinnern wir uns an die Vektoriteration nach v. Mises (siehe Unterabschnitt 5.2.3). Bei dieser hatten wir angenommen, daß $A \in \mathbb{C}^{n \times n}$ diagonalisierbar ist und einen dominanten Eigenwert besitzt. Es wurde also vorausgesetzt, daß ein vollständiges System $\{u_1, \dots, u_n\}$ von Eigenvektoren zu A existiert und daß für die Eigenwerte $\lambda_1, \dots, \lambda_n$ von A gilt $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$. Die Vektoriteration wird gestartet mit einem $y^{(0)} \in \mathbb{C}^n$ mit

$$\text{span}\{y^{(0)}\} \cap \text{span}\{u_2, \dots, u_n\} = \{0\}$$

(dies bedeutet nämlich gerade, daß $y^{(0)}$ eine nicht verschwindende Komponente bezüglich u_1 besitzt). Mit einer Norm $\|\cdot\|$ auf \mathbb{C}^n wurde $x^{(0)} := y^{(0)} / \|y^{(0)}\|$ gesetzt. Anschließend wurden Folgen $\{y^{(k)}\}$ sowie $\{x^{(k)}\}$ mit Hilfe von

$$y^{(k)} := Ax^{(k-1)}, \quad x^{(k)} := \frac{y^{(k)}}{\|y^{(k)}\|}, \quad k = 1, 2, \dots,$$

erzeugt. Dann hatten wir uns überlegt: Ist $\lambda_1 = |\lambda_1| e^{i\phi}$, so ist

$$\lim_{k \rightarrow \infty} e^{-ik\phi} x^k = \frac{\alpha_1}{|\alpha_1| \|u_1\|} u_1.$$

Mit $\mathcal{S} := \text{span}\{x^{(0)}\}$ (und $\|x^{(0)}\| = 1$) wird also bei der Vektoriteration zunächst $A(\mathcal{S}) = \text{span}\{Ax^{(0)}\} = \text{span}\{y^{(1)}\}$ gebildet und dann der Basisvektor $y^{(1)}$ normiert: $A(\mathcal{S}) = \text{span}\{x^{(1)}\}$. In dieser Weise wird fortgefahrene und eine Folge $\{x^{(k)}\}$ mit $A^k(\mathcal{S}) = \text{span}\{x^{(k)}\}$ und $\|x^{(k)}\| = 1$ gewonnen. Ist $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm und u_1 ein durch $\|u_1\|_2 = 1$ normierter Eigenvektor zu λ_1 , so konvergiert der Winkel $\angle(u_1, x^{(k)})$ zwischen u_1 und $x^{(k)}$ wegen

$$\lim_{k \rightarrow \infty} \sin \angle(u_1, x^{(k)}) = \lim_{k \rightarrow \infty} \sqrt{1 - |u_1^H x^{(k)}|^2} = 0$$

gegen Null, was eine Präzisierung der in 5.2.3 gemachten Aussage ist, daß $\{x^{(k)}\}$ der „Richtung nach“ gegen einen zu λ_1 gehörenden normierten Eigenvektor konvergiert.

Man erkennt nun die Analogie zwischen der Vektoriteration und dem QR-Verfahren. Statt mit dem einen eindimensionalen linearen Teilraum $\mathcal{S} = \text{span}\{x^{(0)}\}$, wird mit den linearen Teilaräumen $\mathcal{S}_m = \text{span}\{e_1, \dots, e_m\}$ für $m = 1, \dots, n$ gestartet, $A(\mathcal{S}_m)$ gebildet und anschließend die Basisvektoren orthonormiert. In dieser

Weise wird fortgefahrene (Bilder der Basisvektoren berechnen, dann diese orthonormieren) und hiermit Folgen $\{A^k(\mathcal{S}_m)\}$ von m -dimensionalen linearen Teilräumen des \mathbb{C}^n gewonnen, von denen man sich erhofft, daß sie für $m = 1, \dots, n$ gegen den m -dimensionalen linearen Teilraum $\mathcal{U}_m := \text{span}\{u_1, \dots, u_m\}$ „konvergieren“. Um das präzisieren zu können, wird ein Abstandsbegriff auf der Menge linearer Teilräume des \mathbb{C}^n gleicher Dimension eingeführt.

Durch die folgende Vorschrift ordne man dem m -dimensionalen linearen Teilraum \mathcal{S} mit den Basisvektoren $s_1, \dots, s_m \in \mathbb{C}^n$ zunächst eine Matrix $S \in \mathbb{C}^{n \times m}$ und anschließend eine Matrix $P_S \in \mathbb{C}^{n \times n}$ zu:

$$\mathcal{S} := \text{span}\{s_1, \dots, s_m\} \longrightarrow S := \begin{pmatrix} s_1 & \cdots & s_m \end{pmatrix} \longrightarrow P_S := S(S^H S)^{-1} S^H.$$

Bei gegebenem $x \in \mathbb{C}^n$ ist $P_S x \in \mathcal{S}$ die eindeutige orthogonale Projektion von x auf \mathcal{S} , insbesondere ist P_S von der gewählten Basis von \mathcal{S} unabhängig.

Bindet nun \mathcal{S} und \mathcal{U} zwei m -dimensionale lineare Teilräume des \mathbb{C}^n , so definiere man ihren *Abstand* (siehe auch G. H. GOLUB, C. F. VAN LOAN (1989, S. 76)) durch

$$d(\mathcal{S}, \mathcal{U}) := \|P_S - P_U\|_2 = \|S(S^H S)^{-1} S^H - U(U^H U)^{-1} U^H\|_2.$$

Ganz offensichtlich ist hierdurch auf der Menge der m -dimensionalen linearen Teilräume des \mathbb{C}^n eine Metrik d erklärt. Denn:

1. Natürlich ist $d(\mathcal{S}, \mathcal{U}) \geq 0$. Zum Nachweis der *Definitheit* nehmen wir an, es sei $d(\mathcal{S}, \mathcal{U}) = 0$. Dann ist $P_S = P_U$ wegen der Definitheit der Matrixnorm $\|\cdot\|_2$. Ist $x = Sy \in \mathcal{S}$, so ist folglich $x = P_S x = P_U x \in \mathcal{U}$, so daß $\mathcal{S} \subset \mathcal{U}$. Aus Symmetriegründen ist auch $\mathcal{U} \subset \mathcal{S}$, insgesamt also $\mathcal{S} = \mathcal{U}$.
2. Die *Symmetrieforderung* $d(\mathcal{S}, \mathcal{U}) = d(\mathcal{U}, \mathcal{S})$ ist offensichtlich erfüllt.
3. Sind \mathcal{S} , \mathcal{U} und \mathcal{V} drei m -dimensionale lineare Teilräume des \mathbb{C}^n , so ist

$$d(\mathcal{S}, \mathcal{U}) = \|P_S - P_U\|_2 \leq \|P_S - P_V\|_2 + \|P_V - P_U\|_2 = d(\mathcal{S}, \mathcal{V}) + d(\mathcal{V}, \mathcal{U}).$$

Also ist auch die *Dreiecksungleichung* erfüllt.

Beispiel: Seien $\mathcal{S} = \text{span}\{s\}$ und $\mathcal{U} = \text{span}\{u\}$ mit $\|s\|_2 = \|u\|_2 = 1$ zwei eindimensionale lineare Teilräume des \mathbb{C}^n . Dann ist

$$d(\mathcal{S}, \mathcal{U}) = \|P_S - P_U\|_2 = \|ss^H - uu^H\|_2 = \rho(ss^H - uu^H).$$

Sind s und u linear abhängig, so ist $\mathcal{S} = \mathcal{U}$ und daher $d(\mathcal{S}, \mathcal{U}) = 0$. Daher nehmen wir nun an, s und u seien linear unabhängig und berechnen die Eigenwerte von $A := ss^H - uu^H$. Die hermitesche Matrix A besitzt 0 als $(n-2)$ -fachen Eigenwert mit zugehörigen Eigenvektoren aus dem $(n-2)$ -dimensionalen linearen Raum $\text{span}\{s, u\}^\perp$, dem orthogonalen Komplement von $\text{span}\{s, u\}$. Für die beiden restlichen Eigenvektoren aus $\text{span}\{s, u\}$ macht man den Ansatz $x = \alpha s + \beta u$. Dann ist $Ax = \lambda x$ gleichwertig mit $(\alpha + \beta s^H u)s - (\alpha u^H s + \beta)u = \lambda(\alpha s + \beta u)$ bzw. mit

$$\begin{pmatrix} 1 & s^H u \\ -u^H s & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Für die beiden restlichen Eigenwerte von A erhält man damit

$$\lambda_{1,2} = \pm(1 - |u^H s|^2)^{1/2}.$$

Daher ist $d(\mathcal{S}, \mathcal{U}) = (1 - |u^H s|^2)^{1/2} = \sin \angle(u, s)$, der Abstand zwischen $\mathcal{S} = \text{span}\{s\}$ und $\mathcal{U} = \text{span}\{u\}$ ist also genau der Sinus des Winkels zwischen s und u . \square

Als Verallgemeinerung eines Ergebnisses für die Vektoriteration erhalten wir

Satz 2.7 Sei $A \in \mathbb{C}^{n \times n}$ diagonalähnlich mit Eigenwerten $\lambda_1, \dots, \lambda_n$ und zugehörigen Eigenvektoren u_1, \dots, u_n . Es sei $|\lambda_1| \geq \dots \geq |\lambda_n| > 0$. Für ein $m \in \{1, \dots, n-1\}$ sei $|\lambda_m| > |\lambda_{m+1}|$. Man setze

$$\mathcal{U}_m := \text{span}\{u_1, \dots, u_m\}, \quad \mathcal{T}_m := \text{span}\{u_{m+1}, \dots, u_n\}.$$

Schließlich sei \mathcal{S}_m ein m -dimensionaler linearer Teilraum des \mathbb{C}^n mit $\mathcal{S}_m \cap \mathcal{T}_m = \{0\}$. Dann gibt es eine Konstante $C > 0$ mit

$$d(A^k(\mathcal{S}_m), \mathcal{U}_m) \leq C \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k \quad \text{für alle hinreichend großen } k,$$

d. h. die Folge $\{A^k(\mathcal{S}_m)\}$ konvergiert mit der Konvergenzrate $|\lambda_{m+1}/\lambda_m|$ gegen \mathcal{U}_m .

Beweis: In einem ersten Schritt konstruiert man $v_1^{(k)}, \dots, v_m^{(k)} \in \mathbb{C}^n$ mit

$$A^k(\mathcal{S}_m) = \text{span}\{v_1^{(k)}, \dots, v_m^{(k)}\}, \quad \|v_i^{(k)} - u_i\|_2 \leq \hat{C} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k$$

für $i = 1, \dots, m$ und $k \in \mathbb{N}$, wobei die Konstante \hat{C} von k (und i) unabhängig ist. Mit Hilfe der Definition der Metrik d zeigt man anschließend die Behauptung.

Sei $\{s_1, \dots, s_m\}$ eine Basis von \mathcal{S}_m . Wegen $\mathbb{C}^n = \mathcal{U}_m \oplus \mathcal{T}_m$ besitzt das i -te Basis-
element s_i von \mathcal{S}_m eine Darstellung

$$(*) \quad s_i = \sum_{j=1}^m \alpha_{ij} u_j + t_i \quad \text{mit} \quad t_i \in \mathcal{T}_m, \quad i = 1, \dots, m.$$

Wegen $\mathcal{S}_m \cap \mathcal{T}_m = \{0\}$ ist $(\alpha_{ij}) \in \mathbb{C}^{m \times m}$ nichtsingulär. Denn: Ist $\sum_{i=1}^m \alpha_{ij} \beta_i = 0$ für $j = 1, \dots, m$, so ist

$$\sum_{i=1}^m \beta_i s_i = \sum_{i=1}^m \beta_i \left(\sum_{j=1}^m \alpha_{ij} u_j + t_i \right) = \sum_{j=1}^m \underbrace{\left(\sum_{i=1}^m \alpha_{ij} \beta_i \right)}_{=0} u_j + \sum_{i=1}^m \beta_i t_i = \sum_{i=1}^m \beta_i t_i \in \mathcal{S}_m \cap \mathcal{T}_m$$

und folglich $\beta_1 = \dots = \beta_m = 0$. Da man notfalls $(*)$ mit $(\alpha_{ij})^{-1}$ durchmultiplizieren kann, können wir o. B. d. A. annehmen, daß $\{s_1, \dots, s_m\}$ eine Basis von \mathcal{S}_m mit

$$s_i = u_i + t_i, \quad t_i \in \mathcal{T}_m, \quad i = 1, \dots, m,$$

ist. Wegen $A^k s_i = \lambda_i^k u_i + A^k t_i$ ist durch $\{v_1^{(k)}, \dots, v_m^{(k)}\}$ mit

$$v_i^{(k)} := u_i + \frac{A^k t_i}{\lambda_i^k}, \quad i = 1, \dots, m,$$

eine Basis von $A^k(\mathcal{S}_m)$ gegeben. Das Element $t_i \in \mathcal{T}_m$ lässt sich für $i = 1, \dots, m$ eindeutig darstellen als $t_i = \sum_{j=m+1}^n \beta_{ij} u_j$. Dann ist $A^k t_i = \sum_{j=m+1}^n \beta_{ij} \lambda_j^k u_j$ und folglich

$$\|v_i^{(k)} - u_i\|_2 \leq \sum_{j=m+1}^n \left| \frac{\lambda_j}{\lambda_i} \right|^k |\beta_{ij}| \|u_j\|_2 \leq \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k \sum_{j=m+1}^n |\beta_{ij}| \|u_j\|_2 \leq \hat{C} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k$$

mit einer von k und i unabhängigen Konstanten \hat{C} .

Nun gilt es, den Abstand $d(A^k(\mathcal{S}_m), \mathcal{U}_m)$ der beiden m -dimensionalen linearen Teilräumen $A^k(\mathcal{S}_m)$ und \mathcal{U}_m abzuschätzen. Mit den $n \times m$ -Matrizen

$$V_k := \begin{pmatrix} v_1^{(k)} & \cdots & v_m^{(k)} \end{pmatrix}, \quad U := \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix}$$

ist nach Definition

$$d(A^k(\mathcal{S}_m), \mathcal{U}_m) = \|V_k(V_k^H V_k)^{-1} V_k^H - U(U^H U)^{-1} U^H\|_2 = \|V_k V_k^+ - U U^+\|_2,$$

wobei

$$V_k^+ := (V_k^H V_k)^{-1} V_k^H, \quad U^+ := (U^H U)^{-1} U^H$$

die *Pseudoinverse* von V_k bzw. U ist (siehe Unterabschnitt 1.6.3, wo wir die Pseudoinverse, allerdings unnötigerweise nur für reelle Matrizen, definiert hatten). Für ein beliebiges $y \in \mathbb{C}^m$ ist

$$\|(V_k - U)y\|_2 = \left\| \sum_{i=1}^m (v_i^{(k)} - u_i)y_i \right\|_2 \leq \sum_{i=1}^m \|v_i^{(k)} - u_i\|_2 |y_i| \leq \sqrt{m} \hat{C} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k \|y\|_2$$

und daher

$$\|V_k - U\|_2 = \sup_{y \neq 0} \frac{\|(V_k - U)y\|_2}{\|y\|_2} \leq \sqrt{m} \hat{C} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k.$$

Für alle hinreichend großen k ist $\|U^+\|_2 \|V_k - U\|_2 \leq \frac{1}{2}$, für diese k ist (wegen des verallgemeinerten Störungslemmas 6.6 in Abschnitt 1.6) $\|V_k^+\|_2 \leq 2 \|U^+\|_2$. Wegen

$$\begin{aligned} V_k V_k^+ - U U^+ &= (V_k - U)V_k^+ + U(V_k^+ - U^+) \\ &= (V_k - U)V_k^+ + U[V_k^+(I - U U^+) - V_k^+(V_k - U)U^+] \end{aligned}$$

sowie

$$\|V_k^+(I - U U^+)\|_2 \leq \|V_k^+\|_2^2 \|V_k - U\|_2$$

(siehe den Beweis zu Satz 6.8 in Abschnitt 1.6) existiert eine Konstante $c > 0$ derart, daß für alle hinreichend großen k gilt

$$d(A^k(\mathcal{S}_m), \mathcal{U}_m) = \|V_k V_k^+ - U U^+\|_2 \leq c \|V_k - U\|_2 \leq \underbrace{c \sqrt{m} \hat{C}}_{=: C} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k,$$

das war zu zeigen². □

Der letzte Satz wird in naheliegender Weise angewandt, um eine Konvergenzaussage für das einfache QR-Verfahren zu beweisen.

²Es ist nicht schwierig, für den letzten Teil einen direkten Beweis zu finden, der den Begriff der Pseudoinversen vermeidet und keine Hilfsmittel aus Abschnitt 1.6 benutzt.

Satz 2.8 Sei $A \in \mathbb{C}^{n \times n}$ diagonalähnlich mit Eigenwerten $\lambda_1, \dots, \lambda_n$ und zugehörigen Eigenvektoren u_1, \dots, u_n . Es sei

$$(*) \quad |\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0.$$

Für $m = 1, \dots, n-1$ setze man

$$\mathcal{S}_m := \text{span } \{e_1, \dots, e_m\}, \quad \mathcal{U}_m := \text{span } \{u_1, \dots, u_m\}, \quad \mathcal{T}_m := \text{span } \{u_{m+1}, \dots, u_n\}.$$

Es wird

$$(**) \quad \mathcal{S}_m \cap \mathcal{T}_m = \{0\}, \quad m = 1, \dots, n-1,$$

vorausgesetzt. Dann gilt für die durch das obige einfache QR-Verfahren erzeugte Folge $\{A_k\} = \{(\hat{a}_{ij}^{(k)})\}$:

$$\lim_{k \rightarrow \infty} \hat{a}_{ij}^{(k)} = 0 \quad (1 \leq j < i \leq n), \quad \lim_{k \rightarrow \infty} \hat{a}_{ii}^{(k)} = \lambda_i \quad (i = 1, \dots, n).$$

Beweis: Wegen Satz 2.7 konvergiert $\{A^k(\mathcal{S}_m)\}$ für $m = 1, \dots, n-1$ (mit der Konvergenzrate $|\lambda_{m+1}/\lambda_m|$) gegen \mathcal{U}_m . Nach Lemma 2.6 ist

$$A^k(\mathcal{S}_m) = \text{span } \{\hat{q}_1^{(k)}, \dots, \hat{q}_m^{(k)}\} \quad \text{mit} \quad \hat{Q}_k := Q_1 \cdots Q_k = (\hat{q}_1^{(k)} \ \cdots \ \hat{q}_n^{(k)}).$$

Man kennt also eine Orthonormalbasis von $A^k(\mathcal{S}_m)$. Um den Abstand $d(A^k(\mathcal{S}_m), \mathcal{U}_m)$ besser ausrechnen zu können, liegt es nahe, auch die Eigenvektoren u_1, \dots, u_n sukzessive von links nach rechts zu orthonormieren und hierdurch v_1, \dots, v_n zu erhalten. Dann ist $\mathcal{U}_m = \text{span } \{v_1, \dots, v_m\}$, aus der Definition des Abstandes m -dimensionaler Teilräume folgt

$$\lim_{k \rightarrow \infty} d(A^k(\mathcal{S}_m), \mathcal{U}_m) = \lim_{k \rightarrow \infty} \left\| \sum_{j=1}^m [\hat{q}_j^{(k)}(\hat{q}_j^{(k)})^H - v_j v_j^H] \right\|_2 = 0, \quad m = 1, \dots, n-1.$$

Mit der unitären Matrix $V := (v_1 \ \cdots \ v_n)$ ist ferner

$$\sum_{j=1}^n [\hat{q}_j^{(k)}(\hat{q}_j^{(k)})^H - v_j v_j^H] = \underbrace{\hat{Q}_k \hat{Q}_k^H}_{=I} - \underbrace{V V^H}_{=I} = 0.$$

Sukzessive erhält man daher

$$\lim_{k \rightarrow \infty} \hat{q}_j^{(k)}(\hat{q}_j^{(k)})^H = v_j v_j^H, \quad j = 1, \dots, n,$$

hieraus

$$\lim_{k \rightarrow \infty} (\hat{q}_j^{(k)})^H v_j \hat{q}_j^{(k)} = v_j, \quad \lim_{k \rightarrow \infty} |(\hat{q}_j^{(k)})^H v_j| = 1, \quad j = 1, \dots, n,$$

und damit schließlich

$$\lim_{k \rightarrow \infty} d_j^{(k)} \hat{q}_j^{(k)} = v_j \quad \text{mit} \quad d_j^{(k)} := \frac{(\hat{q}_j^{(k)})^H v_j}{|(\hat{q}_j^{(k)})^H v_j|}, \quad j = 1, \dots, n.$$

Mit der unitären Diagonalmatrix $D_k := \text{diag}(d_1^{(k)}, \dots, d_n^{(k)})$ sowie der unitären Matrix $P_k := \hat{Q}_k D_k$ ist folglich $\lim_{k \rightarrow \infty} P_k = V$ und daher

$$V^H A V \leftarrow P_k^H A P_k = D_k^H \hat{Q}_k^H A \hat{Q}_k D_k = D_k^H A_{k+1} D_k.$$

Nun ist $V^H A V$ eine obere Dreiecksmatrix, die λ_i als i -tes Diagonalelement besitzt. Um das einzusehen, beachten wir zunächst, daß $(V^H A V)_{ij} = v_i^H A v_j$ für $1 \leq i, j \leq n$. Da A den linearen Teilraum \mathcal{U}_j invariant läßt, ist $A v_j \in \mathcal{U}_j = \text{span}\{v_1, \dots, v_j\}$ und damit $v_i^H A v_j = 0$ für $i > j$. Folglich ist $V^H A V$ eine obere Dreiecksmatrix. Schließlich läßt sich v_i in der Form $v_i = \alpha_i u_i + w_i$ mit $w_i \in \mathcal{U}_{i-1}$ darstellen. Daher ist

$$v_i^H A v_i = v_i^H (\alpha_i \lambda_i u_i + A w_i) = \lambda_i \alpha_i v_i^H u_i = \lambda_i v_i^H v_i = \lambda_i.$$

Da gerade die Konvergenz von $\{D_k^H A_{k+1} D_k\}$ gegen $V^H A V$ gezeigt wurde, wobei $\{D_k\}$ eine Folge unitärer Diagonalmatrizen ist, ist die Behauptung bewiesen. \square

Bemerkungen: Die in Satz 2.8 gemachte Voraussetzung

$$(*) \quad |\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0$$

ist zwar sehr einschränkend (z. B. verbietet sie bei *reellem* A die Existenz (konjugiert) komplexer Eigenwerte, andererseits kann in diesem Falle wegen des reellen Schurschen Zerlegungssatzes auch nicht die Existenz einer Folge zu A orthogonal ähnlicher Matrizen erwartet werden, die gegen eine obere Dreiecksmatrix konvergiert), wegen der Analogie des QR-Verfahrens zur Vektoriteration scheint sie trotzdem natürlich zu sein.

Definiert man die (nichtsinguläre) Matrix $U := (u_1 \ \dots \ u_n) \in \mathbb{C}^{n \times n}$, so ist die zweite Voraussetzung

$$(**) \quad \text{span}\{e_1, \dots, e_m\} \cap \text{span}\{u_{m+1}, \dots, u_n\} = \{0\}, \quad m = 1, \dots, n-1,$$

genau dann erfüllt, wenn U^{-1} eine LR-Zerlegung besitzt bzw. das Gaußsche Eliminationsverfahren ohne Spaltenpivotsuche durchführbar ist. Denn es ist

$$x = \sum_{j=1}^m \alpha_j e_j = \sum_{j=m+1}^n \beta_j u_j \in \text{span}\{e_1, \dots, e_m\} \cap \text{span}\{u_{m+1}, \dots, u_n\}$$

genau dann, wenn

$$U^{-1} x = \sum_{j=1}^m \alpha_j U^{-1} e_j = \sum_{j=m+1}^n \beta_j e_j.$$

Hieraus liest man ab, daß $(**)$ genau dann erfüllt ist, wenn die Hauptabschnitts determinanten von U^{-1} sämtlich von Null verschieden sind, was bekanntlich (siehe Satz 3.2 in Abschnitt 1.3) gleichbedeutend mit der Existenz einer LR-Zerlegung von U^{-1} ist. Genau diese Voraussetzung wird in dem üblichen, von Wilkinson stammenden, Konvergenzsatz für das einfache QR-Verfahren gemacht (siehe J. H. WILKINSON (1965, S. 518)).

Wiederholt wurde betont, daß man beim *QR*-Verfahren zunächst einen Reduktionsschritt macht und die gegebene Matrix in eine ähnliche Hessenberg-Matrix überführt. Ist in dieser Hessenberg-Matrix ein Subdiagonalelement gleich Null, so zerfällt das gegebene Eigenwertproblem in zwei niederdimensionale Eigenwertaufgaben für Hessenberg-Matrizen. Daher kann, wenigstens theoretisch, angenommen werden, daß die gegebene Matrix A eine *unreduzierte* Hessenberg-Matrix ist, also sämtliche Subdiagonalelemente von Null verschieden sind. Ist $A \in \mathbb{C}^{n \times n}$ eine unreduzierte, diagonalähnliche Hessenberg-Matrix, so ist die Voraussetzung (**) in Satz 2.8 erfüllt. Denn angenommen, es existiert ein $m \in \{1, \dots, n-1\}$ und ein $x \neq 0$ mit $x \in \mathcal{S}_m \cap \mathcal{T}_m$. O. B. d. A. können wir annehmen, daß $x_m \neq 0$, so daß $x \in \mathcal{S}_m \setminus \mathcal{S}_{m-1}$. Dann ist

$$(Ax)_{m+1} = \sum_{j=1}^n a_{m+1,j} x_j = \sum_{j=1}^m a_{m+1,j} x_j = a_{m+1,m} x_m \neq 0$$

und $(Ax)_k = 0$ für $k > m+1$, so daß $Ax \in \mathcal{S}_{m+1} \setminus \mathcal{S}_m$. Insbesondere sind x und Ax linear unabhängig. Dieses Argument kann fortgesetzt werden und man erhält, daß $\{x, Ax, \dots, A^{n-m}x\}$ linear unabhängig sind. Da A den linearen Teilraum \mathcal{T}_m invariant läßt, hätte man $n-m+1$ linear unabhängige Vektoren in dem $(n-m)$ -dimensionalen linearen Teilraum \mathcal{T}_m gefunden, was ein Widerspruch ist. Diese Beobachtung findet man bei B. N. PARLETT, W. G. POOLE (1973, Lemma 7.2).

Nun soll noch eine Bemerkung zu der vorausgesetzten Nichtsingularität von A gemacht werden (siehe auch das folgende Lemma 2.9). Wir haben gerade eben gesehen, daß $A = A_1$ als eine unreduzierte Hessenberg-Matrix angenommen werden kann. Ist dies der Fall, so sind die ersten $n-1$ Spalten von A_1 linear unabhängig. Ist daher $A_1 = Q_1 R_1$ eine *QR*-Zerlegung von A_1 , so sind die ersten $n-1$ Diagonalelemente $r_{ii}^{(1)}$, $i = 1, \dots, n-1$, von R_1 von Null verschieden und A_1 ist genau dann singulär, wenn $r_{nn}^{(1)} = 0$. Ist also A_1 eine singuläre, unreduzierte Hessenberg-Matrix, so besteht die n -te Zeile in R_1 nur aus Nullen. Im nächsten *QR*-Schritt ist $A_2 = R_1 Q_1$ eine zu A_1 unitär ähnliche Hessenberg-Matrix, deren letzte Zeile eine Nullzeile ist. Man streiche nun in A_2 die letzte Zeile sowie die letzte Spalte und fahre mit dem Verfahren für die so gewonnene Matrix fort. \square

5.2.6 Das *QR*-Verfahren mit Shifts

Die Konvergenzgeschwindigkeit des einfachen *QR*-Verfahrens ist i. allg. für praktische Zwecke zu schlecht. Zur Konvergenzbeschleunigung werden (explizit oder implizit) Shift-Parameter eingeführt. Wie schon mehrfach angegeben, sieht das *QR*-Verfahren zur Berechnung der Eigenwerte einer Matrix $A \in \mathbb{R}^{n \times n}$ (wir wollen uns hier wieder auf den reellen Fall beschränken) folgendermaßen aus:

- Transformiere $A \in \mathbb{R}^{n \times n}$ auf (orthogonal) ähnliche Hessenberg-Form A_1 . Zumindestens für die Berechnung der Eigenwerte kann o. B. d. A. angenommen werden, daß A_1 eine *unreduzierte* Hessenberg-Matrix ist, d. h. daß alle Subdiagonalelemente von A_1 von Null verschieden sind.
- Für $k = 1, 2, \dots$

Bestimme Shift-Parameter $\sigma_k \in \mathbb{R}$.

Bestimme eine QR-Zerlegung $A_k - \sigma_k I = Q_k R_k$ und berechne anschließend $A_{k+1} := R_k Q_k + \sigma_k I$.

Es kommt jetzt darauf an, mögliche Shift-Strategien zu spezifizieren. Die einfachste Strategie besteht darin, $\sigma_k := a_{nn}^{(k)}$ zu setzen. Eine Motivation hierfür ist durch das folgende Lemma gegeben.

Lemma 2.9 Sei $A \in \mathbb{R}^{n \times n}$ eine unreduzierte Hessenberg-Matrix und $\lambda \in \mathbb{R}$ ein Eigenwert von A . Ist $A_+ := RQ + \lambda I$, wobei $A - \lambda I = QR$ eine QR-Zerlegung von $A - \lambda I$ ist, so ist $(A_+)_{nj} = 0$ für $j = 1, \dots, n-1$ und $(A_+)_{nn} = \lambda$.

Beweis: Da A eine unreduzierte Hessenberg-Matrix ist, sind die ersten $n-1$ Spalten von $A - \lambda I$ linear unabhängig. In der QR-Zerlegung $QR = A - \lambda I$ ist daher $r_{ii} \neq 0$ für $i = 1, \dots, n-1$. Da λ ein Eigenwert von A ist, ist $A - \lambda I$ und damit auch R singulär, also notwendig $r_{nn} = 0$. Daher ist die letzte Zeile von RQ eine Nullzeile, woraus die Behauptung folgt. \square

Das letzte Lemma zeigt, daß das QR-Verfahren mit einem exakten Eigenwert als Shift-Parameter in einem Schritt zu einer Reduktion (Deflation) der Eigenwertaufgabe führt. Da die Kenntnis eines exakten Eigenwertes eine irreale Annahme ist, nehmen wir nun an, der untere 2×2 -Block der (unreduzierten) Hessenberg-Matrix

A habe die Form $\begin{pmatrix} * & * \\ \epsilon & a_{nn} \end{pmatrix}$. Ist ϵ „klein“, so wird man a_{nn} als Näherung für einen

Eigenwert ansehen, so daß es nahe liegt, einen Schritt des QR-Verfahrens mit $\sigma = a_{nn}$ als Shift-Parameter durchzuführen (ähnlich dem Vorgehen bei der inversen Iteration nach Wielandt). Wir wollen uns überlegen, wie sich das Element ϵ in der Position $(n, n-1)$ hierbei verändert. Durch Multiplikation von links mit den ersten $n-2$ Givens-Rotationen $G_{12}, \dots, G_{n-2,n-1}$ erhält man aus $A - a_{nn}I$ eine Matrix der Form

$$G_{n-2,n-1} \cdots G_{12}(A - a_{nn}I) = \left(\begin{array}{c|c} R_{n-2} & * \\ \hline 0 & \begin{matrix} a & b \\ \epsilon & 0 \end{matrix} \end{array} \right)$$

mit einer oberen Dreiecksmatrix $R_{n-2} \in \mathbb{R}^{(n-2) \times (n-2)}$. Die letzte Zeile von $A - a_{nn}I$ hat sich hierbei noch nicht verändert. Wir nehmen an, es sei $|\epsilon| \leq |a|$. Die letzte Givens-Rotation $G_{n-1,n}$ ist daher durch die Parameter

$$c_{n-1} = \frac{|a|}{\sqrt{a^2 + \epsilon^2}}, \quad s_{n-1} = \frac{\text{sign}(a)\epsilon}{\sqrt{a^2 + \epsilon^2}}$$

gegeben. Daher ist $R = G_{n-1,n} \cdots G_{12}(A - a_{nn}I)$ eine obere Dreiecksmatrix, deren unterer 2×2 -Block

$$\begin{pmatrix} r_{n-1,n-1} & r_{n-1,n} \\ 0 & r_{nn} \end{pmatrix} = \frac{1}{\sqrt{a^2 + \epsilon^2}} \begin{pmatrix} a|a| & |a|b \\ 0 & -\text{sign}(a)\epsilon b \end{pmatrix}$$

ist. Bei der anschließenden Multiplikation von rechts mit $G_{12}^T, \dots, G_{n-2,n-1}^T$ verändert sich die letzte Spalte von R nicht, ferner bleibt die Null in der Position $(n, n-1)$ erhalten. Also ist

$$RG_{12}^T \cdots G_{n-2,n-1}^T = \left(\begin{array}{c|cc} H_{n-2} & * \\ \hline 0 & * & * \\ & 0 & r_{n-1,n} \\ & & 0 & r_{nn} \end{array} \right)$$

mit einer Hessenberg-Matrix $H_{n-2} \in \mathbb{R}^{(n-2) \times (n-2)}$. Nach der abschließenden Multiplikation mit $G_{n-1,n}^T$ von rechts erhält man wegen

$$\begin{pmatrix} 0 & r_{nn} \end{pmatrix} \begin{pmatrix} c_{n-1} & -s_{n-1} \\ s_{n-1} & c_{n-1} \end{pmatrix} = -\frac{1}{a^2 + \epsilon^2} \begin{pmatrix} \epsilon^2 b & \epsilon a b \end{pmatrix}$$

in

$$A_+ := G_{n-1,n} \cdots G_{12}(A - a_{nn}I)G_{12}^T \cdots G_{n-1,n}^T + a_{nn}I$$

als neues Element in der $(n, n-1)$ -Position gerade

$$(A_+)_{n,n-1} = -\frac{\epsilon^2 b}{a^2 + \epsilon^2}.$$

Daher kann man sagen: Ist in der Hessenberg-Matrix A das Subdiagonalelement ϵ in der $(n-1)$ -ten Spalte klein, so ist das entsprechende Element nach einem Schritt des QR-Verfahrens mit dem Shift-Parameter $\sigma = a_{nn}$ von der Größenordnung ϵ^2 , also wesentlich kleiner (siehe G. W. STEWART (1973, S. 366)). Bei einer Fortsetzung dieses Verfahrens mit dem Shift-Parameter $\sigma_k := a_{nn}^{(k)}$ wird man erwarten, daß die Folge $\{a_{n,n-1}^{(k)}\}$ schnell gegen Null konvergiert. Man kann $a_{n-1,n}^{(k)}$ „zu Null erklären“ (und $a_{nn}^{(k)}$ als Eigenwert der Ausgangsmatrix A akzeptieren), wenn

$$|a_{n-1,n}^{(k)}| \leq tol(|a_{n-1,n-1}^{(k)}| + |a_{nn}^{(k)}|)$$

mit einer kleinen Toleranz tol (siehe G. W. STEWART (1973, S. 363) und G. H. GOLUB, C. F. VAN LOAN (1989, S. 373)). Streicht man anschließend in A_k die letzte Zeile und letzte Spalte, so hat man das Eigenwertproblem auf eine $(n-1)$ -dimensionale Aufgabe reduziert.

Die einfache Shift-Strategie $\sigma_k := a_{nn}^{(k)}$ verliert ihre Berechtigung, wenn das Element $a_{n,n-1}^{(k)}$ nicht klein ist verglichen mit $a_{n-1,n-1}^{(k)}$ und $a_{nn}^{(k)}$. Daher wird i. allg. vorgezogen, die beiden Eigenwerte σ_k und τ_k des unteren 2×2 -Blocks

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix}$$

zu berechnen, und anschließend einen sogenannten *QR-Doppelschritt* zu machen. Hierbei wird A_{k+2} aus A_k durch den folgenden Prozeß berechnet:

$$A_k - \sigma_k I = Q_k R_k, \quad A_{k+1} := R_k Q_k + \sigma_k I,$$

d. h. man mache einen QR-Schritt mit σ_k als Shift, anschließend einen Schritt des QR-Verfahrens mit τ_k als Shift-Parameter:

$$A_{k+1} - \tau_k I = Q_{k+1} R_{k+1}, \quad A_{k+2} := R_{k+1} Q_{k+1} + \tau_k I.$$

Nun können σ_k und τ_k (konjugiert) komplex sein, so daß auch die unitären Matrizen Q_k und Q_{k+1} sowie die oberen Dreiecksmatrizen R_k und R_{k+1} i. allg. komplex sein werden, obwohl A_k reell ist. Wir wollen uns überlegen, daß man A_{k+2} aus A_k durch eine reelle Rechnung erhalten kann. Denn

$$\begin{aligned} Q_k Q_{k+1} R_{k+1} R_k &= Q_k (A_{k+1} - \tau_k I) R_k \\ &= Q_k (A_{k+1} - \tau_k I) Q_k^H (A_k - \sigma_k I) \\ &= Q_k (R_k Q_k + \sigma_k I + \tau_k I) Q_k^H (A_k - \sigma_k I) \\ &= (A_k - \tau_k I)(A_k - \sigma_k I) \\ &= A_k^2 - (\sigma_k + \tau_k) A_k + \sigma_k \tau_k I \\ &=: \tilde{A}_k \end{aligned}$$

ist eine *reelle* Matrix, da

$$\sigma_k + \tau_k = a_{n-1,n-1}^{(k)} + a_{nn}^{(k)}, \quad \sigma_k \tau_k = a_{n-1,n-1}^{(k)} a_{nn}^{(k)} - a_{n-1,n}^{(k)} a_{n,n-1}^{(k)}$$

reell sind. Eine reelle Matrix besitzt eine reelle QR-Zerlegung, so daß die (komplex) unitären Matrizen Q_k und Q_{k+1} so gewählt werden können, daß $Q_k Q_{k+1}$ (reell) orthogonal ist. Wegen

$$A_{k+2} = Q_{k+1}^H A_{k+1} Q_{k+1} = (Q_k Q_{k+1})^H A_k (Q_k Q_{k+1})$$

muß es daher möglich sein, A_{k+2} aus A_k auch bei (konjugiert) komplexen Shift-Parametern σ_k und τ_k durch eine reelle, orthogonale Ähnlichkeitstransformation zu erhalten. Eine naive Vorgehensweise mit einem Aufwand an Multiplikationen, der im wesentlichen proportional zu n^3 ist, würde darin bestehen, die Matrix \tilde{A}_k zu bilden (was alleine schon im wesentlichen einen zu n^3 proportionalen Aufwand erfordert), hiervon eine (reelle) QR-Zerlegung $\tilde{A}_k = \tilde{Q}_k \tilde{R}_k$ zu berechnen und $A_{k+2} := \tilde{Q}_k^T A_k \tilde{Q}_k$ zu setzen. Grundlage für ein wesentlich effizienteres Verfahren zur Durchführung dieses sogenannten *QR-Doppelschrittes* ist der folgende Satz.

Satz 2.10 Seien $Q = (q_1 \ \dots \ q_n)$ und $V = (v_1 \ \dots \ v_n)$ orthogonale Matrizen, die eine gegebene Matrix $A \in \mathbb{R}^{n \times n}$ jeweils in eine Hessenberg-Matrix $H := Q^T A Q$ bzw. $G := V^T A V$ transformieren, wobei G unreduziert sei. Ist dann $q_1 = \pm v_1$, stimmen die ersten Spalten von Q und V also (eventuell bis auf einen Faktor -1) überein, so ist auch H eine unreduzierte Hessenberg-Matrix. Ferner sind dann Q und V sowie H und G im wesentlichen gleich, d. h. $D := V^T Q$ ist eine (orthogonale) Diagonalmatrix (besitzt also nur $+1$ oder -1 als Diagonalelemente) und $H = D^T G D$.

Beweis: Man definiere die orthogonale Matrix $D := V^T Q = (d_1 \ \dots \ d_n)$. Da $q_1 = \pm v_1$ vorausgesetzt wurde, ist $d_1 = \pm e_1$ (eventuell bis auf das Vorzeichen) der erste Einheitsvektor. Angenommen, es sei $d_i = \pm e_i$ für $i = 1, \dots, k$. Aus

$$G D = V^T A V V^T Q = V^T A Q = V^T Q Q^T A Q = V^T Q H = D H$$

erhält man beim Vergleich der k -ten Spalte

$$\pm Ge_k = Gd_k = (GD)_k = (DH)_k = h_{k+1,k}d_{k+1} + \sum_{i=1}^k h_{ik}d_i.$$

Multipliziert man diese Gleichung von links mit $d_i^T = \pm e_i^T$, so erhält man $\pm g_{ik} = h_{ik}$ für $i = 1, \dots, k$ und anschließend $\pm g_{k+1,k}e_{k+1} = h_{k+1,k}d_{k+1}$. Da G unreduziert ist, ist auch $h_{k+1,k} \neq 0$ und $d_{k+1} = \pm e_{k+1}$. Daher ist auch H unreduziert und D eine (orthogonale) Diagonalmatrix. \square

Nun kommen wir zu einer effizienten Lösung der folgenden Aufgabenstellung.

- Input: Gegeben sei eine Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$. Seien σ und τ die beiden Eigenwerte des unteren 2×2 -Blocks von A , also von

$$\begin{pmatrix} a_{n-1,n-1} & a_{n-1,n} \\ a_{n,n-1} & a_{nn} \end{pmatrix}.$$

Ferner sei

$$\tilde{A} := (A - \sigma I)(A - \tau I) = A^2 - \underbrace{(a_{n-1,n-1} + a_{nn})}_{=\sigma+\tau} A + \underbrace{(a_{n-1,n-1}a_{nn} - a_{n-1,n}a_{n,n-1})}_{=\sigma\tau} I$$

(dies ist nur eine *Bezeichnung*, es soll *nicht* suggeriert werden, daß \tilde{A} berechnet werden soll).

- Output: Die Matrix A wird mit einer Hessenberg-Matrix überschrieben, die „im wesentlichen“ (im Sinne von Satz 2.10, also bis auf eine Ähnlichkeitstransformation mit einer orthogonalen Diagonalmatrix) mit der Hessenberg-Matrix $A_+ := Q^T A Q$ übereinstimmt. Hierbei ist Q der orthogonale Anteil einer QR -Zerlegung von \tilde{A} , also $\tilde{A} = QR$ mit einer oberen Dreiecksmatrix R .

Man beachte, daß in der ersten Spalte von $\tilde{A} = (A - \sigma I)(A - \tau I)$ nur die ersten drei Elemente von Null verschieden sind. Diese sind gegeben durch

$$\begin{aligned} \tilde{a}_{11} &= a_{11}^2 - (\sigma + \tau)a_{11} + \sigma\tau + a_{12}a_{21} \\ \tilde{a}_{21} &= a_{21}[a_{11} + a_{22} - (\sigma + \tau)] \\ \tilde{a}_{31} &= a_{21}a_{32}. \end{aligned}$$

Damit ist die erste Spalte des orthogonalen Anteils Q in einer QR -Zerlegung von \tilde{A} bekannt, denn diese stimmt bis auf eine Normierung mit der ersten Spalte von \tilde{A} überein. Die Lösung der obigen Aufgabenstellung erfolgt durch die folgenden Schritte.

- Gegeben sei die Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$, ferner seien die ersten drei Elemente \tilde{a}_{11} , \tilde{a}_{21} und \tilde{a}_{31} der ersten Spalte von \tilde{A} berechnet.
- Bestimme Householder-Matrix $\bar{P}_0 \in \mathbb{R}^{3 \times 3}$ mit $\bar{P}_0(\tilde{a}_{11}, \tilde{a}_{21}, \tilde{a}_{31})^T = (*, 0, 0)^T$. Setze $P_0 := \text{diag}(\bar{P}_0, I_{n-3})$, berechne $A := P_0 A \bar{P}_0$.

Nach Konstruktion ist P_0 eine Householder-Matrix, welche die erste Spalte von \tilde{A} in ein Vielfaches des ersten Einheitsvektors überführt. Daher stimmt die erste Spalte von P_0 (eventuell bis auf einen Faktor -1) mit der ersten Spalte von Q , dem orthogonalen Anteil einer QR -Zerlegung von \tilde{A} , überein. In der orthogonal ähnlichen, transformierten Matrix $A := P_0 A P_0$ wird die Hessenberg-Gestalt nur in drei Elementen gestört, und zwar denen in den Positionen $(3, 1)$, $(4, 1)$ und $(4, 2)$. Die Idee für die weiteren Schritte besteht darin, die drei störenden Elemente unterhalb der Subdiagonalen sukzessive nach unten zu schieben und schließlich aus der Matrix zu verdrängen.

- Für $k = 1, \dots, n - 2$:

Falls $k \leq n - 3$, dann:

Bestimme Householdermatrix $\bar{P}_k \in \mathbb{R}^{3 \times 3}$ mit

$$\bar{P}_k(a_{k+1,k}, a_{k+2,k}, a_{k+3,k})^T = (*, 0, 0)^T.$$

Setze $P_k := \text{diag}(I_k, \bar{P}_k, I_{n-k-3})$, berechne $A := P_k A P_k$

Andernfalls:

Bestimme Householder-Matrix $\bar{P}_{n-2} \in \mathbb{R}^{2 \times 2}$ mit

$$\bar{P}_{n-2}(a_{n-1,n-2}, a_{n,n-2})^T = (*, 0)^T.$$

Setze $P_{n-2} := \text{diag}(I_{n-2}, \bar{P}_{n-2})$, berechne $A := P_{n-2} A P_{n-2}$.

- Output: Mit der orthogonalen Matrix $V := P_0 P_1 \cdots P_{n-2}$ ist die Ausgangsmatrix A mit der orthogonal ähnlichen Hessenberg-Matrix $V^T A V$ überschrieben.

Für $k = 1, \dots, n - 2$ ist e_1 die erste Spalte der Householder-Matrizen P_k . Die erste Spalte von $V := P_0 P_1 \cdots P_{n-2}$ ist daher $P_0 e_1$, die erste Spalte von P_0 . Diese wiederum stimmt nach Wahl von P_0 (eventuell bis auf den Faktor -1) mit der ersten Spalte von Q , dem orthogonalen Anteil in einer QR -Zerlegung von \tilde{A} , überein. Ist daher $V^T A V$ unreduziert, so stimmen V und Q sowie $V^T A V$ und $Q^T A Q = A_+$ wegen Satz 2.10 im wesentlichen (und nur hierauf kommt es an) überein.

Diesen Prozeß, der von der Hessenberg-Matrix A zunächst zu $P_0 A P_0$ und dann in $n - 2$ weiteren Ähnlichkeitstransformationen mit Householder-Matrizen zu der orthogonal ähnlichen Hessenberg-Matrix $V^T A V$ führt, wollen wir uns für $n = 6$ veranschaulichen. Bei einer Transformation fest bleibende Elemente werden mit \bullet , sich verändernde mit $*$ bezeichnet.

$$\left(\begin{array}{cccccc} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array} \right) \xrightarrow{P_0} \left(\begin{array}{cccccc} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array} \right) \xrightarrow{P_1} \left(\begin{array}{cccccc} \bullet & * & * & * & \bullet & \bullet \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array} \right) \xrightarrow{P_2} \left(\begin{array}{cccccc} \bullet & * & * & * & \bullet & \bullet \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array} \right)$$

$$\left(\begin{array}{cccccc} \bullet & \bullet & * & * & * & \bullet \\ \bullet & \bullet & * & * & * & \bullet \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & \bullet \end{array} \right) \xrightarrow{P_3} \left(\begin{array}{cccccc} \bullet & \bullet & \bullet & * & * & * \\ \bullet & \bullet & \bullet & * & * & * \\ \bullet & \bullet & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{array} \right) \xrightarrow{P_4} \left(\begin{array}{ccccc} \bullet & \bullet & \bullet & \bullet & * & * \\ \bullet & \bullet & \bullet & \bullet & * & * \\ \bullet & \bullet & \bullet & \bullet & * & * \\ \bullet & \bullet & \bullet & * & * & * \\ \bullet & \bullet & * & * & * & * \\ * & * & * & * & * & * \end{array} \right)$$

Hiermit ist ein QR -Doppelschritt beschrieben, von dem man leicht nachweist, daß sein Aufwand im wesentlichen proportional zu n^2 ist. Einer Implementation dürfte nun nichts mehr im Wege stehen. Für Programme in Pseudo-Code sei auf G. W. STEWART (1973, S. 378) und G. H. GOLUB, C. F. VAN LOAN (1989, S. 376 ff.) verwiesen. Algol-Programme findet man bei J. H. WILKINSON, C. REINSCH (1971), hierauf basierende Fortran-Programme bei B. T. SMITH ET AL. (1974).

Nun sollen noch einige wenige Bemerkungen zur Berechnung der Eigenvektoren einer gegebenen Matrix $A \in \mathbb{R}^{n \times n}$ gemacht werden. In einem ersten Schritt wird A durch eine orthogonale Matrix Q_0 , einem Produkt von $n - 2$ Householder-Matrizen, auf die Hessenberg-Gestalt $A_1 = Q_0^T A Q_0$ transformiert. Will man die Eigenvektoren berechnen, so sollte man sich Q_0 merken. Es gibt im wesentlichen zwei Methoden zur Berechnung der Eigenvektoren von A , die hier nur ganz kurz angedeutet werden sollen (für nähere Erläuterungen und Algol-Programme siehe J. H. WILKINSON, C. REINSCH (1971, S. 418 ff. und S. 372 ff.)).

- Ist durch das QR -Verfahren eine Näherung μ für einen Eigenwert von A bzw. A_1 berechnet, so wende man die inverse Iteration nach Wielandt an, um einen zugehörigen Eigenvektor z von A_1 zu berechnen und damit schließlich den Eigenvektor $x := Q_0 z$ von A zu erhalten. Hier ist es also *nicht* nötig, sich die im Verlauf des QR -Verfahrens auftretenden orthogonalen Transformationsmatrizen zu merken oder gar deren Produkte zu bilden.
- Durch das QR -Verfahren mit Doppelschritt sei eine reelle Schur-Zerlegung von A_1 berechnet, also eine orthogonale Matrix Q (Produkt sämtlicher auftretender Givens-Rotationen) derart, daß $Q^T A_1 Q = R$ eine obere Block-Dreiecksmatrix ist, die in der Diagonalen 1×1 -Blöcke (reeller Eigenwert) oder 2×2 -Blöcke (zwei reelle Eigenwerte oder ein paar konjugiert komplexer Eigenwerte) enthält. Ist z ein Eigenvektor von R , so ist Qz ein Eigenvektor von A_1 und $Q_0 Qz$ ein Eigenvektor der Ausgangsmatrix A (jeweils zum gleichen Eigenwert). Es kommt also nur noch darauf an, die Eigenvektoren der oberen Block-Dreiecksmatrix R zu berechnen. Hinweise hierzu finden sich in den Aufgaben.

Aufgaben

1. Man programmiere den in 5.2.1 angegebenen Algorithmus zur Transformation einer Matrix $A \in \mathbb{R}^{n \times n}$ auf orthogonal ähnliche Hessenberg-Form (mit Hilfe von $n - 2$

Householder-Matrizen). Zur Kontrolle wende man das Programm auf die Matrix

$$A = \begin{pmatrix} 1 & 5 & 7 \\ 3 & 0 & 6 \\ 4 & 3 & 1 \end{pmatrix}$$

an und gebe die bei der Ähnlichkeitstransformation benutzte Householder-Matrix P_1 an.

Hinweis: Als Output sollten Sie

$$A = \left(\begin{array}{ccc} 1.00 & -8.60 & 0.20 \\ -5.00 & 4.96 & -0.72 \\ \hline 1.00 & 2.28 & -3.96 \end{array} \right), \quad d_1 = 2.00, \quad \beta_1 = 0.40$$

erhalten. Hieraus liest man

$$\bar{P}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.4 \begin{pmatrix} 2 \\ 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \end{pmatrix}^T = \begin{pmatrix} -0.6 & -0.8 \\ -0.8 & 0.6 \end{pmatrix}, \quad P_1 = \text{diag}(I_1, \bar{P}_1)$$

ab.

2. Man zeige, daß die Anzahl der Multiplikationen bzw. „flops“ zur Reduktion einer Matrix $A \in \mathbb{R}^{n \times n}$ (mit Hilfe von $n-2$ Householder-Matrizen P_1, \dots, P_{n-2}) auf Hessenberg-Form im wesentlichen durch $\frac{5}{3}n^3$ gegeben ist. Was ist die entsprechende Anzahl, wenn man auch noch $P_1 \cdots P_{n-2}$ berechnet?
3. Für $1 \leq r \leq s \leq n$ heißt eine Matrix der Form

$$P_{rs} := (e_1 \ \dots \ e_{r-1} \ e_s \ e_{r+1} \ \dots \ e_{s-1} \ e_r \ e_{s+1} \ \dots \ e_n)$$

eine *Vertauschungsmatrix*. Für $1 \leq k < n$ heißt eine Matrix der Form

$$M_k := I - ue_k^T \quad \text{mit} \quad u := (\underbrace{0, \dots, 0}_k, u_{k+1}, \dots, u_n)^T \in \mathbb{R}^n$$

eine *Gauß-Matrix*.

Man zeige, daß man eine gegebene Matrix $A \in \mathbb{R}^{n \times n}$ durch $n-2$ Ähnlichkeitstransformationen mit dem Produkt aus einer (symmetrischen und orthogonalen) Vertauschungsmatrix und einer Gauß-Matrix in eine ähnliche Hessenberg-Matrix überführen kann.

Hinweis: Das Verfahren könnte folgendermaßen aussehen:

- Input: Gegeben $A = (a_{ij}) \in \mathbb{R}^{n \times n}$.
- Für $k = 1, \dots, n-2$:
Bestimme $r \in \{k+1, \dots, n\}$ mit $|a_{rk}| = \max_{i=k+1, \dots, n} |a_{ik}|$
Falls $a_{rk} \neq 0$, dann:

Berechne $A := P_{k+1,r} A P_{k+1,r}$

Für $i = k+2, \dots, n$:

Berechne $l_{ik} := a_{ik}/a_{k+1,k}$

Setze $l_{k+1} := (\underbrace{0, \dots, 0}_{k+1}, l_{k+2,k}, \dots, l_{nk})^T$, $M_{k+1} := I - l_{k+1}e_{k+1}^T$

Berechne $A := M_{k+1}AM_{k+1}^{-1}$

- Output: In A steht eine zur Ausgangsmatrix ähnliche Hessenberg-Matrix.

4. Man programmiere das im Hinweis zu Aufgabe 3 skizzierte Verfahren, eine gegebene Matrix $A \in \mathbb{R}^{n \times n}$ in eine ähnliche Hessenberg-Matrix zu überführen. Anschließend teste man das Verfahren an den Matrizen

$$A = \begin{pmatrix} 1 & 5 & 7 \\ 3 & 0 & 6 \\ 4 & 3 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}.$$

Schließlich zeige man noch, daß die Anzahl der benötigten Multiplikationen bzw. flops im wesentlichen durch $\frac{5}{6}n^3$ gegeben ist, also nur halb so groß ist wie bei der entsprechenden Methode, die Householder-Matrizen benutzt.

Hinweis: Im k -ten Schritt sollte man unterhalb des Subdiagonalelementes in der k -ten Spalte die Elemente l_{ik} , $i = k + 2, \dots, n$, also die relevanten Informationen über die benutzten Gauß-Matrizen speichern. Für die angegebene 3×3 -Matrix A erhält man

$$\begin{pmatrix} 1 & 5 & 7 \\ 3 & 0 & 6 \\ 4 & 3 & 1 \end{pmatrix} \xrightarrow{P_{23}} \begin{pmatrix} 1 & 7 & 5 \\ 4 & 1 & 3 \\ 3 & 6 & 0 \end{pmatrix} \xrightarrow{M_2} \begin{pmatrix} 1 & \frac{43}{4} & 5 \\ 4 & \frac{13}{4} & 3 \\ \frac{3}{4} & \frac{57}{16} & -\frac{9}{4} \end{pmatrix}.$$

Bei J. H. WILKINSON, C. REINSCH (1971, S. 339 ff.) findet man ein Algol-Programm zu diesem Verfahren, die zweite 4×4 -Matrix dient hier als ein Test-Beispiel. Als Output berechnet man:

$$\left(\begin{array}{cccc} 1.000000000000 & 8.500000000000 & 5.321428571428 & 3.000000000000 \\ 2.000000000000 & 10.500000000000 & 6.107142857145 & 1.000000000000 \\ \hline 0.500000000000 & -7.000000000000 & -3.000000000000 & 0.000000000000 \\ 1.000000000000 & 0.107142857143 & 0.160714285714 & -1.500000000000 \end{array} \right).$$

5. Man programmiere den in 5.2.1 angegebenen Algorithmus zur Reduktion einer symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$ auf eine orthogonal ähnliche Tridiagonalmatrix. Zur Kontrolle wende man das Programm auf die Matrix

$$A = \begin{pmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 6 & 0 & 4 & 3 \\ 3 & 0 & 7 & 6 & 5 \\ 2 & 4 & 6 & 8 & 7 \\ 1 & 3 & 5 & 7 & 9 \end{pmatrix}$$

an (siehe H. R. SCHWARZ (1988, S. 252)).

Hinweis: Als Haupt- und Nebendiagonalelemente der resultierenden Tridiagonalmatrix erhält man (eventuell bis auf das Vorzeichen)

$$\delta = \begin{pmatrix} 5.0000000000 \\ 13.9333333334 \\ 9.2024742127 \\ 4.2077060891 \\ 2.6564863649 \end{pmatrix}, \quad \gamma = \begin{pmatrix} -5.4772255751 \\ 9.2985064512 \\ -2.6649567101 \\ -2.1548256624 \end{pmatrix}.$$

6. Sei $A \in \mathbb{R}^{n \times n}$ eine unreduzierte Hessenbergmatrix mit einer QR-Zerlegung $A = QR$. Dann ist $|r_{jj}| \geq |a_{j+1,j}|$ für $j = 1, \dots, n-1$.

Hinweis: Die ersten $n-1$ Spalten a_1, \dots, a_{n-1} der unreduzierten Hessenberg-Matrix A sind linear unabhängig. Bezeichnet man mit q_1, \dots, q_n die Spalten von Q , so ist daher

$$\text{span } \{q_1, \dots, q_j\} = \text{span } \{a_1, \dots, a_j\} \subset \text{span } \{e_1, \dots, e_j\}, \quad j = 1, \dots, n-1.$$

In $a_j = \sum_{i=1}^j r_{ij} q_i$ betrachte man die $(j+1)$ -te Komponente und schließe hieraus auf die Behauptung.

7. Sei $A \in \mathbb{R}^{n \times n}$ eine Hessenberg-Matrix und $\sigma \in \mathbb{R}$ kein Eigenwert von A . Durch Gauß-Elimination mit Spaltenpivotsuche berechne man eine LR-Zerlegung von $A - \sigma I$, also eine untere Dreiecksmatrix L mit Einsen in der Diagonalen, eine obere Dreiecksmatrix R und eine Permutationsmatrix P mit $P(A - \sigma I) = LR$. Anschließend setze man $A_+ := RPT L + \sigma I$. Man zeige: A_+ ist eine zu A ähnliche Hessenberg-Matrix.
8. Sei $A \in \mathbb{R}^{n \times n}$ eine Matrix mit reellen Eigenwerten $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n-1} \geq \lambda_n$. Man zeige:

- (a) Ist $\sigma < \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_1 > \lambda_2$, so ist $\lambda_1 - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.
- (b) Ist $\sigma > \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_n < \lambda_{n-1}$, so ist $\lambda_n - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.

Daher kann man bei der Anwendung der Vektoriteration auf $A - \sigma I$ nur Konvergenz gegen λ_1 oder λ_n (bzw. $\lambda_1 - \sigma$ oder $\lambda_n - \sigma$) erwarten. Wie sollte man σ in den beiden Fällen wählen, um die Konvergenzrate zu minimieren?

9. Mit Hilfe der Vektoriteration bestimme man Näherungen für den dominanten Eigenwert und einen dazugehörigen Eigenvektor der Matrix

$$A = \begin{pmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{pmatrix}$$

(siehe G. H. GOLUB, C. F. VAN LOAN (1989, S. 352)).

Hinweis: Man wähle z. B. $\|\cdot\| = \|\cdot\|_\infty$, $y^{(0)} = (1, 1, 1)^T$ und $l = 3$. Da A die Eigenwerte $\lambda_1 = 10$, $\lambda_2 = 4$ und $\lambda_3 = 3$ besitzt, wird man die Konvergenzrate $q = 0.4$ „beobachten“. Was ist der beste Shift-Parameter bei der Berechnung von λ_1 mit der Vektoriteration? Welche Konvergenzrate kann man erhalten?

10. Man führe 5 Schritte der Vektoriteration zur Bestimmung des dominanten Eigenwertes und eines zugehörigen Eigenvektors der symmetrischen Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{pmatrix}$$

durch. Mit Hilfe von Korollar 1.8 mache man eine Fehlerabschätzung.

11. Sei $A \in \mathbb{R}^{n \times n}$ eine Matrix mit reellen Eigenwerten $\lambda_1 > \dots > \lambda_n$. Man zeige:

- (a) Ist $\lambda > \frac{1}{2}(\lambda_1 + \lambda_2)$, $\lambda \neq \lambda_1$, so ist $1/(\lambda_1 - \lambda)$ ein dominanter Eigenwert von $(A - \lambda I)^{-1}$.
- (b) Sei $j \in \{2, \dots, n-1\}$. Ist $\frac{1}{2}(\lambda_{j+1} + \lambda_j) < \lambda < \frac{1}{2}(\lambda_j + \lambda_{j-1})$, $\lambda \neq \lambda_j$, so ist $1/(\lambda_j - \lambda)$ ein dominanter Eigenwert von $(A - \lambda I)^{-1}$.
- (c) Ist $\lambda < \frac{1}{2}(\lambda_n + \lambda_{n-1})$, $\lambda \neq \lambda_n$, so ist $1/(\lambda_n - \lambda)$ ein dominanter Eigenwert von $(A - \lambda I)^{-1}$.

Was bedeutet dieses Ergebnis für die inverse Iteration? Man vergleiche es mit der entsprechenden Aussage in Aufgabe 8 für die (einfache) Vektoriteration!

12. Ein Verfahren liefere als Näherungen für die Eigenwerte der Matrix

$$A = \begin{pmatrix} -2 & 2 & 2 & 2 \\ -3 & 3 & 2 & 2 \\ -2 & 0 & 4 & 2 \\ -1 & 0 & 0 & 5 \end{pmatrix}$$

die Werte $(0.99, 2.01, 2.95, 4.02)$. Mit Hilfe der inversen Iteration nach Wielandt verbessere man diese Näherungen und berechne gleichzeitig Näherungen für die Eigenvektoren. Der Startvektor $y^{(0)}$ sei zufällig erzeugt.

13. Man programmiere das in 5.2.2 angegebene Verfahren, mit Hilfe von $n-1$ Givens-Rotationen einen QR -Schritt auf die Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$ mit (explizitem) Shift $\sigma \in \mathbb{R}$ anzuwenden. Anschließend wende man das QR -Verfahren auf die Matrix

$$A = \begin{pmatrix} -3 & 9 & 0 & 1 \\ 1 & 6 & 0 & 0 \\ -23 & 23 & 4 & 3 \\ -12 & 15 & 1 & 3 \end{pmatrix}$$

(nachdem diese auf Hessenberg-Gestalt transformiert ist) an. An diesem Beispiel vergleiche man das einfache QR -Verfahren (kein Shift) mit dem mit $\sigma_k = a_{nn}^{(k)}$ geshifteten QR -Verfahren (da A nur reelle Eigenwerte besitzt, ist dieser Shift sinnvoll).

14. Zu einer Matrix $A \in \mathbb{R}^{n \times n}$ sei eine orthogonale Matrix $Q \in \mathbb{R}^{n \times n}$ bestimmt worden derart, daß $Q^T A Q = R$ eine obere Dreiecksmatrix ist. In der Diagonalen von R stehen also die (notwendigerweise reellen) Eigenwerte von A . Wir nehmen an, daß diese sogar

paarweise verschieden voneinander sind. Wie kann aus Q und R das vollständige System von Eigenvektoren zu A berechnet werden?

Hinweis: Es kommt offenbar darauf an, die Eigenvektoren der oberen Dreiecksmatrix R zu berechnen. Denn ist y ein Eigenvektor von R , so ist Qy ein Eigenvektor von A . Für den j -ten Eigenvektor von R zum Eigenwert r_{jj} mache man den Ansatz, die letzten $n - j$ Komponenten auf Null und die j -te Komponente auf Eins zu setzen, danach die ersten $j - 1$ Komponenten durch Rückwärtseinsetzen zu bestimmen.

15. Die 2×2 -Matrix $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ besitze reelle Eigenwerte. Man bestimme eine Givens-Rotation G_{12} derart, daß $G_{12}AG_{12}^T$ eine obere Dreiecksmatrix (mit den beiden reellen Eigenwerten von A in der Diagonalen) ist.
16. Die Matrix $R \in \mathbb{R}^{n \times n}$ sei eine obere Block-Dreiecksmatrix, bei der die Diagonalblöcke entweder 1×1 - oder 2×2 -Blöcke sind. Man bestimme eine orthogonal ähnliche obere Block-Dreiecksmatrix, bei der die vorhandenen 2×2 -Blöcke nur (konjugiert) komplexe Eigenwerte besitzen.
17. Man entwickle ein Verfahren, das unter geeigneten Voraussetzungen die Eigenvektoren einer oberen Block-Dreiecksmatrix $R \in \mathbb{R}^{n \times n}$ bestimmt, die in der Diagonalen nur 1×1 - und 2×2 -Blöcke besitzt.

Hinweis: Eventuell konsultiere man J. H. WILKINSON, C. REINSCH (1971, S. 372ff.).

18. Zu einer Matrix $A \in \mathbb{R}^{n \times n}$ sei eine orthogonale Matrix $Q_0 \in \mathbb{R}^{n \times n}$ bestimmt derart, daß $Q_0^T A Q_0$ die Form

$$Q_0^T A Q_0 = \left(\begin{array}{c|cc} B & C \\ \hline 0 & \alpha & \beta \\ 0 & \gamma & \delta \end{array} \right)$$

hat, wobei $B \in \mathbb{R}^{(n-2) \times (n-2)}$ eine Hessenberg-Matrix ist. Typischerweise tritt dies (zumindestens näherungsweise) auf, wenn nach einer Folge von QR-Doppelschritten eine Reduktion auf eine $(n - 2)$ -dimensionale Eigenwertaufgabe erfolgt. Ferner sei eine orthogonale Matrix $Q_1 \in \mathbb{R}^{(n-2) \times (n-2)}$ bekannt, für die $Q_1^T B Q_1 = R_1$ eine obere Block-Dreiecksmatrix ist, in deren Diagonale nur 1×1 - oder 2×2 -Blöcke stehen. Aus diesen Daten bestimme man eine orthogonale Matrix $Q \in \mathbb{R}^{n \times n}$ und eine obere Block-Dreiecksmatrix R (mit 1×1 - oder 2×2 -Blöcken in der Diagonalen) mit $Q^T A Q = R$.

19. Man programmiere den in 5.2.6 angegebenen QR-Doppelschritt. Man teste das Programm, indem man die Matrix

$$A = \begin{pmatrix} 4 & -5 & 0 & 3 \\ 0 & 4 & -3 & -5 \\ 5 & -3 & 4 & 0 \\ 3 & 0 & 5 & 4 \end{pmatrix}$$

zunächst auf Hessenberg-Gestalt transformiert und dann auf die resultierende Matrix eine Folge von QR-Doppelschritten durchführt.

Hinweis: A hat die Eigenwerte $\lambda_1 = 12$, $\lambda_{2,3} = 1 \pm 5i$, $\lambda_4 = 2$.

5.3 Eigenwertaufgaben für symmetrische Matrizen

Ziel dieses Abschnittes ist es, Verfahren zur Berechnung der Eigenwerte und (eventuell) der Eigenvektoren einer *symmetrischen*³ Matrix $A \in \mathbb{R}^{n \times n}$ anzugeben, zu motivieren und zu analysieren. Grundlegend für die numerische Behandlung der Eigenwertaufgabe für symmetrische Matrizen ist:

- Die Eigenwerte einer symmetrischen Matrix sind reell.
- Eine symmetrische Matrix ist einer Diagonalmatrix orthogonal ähnlich. D. h. zu jeder symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$ existiert eine orthogonale Matrix Q derart, daß $Q^T A Q = \Lambda$ eine Diagonalmatrix ist. Die Diagonalelemente von Λ sind Eigenwerte, die Spalten von Q bilden ein orthonormiertes System von Eigenvektoren von A .

In 5.3.1 werden wir eines der ältesten Verfahren (es stammt aus dem Jahre 1846) zur Berechnung der Eigenwerte und Eigenvektoren einer symmetrischen Matrix, das Jacobi-Verfahren, schildern. Wegen seiner Einfachheit (insbesondere was die Implementation angeht) und Stabilität ist es nach wie vor beliebt, auch wenn ihm das in 5.3.3 zu besprechende QR-Verfahren (für symmetrische Matrizen) inzwischen weitgehend den Rang abgelaufen hat. Das Bisektionsverfahren in 5.3.2 erlaubt es, gewisse Eigenwerte (z. B. den größten oder kleinsten Eigenwert) einer symmetrischen Matrix zu berechnen. Schließlich werden wir in 5.3.4 noch kurz auf die Berechnung der Singulärwertzerlegung, insbesondere also die der singulären Werte, einer gegebenen Matrix eingehen.

Für eine wesentlich ausführlichere Darstellung des symmetrischen Eigenwertproblems sei insbesondere auf B. N. PARLETT (1980) verwiesen.

5.3.1 Das Jacobi-Verfahren

Im folgenden sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ stets eine symmetrische Matrix. Als Maß für die Abweichung von A von Diagonalgestalt wird die Summe der quadrierten Außerdia-
gonalelemente angesehen und mit $N(A)$ bezeichnet:

$$N(A) := \sum_{\substack{i,j=1 \\ i \neq j}}^n a_{ij}^2.$$

Die Idee des *klassischen Jacobi-Verfahrens* ist einfach. Man bestimme ein Paar (p, q) mit $1 \leq p < q \leq n$ und $|a_{pq}| = \max_{1 \leq i < j \leq n} |a_{ij}|$, also ein dem Betrage nach maximales Außerdia-
gonalelement, und anschließend eine Givens-Rotation⁴ G_{pq} derart, daß in $B := G_{pq}^T A G_{pq}$ das (p, q) -Element annulliert ist. Dann kann leicht gezeigt werden, daß eine von A unabhängige Konstante $c \in (0, 1)$ mit $N(B) \leq c N(A)$ existiert, die

³Die Übertragung der Verfahren auf (komplexe) hermitesche Matrizen ist meistens naheliegend, wir werden darauf verzichten.

⁴Häufig spricht man in diesem Zusammenhang, Jacobi zu Ehren, auch von *Jacobi-Rotationen*.

Summe $N(B)$ der quadrierten Außendiagonalelemente von B gegenüber der von A also gleichmäßig verkleinert wird. Setzt man dieses Verfahren mit B statt A fort (hierbei werden allerdings in einem Schritt annullierte Elemente im Laufe des Verfahrens i. allg. wieder ungleich Null!), so erhält man eine Folge $\{A_k\}$ von zur Ausgangsmatrix A orthogonal ähnlichen Matrizen mit $\lim_{k \rightarrow \infty} N(A_k) = 0$ bzw. $\lim_{k \rightarrow \infty} a_{ij}^{(k)} = 0$ für $i \neq j$. Mit Hilfe des Satzes von Gerschgorin (Satz 1.3) oder Korollar 1.12 schließt man hieraus, daß die Diagonalelemente $a_{ii}^{(k)}$, $i = 1, \dots, n$, (eventuell nach geeigneter Umnumerierung) gegen die Eigenwerte von A konvergieren.

Da die Suche nach einem betragsmaximalen Element verhältnismäßig aufwendig ist, geht man in der Praxis vom klassischen zum *zyklischen Jacobi-Verfahren* kombiniert mit einer sogenannten *Schwellenmethode* über. Hierbei werden die Außendiagonalelemente in einer festen Reihenfolge durchlaufen, etwa zeilenweise von $(1, 2)$ über $(1, n), (2, 3)$ über $(2, n)$ bis $(n-1, n)$, eine Transformation aber nur dann durchgeführt, wenn das zu annullierende Element betragsmäßig nicht kleiner als ein vorgegebener Schwellenwert $\epsilon > 0$ ist. Sobald alle Außendiagonalelemente betragsmäßig kleiner als der Schwellenwert ϵ sind, so wird dieser heruntergesetzt und das Verfahren entsprechend fortgesetzt.

Für $1 \leq p < q \leq n$ und $c = \cos(\phi)$, $s = \sin(\phi)$ ist die zugehörige Givens-Rotation durch

$$G_{pq}(\phi) = \begin{pmatrix} 1 & & & & & \\ & \vdots & & & & \\ \cdots & c & \cdots & s & \cdots & \\ & \vdots & & \vdots & & \\ \cdots & -s & \cdots & c & \cdots & \\ & \vdots & & \vdots & & 1 \end{pmatrix} \quad \begin{matrix} p \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ p & & & & & q \end{matrix}$$

gegeben. Nun ist es leicht, das folgende Lemma zu beweisen.

Lemma 3.1 Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ symmetrisch, ferner sei (p, q) ein Indexpaar mit $1 \leq p < q \leq n$ und $a_{pq} \neq 0$. Ein Drehwinkel $\phi \in [-\pi/4, \pi/4]$ sei durch

$$\begin{aligned} \cot 2\phi &= \frac{a_{qq} - a_{pp}}{2a_{pq}} \quad \text{falls } a_{pp} \neq a_{qq}, \\ \phi &= \frac{\pi}{4} \quad \text{falls } a_{pp} = a_{qq} \end{aligned}$$

gegeben. Anschließend seien die Givens-Rotation $G_{pq} := G_{pq}(\phi)$ sowie

$$B := G_{pq}^T A G_{pq} = (b_{ij})$$

definiert. Dann gilt:

1. Es ist $b_{pq} = b_{qp} = 0$, in B wird also das (p, q) -Element annulliert.
2. Es ist $N(B) = N(A) - 2a_{pq}^2 < N(A)$.

3. Ist $|a_{pq}| = \max_{1 \leq i < j \leq n} |a_{ij}|$, so ist $N(B) \leq [1 - 2/(n^2 - n)] N(A)$.

Beweis: Offenbar ist

$$(*) \quad \begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix}.$$

Daher ist

$$\begin{aligned} b_{pq} = b_{qp} &= (a_{pp} - a_{qq}) \cos \phi \sin \phi + a_{pq}(\cos^2 \phi - \sin^2 \phi) \\ &= \frac{1}{2}(a_{pp} - a_{qq}) \sin 2\phi + a_{pq} \cos 2\phi \\ &= 0 \quad \text{nach Definition des Drehwinkels } \phi. \end{aligned}$$

Die Frobenius-Norm $\|\cdot\|_F$ ist invariant unter orthogonalen Ähnlichkeitstransformationen, wie man z. B. aus

$$\|C\|_F^2 = \sum_{i,j=1}^n c_{ij}^2 = \text{Spur}(C^T C) \quad \text{für } C = (c_{ij}) \in \mathbb{R}^{n \times n}$$

erkennt. In der zu A orthogonal ähnlichen Matrix $B = G_{pq}^T A G_{pq}$ verändern sich gegenüber A lediglich die p -te und q -te Zeile sowie Spalte. Wegen $(*)$, $b_{pq} = b_{qp} = 0$ und der Invarianz der Frobenius-Norm unter orthogonalen Ähnlichkeitstransformationen ist ferner $b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2$. Daher ist

$$N(B) = \|B\|_F^2 - \sum_{i=1}^n b_{ii}^2 = \|A\|_F^2 - \sum_{\substack{i=1 \\ i \neq p,q}}^n a_{ii}^2 - b_{pp}^2 - b_{qq}^2 = \|A\|_F^2 - \sum_{i=1}^n a_{ii}^2 - 2a_{pq}^2 = N(A) - 2a_{pq}^2.$$

Wählt man (p, q) so, daß $|a_{pq}| = \max_{1 \leq i < j \leq n} |a_{ij}|$, so ist $a_{ij}^2 \leq a_{pq}^2$ für $i \neq j$ und daher $N(A) \leq (n^2 - n)a_{pq}^2$. Folglich ist

$$N(B) = N(A) - 2a_{pq}^2 \leq \left(1 - \frac{2}{n^2 - n}\right) N(A),$$

womit das Lemma bewiesen ist. \square

Es ist *nicht* sinnvoll, bei gegebenem Indexpaar (p, q) mit $a_{pq} \neq 0$ den Drehwinkel ϕ , wie in Lemma 3.1 angegeben, aus $\cot 2\phi = (a_{qq} - a_{pp})/(2a_{pq})$ zu berechnen, anschließend $\cos \phi$ sowie $\sin \phi$ zu erhalten, und die neuen p -ten und q -ten Spalten (und Zeilen) von $B = G_{pq}^T A G_{pq}$ sozusagen auf naive Weise zu bestimmen. Statt dessen geht man nach H. RUTISHAUSER (1966) (siehe auch J. H. WILKINSON, C. REINSCH (1971, S. 202 ff.)) folgendermaßen vor:

- Berechne

$$\vartheta := \frac{a_{qq} - a_{pp}}{2a_{pq}} \quad (= \cot 2\phi).$$

- Es ist $\tan^2 \phi + 2 \cot 2\phi \tan \phi = 1$. Daher kann man $t := \tan \phi$ aus der quadratischen Gleichung $t^2 + 2\vartheta t = 1$ mit den Lösungen $t_{1,2} = -\vartheta \pm \sqrt{1 + \vartheta^2}$ als die betragsmäßig kleinere Lösung

$$t := \frac{\operatorname{sign}(\vartheta)}{|\vartheta| + \sqrt{1 + \vartheta^2}} = -\vartheta + \operatorname{sign}(\vartheta) \sqrt{1 + \vartheta^2}$$

berechnen. Für kleine $|\vartheta|$ kann man $t := 1$ setzen, für große $|\vartheta|$ setzt Rutishauser (um einen overflow bei der Berechnung von ϑ^2 zu vermeiden) $t := 0.5/\vartheta$.

- Berechne

$$\begin{aligned} c &:= \frac{1}{\sqrt{1+t^2}} = \cos \phi, \\ s &:= ct = \sin \phi, \\ \tau &:= \frac{s}{1+c} = \tan \frac{\phi}{2}. \end{aligned}$$

- Berechne die p -te und die q -te Zeile sowie Spalte von $B := G_{pq}^T A G_{pq}$, also b_{pp} und b_{qq} sowie $b_{pj} = b_{jp}$, $b_{qj} = b_{jq}$ für $j \neq p, q$ durch

$$b_{pp} := a_{pp} - ta_{pq}, \quad b_{qq} := a_{qq} + ta_{pq}$$

sowie

$$b_{pj} := a_{pj} - s(a_{qj} + \tau a_{pj}), \quad b_{qj} := a_{qj} + s(a_{pj} - \tau a_{qj}) \quad (j \neq p, q).$$

Die Korrektheit dieser Beziehungen weist man leicht nach.

Nun ist es einfach, den angekündigten Konvergenzsatz für das klassische Jacobi-Verfahren zu beweisen.

Satz 3.2 Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch mit Eigenwerten $\lambda_1 \geq \dots \geq \lambda_n$. Sei $\{Q_k\}$ eine Folge von Givens-Rotationen und $\{A_k\}$ eine Folge zu A orthogonal ähnlicher Matrizen, die auf die folgende Weise konstruiert seien:

- Setze $A_1 := A$.

- Für $k = 1, 2, \dots$

Bestimme (p, q) mit $1 \leq p < q \leq n$ und $|a_{pq}^{(k)}| = \max_{1 \leq i < j \leq n} |a_{ij}^{(k)}|$.

Bestimme eine Givens-Rotation $Q_k := G_{pq}$ mit $(G_{pq}^T A_k G_{pq})_{pq} = 0$.

Berechne $A_{k+1} := Q_k^T A_k Q_k$.

Dann gilt:

1. $\lim_{k \rightarrow \infty} N(A_k) = 0$, für $i \neq j$ konvergiert also die Folge der Außerdagonalelemente $\{a_{ij}^{(k)}\}$ gegen Null.
2. Sind $a_1^{(k)} \geq \dots \geq a_n^{(k)}$ die Diagonalelemente von A_k , so ist $\lim_{k \rightarrow \infty} a_j^{(k)} = \lambda_j$ für $j = 1, \dots, n$.

Beweis: Wegen Lemma 3.1 ist $N(A_{k+1}) \leq [1 - 2/(n^2 - n)] N(A_k)$, woraus (für $n \geq 2$, was implizit vorausgesetzt sei) $\lim_{k \rightarrow \infty} N(A_k) = 0$ folgt. Wegen Korollar 1.12 ist $|\lambda_j - a_j^{(k)}| \leq N(A_k)^{1/2}$ für $j = 1, \dots, n$. Damit ist der Konvergenzsatz für das klassische Jacobi-Verfahren bewiesen. \square

Bemerkungen: Akkumuliert man die beim Jacobi-Verfahren auftretenden Rotationen in einer Matrix, bildet man also $\hat{Q}_k := Q_0 \cdots Q_k$, so erhält man in den Spalten von \hat{Q}_k Näherungen an das System der Eigenvektoren der Ausgangsmatrix.

Ein Nachteil des Jacobi-Verfahrens besteht darin, daß eine spezielle Struktur der Ausgangsmatrix, wie etwa eine Tridiagonalgestalt, nicht respektiert bzw. im Laufe des Verfahrens zerstört wird.

Oben ist die Konvergenz des klassischen Jacobi-Verfahrens bewiesen worden, bei dem in jedem Schritt das betragsgrößte Außerdiagonalelement von A_k durch eine zweidimensionale Rotation annulliert wird. Offenbar „zieht“ der Konvergenzbeweis immer noch, wenn man in jedem Schritt (p, q) so wählt, daß $(a_{pq}^{(k)})^2$ nicht kleiner als der Mittelwert der quadrierten Außerdiagonalelemente von A_k ist, wenn also $(a_{pq}^{(k)})^2 \geq N(A_k)/(n^2 - n)$. Angemerkt sei ferner, daß für $i = 1, \dots, n$ die Konvergenz von $\{a_{ii}^{(k)}\}$ gegen einen Eigenwert von A gezeigt werden kann (siehe J. H. WILKINSON (1965, S. 268 ff.), dort findet man auch Hinweise zur Konvergenzgeschwindigkeit).

Wie zu Beginn schon angegeben, wird man in der Praxis nicht das klassische Jacobi-Verfahren, sondern das zyklische Jacobi-Verfahren in Kombination mit einer Schwellenmethode anwenden. Hierbei ist eine Nullfolge $\{\epsilon_k\}$ positiver Zahlen (z. B. ist $\epsilon_k := \epsilon^k$ mit $\epsilon \in (0, 1)$) vorgegeben. Man durchläuft die Außerdiagonalelemente (i, j) mit $1 \leq i < j \leq n$ in einer festen Reihenfolge und führt für ein Indexpaar (p, q) eine Rotation mit G_{pq} nur dann durch, wenn $|a_{pq}^{(k)}| \geq \epsilon_k$. Hierbei wird $N(A_k)$ um mindestens $2\epsilon_k^2$ vermindert und $A_k := G_{pq}^T A_k G_{pq}$ gesetzt. Nach endlich vielen Schritten ist $|a_{ij}^{(k)}| < \epsilon_k$ für alle Außerdiagonalelemente von A_k . Dann wird $k := k + 1$ gesetzt und entsprechend fortgefahrene. Trivialerweise ist dieses Verfahren konvergent, da die Außerdiagonalelemente gegen Null konvergieren. \square

5.3.2 Das Bisektions-Verfahren

In 5.2.1 wurde beschrieben, wie man durch $n - 2$ Ähnlichkeitstransformationen mit Householder-Matrizen P_1, \dots, P_{n-2} eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ auf orthogonal ähnliche Tridiagonalgestalt transformieren kann. Ziel ist es nun, das Bisektionsverfahren zur Bestimmung bestimmter Eigenwerte einer symmetrischen Tridiagonalmatrix zu beschreiben.

Im folgenden Satz wird u. a. eine Methode angegeben, den Wert des charakteristischen Polynoms einer symmetrischen Tridiagonalmatrix rekursiv zu berechnen.

Satz 3.3 Bei vorgegebenen reellen Zahlen $\delta_1, \dots, \delta_n$ und von Null verschiedenen Zahlen $\gamma_1, \dots, \gamma_{n-1}$ seien für $i = 1, \dots, n$ die unreduzierten, symmetrischen Tridiagonalmatrizen $A_i \in \mathbb{R}^{i \times i}$ mit den Hauptdiagonalelementen $\delta_1, \dots, \delta_i$ sowie den Nebendiagonalelementen $\gamma_1, \dots, \gamma_{i-1}$ definiert. Sei $p_i(\mu) := \det(A_i - \mu I)$ das zugehörige charakteristische Polynom. Dann gilt die Rekursionsformel

$$p_i(\mu) = (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu), \quad i = 2, \dots, n,$$

mit $p_0(\mu) := 1$ und $p_1(\mu) := \delta_1 - \mu$. Ferner gilt:

1. p_n besitzt nur einfache (reelle) Nullstellen $\lambda_1 > \dots > \lambda_n$. Insbesondere besitzt eine unreduzierte, symmetrische Tridiagonalmatrix paarweise verschiedene Eigenwerte.
2. Für $j = 1, \dots, n$ ist $\text{sign } p_{n-1}(\lambda_j) = -\text{sign } p'_n(\lambda_j)$.
3. Für $i = 1, \dots, n-1$ folgt aus $p_i(\xi) = 0$, daß $p_{i+1}(\xi) p_{i-1}(\xi) < 0$.

Beweis: Die Rekursionsformel ist offenbar für $i = 2$ richtig. Für $i = 3, \dots, n$ erhält man durch Determinantenentwicklung nach der letzten Spalte bzw. Zeile:

$$\begin{aligned} p_i(\mu) &= \det \begin{pmatrix} \delta_1 - \mu & \gamma_1 & & \\ \gamma_1 & \delta_2 - \mu & \gamma_2 & \\ & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \gamma_{i-1} \\ & \gamma_{i-1} & \delta_i - \mu & \end{pmatrix} \\ &= (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1} \det \begin{pmatrix} \delta_1 - \mu & \gamma_1 & & \\ \gamma_1 & \delta_2 - \mu & \gamma_2 & \\ & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \gamma_{i-2} \\ & 0 & \gamma_{i-1} & \end{pmatrix} \\ &= (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu). \end{aligned}$$

Damit ist die Rekursionsformel bewiesen.

Die Nullstellen $\lambda_1, \dots, \lambda_n$ von p_n sind natürlich als Eigenwerte der symmetrischen Matrix A_n reell. Formal setze man $\gamma_n := 1$ und definiere

$$q_i(\mu) := \begin{cases} 1 & \text{für } i = 0, \\ (-1)^i \frac{p_i(\mu)}{\gamma_1 \cdots \gamma_i} & \text{für } i = 1, \dots, n, \end{cases} \quad q(\mu) := (q_0(\mu), \dots, q_{n-1}(\mu))^T.$$

Dann ist

$$(*) \quad (A_n - \mu I)q(\mu) = (0, \dots, 0, -q_n(\mu))^T,$$

wie man sehr leicht mit Hilfe der Rekursionsformel nachweist. Eine Nullstelle λ_j von p_n ist auch eine Nullstelle von q_n , so daß $(A_n - \lambda_j I)q(\lambda_j) = 0$ wegen (*). Da $q_0(\lambda_j) \neq 0$, ist $q(\lambda_j)$ ein Eigenvektor von A_n zum Eigenwert λ_j . Eine Differentiation von (*) nach μ liefert

$$(**) \quad -q(\mu) + (A_n - \mu I)q'(\mu) = (0, \dots, 0, -q'_n(\mu))^T.$$

Eine Multiplikation der Gleichung (**) von links mit $q(\mu)^T$ ergibt

$$-q(\mu)^T q(\mu) + [(A_n - \mu I)q(\mu)]^T q'(\mu) = -q_{n-1}(\mu) q'_n(\mu).$$

Setzt man hier $\mu = \lambda_j$, so folgt

$$0 < \|q(\lambda_j)\|_2^2 = q_{n-1}(\lambda_j) q'_n(\lambda_j) = -\frac{p_{n-1}(\lambda_j) p'_n(\lambda_j)}{\gamma_1^2 \cdots \gamma_{n-1}^2}.$$

Hieraus folgt:

1. Es ist $p'_n(\lambda_j) \neq 0$, d. h. die Nullstellen von p_n bzw. die Eigenwerte von A_n sind einfach.
2. Es ist $p_{n-1}(\lambda_j) p'_n(\lambda_j) < 0$ bzw. $\text{sign } p_{n-1}(\lambda_j) = -\text{sign } p'_n(\lambda_j)$ für jede Nullstelle λ_j von p_n .

Sei schließlich $p_i(\xi) = 0$ für ein $i \in \{1, \dots, n-1\}$. Wegen der Rekursionsformel ist

$$p_{i+1}(\xi) = (\delta_{i+1} - \xi) \underbrace{p_i(\xi)}_{=0} - \gamma_i^2 p_{i-1}(\xi) = -\gamma_i^2 p_{i-1}(\xi).$$

Wäre $p_{i+1}(\xi) = 0$, so wäre auch $p_{i-1}(\xi) = 0$. Sukzessive würde man $p_0(\xi) = 0$ erhalten, was wegen $p_0(\xi) = 1$ einen Widerspruch bedeutet. Insgesamt folgt aus $p_i(\xi) = 0$ daher $p_{i+1}(\xi) p_{i-1}(\xi) < 0$. Damit ist der Satz bewiesen. \square

Bemerkung: Aus der Rekursionsformel für die Polynome p_i in Satz 3.3 erhält man durch Differenzieren $p'_0(\mu) = 0$, $p'_1(\mu) = -1$ und

$$p'_i(\mu) = -p_{i-1}(\mu) + (\delta_i - \mu)p'_{i-1}(\mu) - \gamma_{i-1}^2 p'_{i-2}(\mu), \quad i = 2, \dots, n.$$

Daher kann auch $p'_n(\mu)$ rekursiv berechnet und zur Bestimmung einer Nullstelle von p_n bzw. eines Eigenwertes von A_n das Newton-Verfahren angesetzt werden. \square

Bemerkung: Entsprechend den Aussagen von Satz 3.3 gilt natürlich für $i = 1, \dots, n$:

1. p_i besitzt i einfache Nullstellen $\lambda_1^{(i)} > \dots > \lambda_i^{(i)}$.
2. $p_{i-1}(\lambda_j^{(i)}) p'_i(\lambda_j^{(i)}) < 0$ für $j = 1, \dots, i$.
3. $p_{i+1}(\lambda_j^{(i)}) p_{i-1}(\lambda_j^{(i)}) < 0$ für $j = 1, \dots, i$.

Hieraus erhält man:

- Für $i = 1, \dots, n$ werden die Nullstellen von p_i streng durch die Nullstellen von p_{i-1} getrennt, d. h. es gilt

$$\lambda_1^{(i)} > \lambda_1^{(i-1)} > \lambda_2^{(i)} > \lambda_2^{(i-1)} > \dots > \lambda_{i-1}^{(i)} > \lambda_{i-1}^{(i-1)} > \lambda_i^{(i)}, \quad i = 1, \dots, n.$$

Denn: Da $\lambda_j^{(i)}$ und $\lambda_{j+1}^{(i)}$ aufeinanderfolgende einfache Nullstellen von p_i sind, hat die Ableitung p'_i in diesen Nullstellen unterschiedliches Vorzeichen:

$$p'_i(\lambda_j^{(i)}) p'_i(\lambda_{j+1}^{(i)}) < 0, \quad j = 1, \dots, i-1.$$

Da ferner

$$p_{i-1}(\lambda_j^{(i)}) p'_i(\lambda_j^{(i)}) < 0, \quad p_{i-1}(\lambda_{j+1}^{(i)}) p'_i(\lambda_{j+1}^{(i)}) < 0, \quad j = 1, \dots, i-1,$$

ist

$$p_{i-1}(\lambda_j^{(i)}) p_{i-1}(\lambda_{j+1}^{(i)}) < 0, \quad j = 1, \dots, i-1.$$

Also hat p_{i-1} in aufeinanderfolgenden Nullstellen von p_i unterschiedliches Vorzeichen, woraus $\lambda_j^{(i)} > \lambda_{j-1}^{(i-1)} > \lambda_{j+1}^{(i)}$ und damit die behauptete Trennungseigenschaft folgt. \square

Der folgende Satz bildet die Grundlage für das Bisektionsverfahren.

Satz 3.4 Gegeben sei die unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ mit den Hauptdiagonalelementen $\delta_1, \dots, \delta_n$ und den (von Null verschiedenen) Nebendiagonalelementen $\gamma_1, \dots, \gamma_{n-1}$. Die Polynome $p_i \in \Pi_i$ seien für $i = 0, \dots, n$ durch die Rekursionsformel

$$p_0(\mu) := 1, \quad p_1(\mu) := \delta_1 - \mu, \quad p_i(\mu) := (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu), \quad i = 2, \dots, n,$$

definiert. Für $\xi \in \mathbb{R}$ sei $N_n(\xi)$ die Anzahl aufeinanderfolgender Vorzeichenübereinstimmungen in $(p_0(\xi), p_1(\xi), \dots, p_n(\xi))$. Hierbei wird vereinbart: Ist $p_k(\xi) = 0$ für ein $k \in \{1, \dots, n\}$, so erhält $p_k(\xi)$ das Vorzeichen von $p_{k-1}(\xi) \neq 0$. Dann gibt es genau $N_n(\xi)$ Eigenwerte von A (bzw. Nullstellen von p_n), welche größer oder gleich ξ sind.

Beispiel: Für $n = 3$ erhält man z. B.

$$(p_0(\xi), p_1(\xi), p_2(\xi), p_3(\xi)) \longrightarrow \begin{cases} (+, 0, -, 0) & \longrightarrow (+, +, -, -) \implies N_3(\xi) = 2, \\ (+, -, 0, +) & \longrightarrow (+, -, -, +) \implies N_3(\xi) = 1, \\ (+, -, +, -) & \implies N_3(\xi) = 0. \end{cases}$$

\square

Beweis: Der Satz wird durch vollständige Induktion nach n bewiesen.

Wegen $p_0(\xi) = 1$ und $p_1(\xi) = \delta_1 - \xi$ sind für $n = 1$ zwei Fälle möglich:

$$(p_0(\xi), p_1(\xi)) \longrightarrow \begin{cases} (+, +) \implies N_1(\xi) = 1, & \delta_1 - \xi \geq 0 \text{ bzw. } \xi \leq \lambda_1^{(1)}, \\ (+, -) \implies N_1(\xi) = 0, & \delta_1 - \xi < 0 \text{ bzw. } \xi > \lambda_1^{(1)}. \end{cases}$$

In beiden Fällen ist die Behauptung richtig.

Wir nehmen nun an, die Behauptung sei für $n-1$ richtig. Sei $q := N_{n-1}(\xi)$, so daß p_{n-1} nach Induktionsvoraussetzung q Nullstellen besitzt, die größer oder gleich ξ sind. Für $j = 1, \dots, n-1$ seien $\mu_j := \lambda_j^{(n-1)}$ die der Größe nach geordneten (einfachen) Nullstellen von p_{n-1} . Dann ist also

$$\mu_{n-1} < \dots < \mu_{q+1} < \xi \leq \mu_q < \dots < \mu_1 < \lambda_1$$

Sind $\lambda_n < \dots < \lambda_1$ die Nullstellen von p_n , so werden diese nach obiger Bemerkung durch die Nullstellen von p_{n-1} getrennt. Daher ist

$$\lambda_n < \mu_{n-1} < \lambda_{n-1} < \dots < \lambda_{q+2} < \mu_{q+1} < \xi \leq \mu_q < \lambda_q < \dots < \mu_1 < \lambda_1$$

und $\lambda_{q+1} \in (\mu_{q+1}, \mu_q)$. Wir zeigen nun:

- (a) Ist $\xi \leq \lambda_{q+1}$, so ist $N_n(\xi) = N_{n-1}(\xi) + 1$ bzw. $\operatorname{sign} p_n(\xi) = \operatorname{sign} p_{n-1}(\xi)$ oder $p_n(\xi) = 0$.

Denn: Ist $\xi < \lambda_{q+1}$ (andernfalls ist $p_n(\xi) = 0$ und die Aussage richtig), so ist $p_n(\xi) \neq 0$ und $p_{n-1}(\xi) \neq 0$. Wegen

$$p_{n-1}(\xi) = \prod_{j=1}^{n-1} (\mu_j - \xi) = \prod_{j=1}^q \underbrace{(\mu_j - \xi)}_{>0} \prod_{j=q+1}^{n-1} \underbrace{(\mu_j - \xi)}_{<0}$$

ist $\operatorname{sign} p_{n-1}(\xi) = (-1)^{n-q-1}$. Entsprechend erkennt man an

$$p_n(\xi) = \prod_{j=1}^n (\lambda_j - \xi) = \prod_{j=1}^{q+1} \underbrace{(\lambda_j - \xi)}_{>0} \prod_{j=q+2}^n \underbrace{(\lambda_j - \xi)}_{<0},$$

daß auch $\operatorname{sign} p_n(\xi) = (-1)^{n-q-1}$. Also ist $N_n(\xi) = N_{n-1}(\xi) + 1$.

- (b) Ist $\lambda_{q+1} < \xi$, so ist $N_n(\xi) = N_{n-1}(\xi)$ bzw. $\operatorname{sign} p_n(\xi) = -\operatorname{sign} p_{n-1}(\xi)$ falls $p_{n-1}(\xi) \neq 0$ oder $\operatorname{sign} p_n(\xi) = -\operatorname{sign} p_{n-2}(\xi)$ falls $p_{n-1}(\xi) = 0$.

Denn: Ist $\lambda_{q+1} < \xi < \mu_q$, so erhält man wie in (a), daß $\operatorname{sign} p_{n-1}(\xi) = (-1)^{n-q-1}$ und $\operatorname{sign} p_n(\xi) = (-1)^{n-q}$, also $N_n(\xi) = N_{n-1}(\xi)$. Ist dagegen $\lambda_{q+1} < \xi = \mu_q$, so ist $p_{n-1}(\xi) = 0$ und wegen Satz 3.3 folgt $p_n(\xi) p_{n-2}(\xi) < 0$, so daß auch in diesem Falle $N_n(\xi) = N_{n-1}(\xi)$ folgt.

Insgesamt haben wir gezeigt, daß $N_n(\xi)$ die Anzahl der Nullstellen von p_n angibt, die größer oder gleich ξ sind. Damit ist der Satz bewiesen. \square

Mit den Bezeichnungen von Satz 3.4 gilt daher:

Korollar 3.5 Seien $\lambda_1 > \dots > \lambda_n$ die Eigenwerte der unreduzierten, symmetrischen Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$. Es sei $\lambda_j \in [a, b]$ und $\xi \in (a, b)$ (z.B. $\xi = (a+b)/2$). Dann gilt: Ist $N_n(\xi) < j$, so ist $\lambda_j \in [a, \xi]$. Ist dagegen $N_n(\xi) \geq j$, so ist $\lambda_j \in [\xi, b]$.

Damit erhält man das *Bisektionsverfahren* zur Berechnung des j -ten Eigenwerts λ_j einer unreduzierten, symmetrischen Tridiagonalmatrix.

- Gegeben sei die unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ mit Hauptdiagonalelementen $\delta_1, \dots, \delta_n$ und Nebendiagonalelementen $\gamma_1, \dots, \gamma_{n-1}$. Die Polynome $p_i \in \Pi_i$, $i = 0, \dots, n$ seien rekursiv durch

$$p_0(\mu) := 1, \quad p_1(\mu) := \delta_1 - \mu$$

und

$$p_i(\mu) := (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu), \quad i = 2, \dots, n,$$

definiert. Ferner sei $j \in \{1, \dots, n\}$ und $[a_1, b_1]$ ein gegebenes Intervall mit $\lambda_j \in [a_1, b_1]$.

- Für $k = 1, 2, \dots$:

Sei $\xi_k := (a_k + b_k)/2$.

Berechne $(p_0(\xi_k), \dots, p_n(\xi_k))$ und hiermit die Anzahl $N_n(\xi_k)$ aufeinanderfolgender Vorzeichenübereinstimmungen. Hierbei wird vereinbart: Ist $p_i(\xi_k) = 0$, so erhält $p_i(\xi_k)$ das Vorzeichen von $p_{i-1}(\xi_k)$.

Falls $N_n(\xi_k) < j$, dann: $a_{k+1} := a_k, b_{k+1} := \xi_k$
 Andernfalls: $a_{k+1} := \xi_k, b_{k+1} := b_k$.

Bemerkung: Alle Eigenwerte der symmetrischen Tridiagonalmatrix A mit den Hauptdiagonalelementen $\delta_1, \dots, \delta_n$ und den Nebendiagonalelementen $\gamma_1, \dots, \gamma_{n-1}$ liegen wegen des Satzes von Gerschgorin (Satz 1.3) in dem Intervall $[a_{\min}, b_{\max}]$ mit

$$a_{\min} := \min_{i=1, \dots, n} [\delta_i - (|\gamma_{i-1}| + |\gamma_i|)], \quad b_{\max} := \max_{i=1, \dots, n} [\delta_i + (|\gamma_{i-1}| + |\gamma_i|)],$$

wobei $\gamma_0 := 0, \gamma_n := 0$ gesetzt wird. Dieses Intervall kann daher stets als Ausgangsintervall $[a_1, b_1]$ für das Bisektionsverfahren genommen werden. Für genauere Hinweise zur Implementation des Bisektionsverfahrens sei auf J. H. WILKINSON, C. REINSCH (1971, S. 249 ff.) hingewiesen (siehe auch Aufgabe 2). \square

5.3.3 Das QR-Verfahren für symmetrische Matrizen

Da zu Beginn des QR-Verfahrens grundsätzlich ein Reduktionsschritt gemacht wird (dieser wurde in 5.2.1 beschrieben), sei $A \in \mathbb{R}^{n \times n}$ nun eine symmetrische Tridiagonalmatrix. Uns interessiert zunächst, wie ein Schritt des QR-Verfahrens effizient ausgeführt werden kann, also eine Lösung der folgenden Aufgabenstellung.

- Input: Die unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ mit den Haupt- bzw. Nebendiagonalelementen $\delta_1, \dots, \delta_n$ bzw. $\gamma_1, \dots, \gamma_{n-1}$ sei gegeben. Ferner ist ein Shift-Parameter $\sigma \in \mathbb{R}$ vorgegeben.
- Output: Die Matrix A wird mit einer orthogonalen ähnlichen (symmetrischen) Tridiagonalmatrix überschrieben, die im wesentlichen mit $A_+ = Q^T A Q$ übereinstimmt, wobei Q der orthogonale Anteil einer QR-Zerlegung von $A - \sigma I$ ist.

Im Prinzip geht man hier ganz ähnlich vor wie beim QR-Doppelschritt. Der Schlüssel für eine effiziente Realisierung ist wieder durch Satz 2.10 gegeben.

- Bestimme eine Givens-Rotation $G_{12} = G_{12}(c_1, s_1)$ mit

$$\begin{pmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{pmatrix} \begin{pmatrix} \delta_1 - \sigma \\ \gamma_1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

und berechne $A := G_{12} A G_{12}^T$.

Die Givens-Rotation G_{12} transformiert die erste Spalte von $A - \sigma I$ in ein Vielfaches des ersten Einheitsvektors. Daher stimmen die erste Spalte von G_{12}^T und die erste Spalte von Q , dem orthogonalen Anteil in einer QR-Zerlegung von $A - \sigma I$, (eventuell bis auf den Faktor -1) überein. Ferner wird die Tridiagonalgestalt der transformierten Matrix $G_{12} A G_{12}^T$ nur in der Position $(3, 1)$ (und symmetrisch dazu in $(1, 3)$)

gestört. Die weitere Idee besteht einfach darin, das die Tridiagonalgestalt störende Element sukzessive durch Ähnlichkeitstransformationen mit Givens-Rotationen von der Position (3, 1) über (4, 2) nach ($n, n-2$) zu vertreiben und in einem letzten Schritt auch noch das Element an der Stelle ($n, n-2$) zu annullieren.

- Für $k = 2, \dots, n-1$:

Bestimme Givens-Rotation $G_{k,k+1} = G_{k,k+1}(c_k, s_k)$ mit

$$\begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,k-1} \\ a_{k+1,k-1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}, \text{ berechne } A := G_{k,k+1} A G_{k,k+1}^T.$$

- Output: Die Ausgangsmatrix A wird mit der orthogonal ähnlichen (symmetrischen) Tridiagonalmatrix $V^T A V$ überschrieben, wobei $V := G_{12}^T \cdots G_{n-1,n}^T$.

Die erste Spalte von $G_{23}^T, \dots, G_{n-1,n}^T$ ist offenbar jeweils der erste Einheitsvektor e_1 , so daß die erste Spalte von V genau die erste Spalte von G_{12}^T ist. Nach Konstruktion stimmt diese aber mit der ersten Spalte von Q , dem orthogonalen Anteil einer QR-Zerlegung von $A - \sigma I$, (eventuell bis auf das Vorzeichen) überein. Aus Satz 2.10 folgt, daß $V^T A V$ und $A_+ = Q^T A Q$ im wesentlichen (d. h. bis auf eine Ähnlichkeitstransformation mit einer orthogonalen Diagonalmatrix) gleich sind, wenn $V^T A V$ unreduziert ist.

Ähnlich wie beim QR-Doppelschritt wollen wir uns den Prozeß, der von der symmetrischen Tridiagonalmatrix A zunächst zu $G_{12} A G_{12}^T$ und dann nach weiteren Ähnlichkeitstransformationen mit Givens-Rotationen $G_{23}, \dots, G_{n-1,n}$ zur Tridiagonalmatrix $V^T A V$ führt, für $n = 6$ verdeutlichen. Wieder bezeichne • ein bei der Transformation festbleibendes und * ein sich veränderndes Element.

$$\begin{array}{ccc} \left(\begin{array}{cccccc} \bullet & \bullet & & & & & \\ \vdots & \vdots & \ddots & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \ddots & \ddots \end{array} \right) & \xrightarrow{G_{12}} & \left(\begin{array}{cccccc} * & * & * & & & & \\ * & * & * & & & & \\ * & * & \bullet & \bullet & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots \end{array} \right) & \xrightarrow{G_{23}} & \left(\begin{array}{cccccc} \bullet & * & & & & & \\ * & * & * & * & & & \\ * & * & \bullet & \bullet & \bullet & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots \end{array} \right) & \xrightarrow{G_{34}} \\ \left(\begin{array}{cccccc} \bullet & \bullet & & & & & \\ \vdots & \vdots & * & & & & \\ * & * & * & * & & & \\ * & * & * & * & & & \\ * & * & \bullet & \bullet & & & \\ & \ddots & \ddots & \ddots & & & \end{array} \right) & \xrightarrow{G_{45}} & \left(\begin{array}{cccccc} \bullet & \bullet & & & & & \\ \vdots & \vdots & \bullet & & & & \\ \bullet & \bullet & * & & & & \\ & \ddots & \ddots & * & * & * & \\ & & * & * & * & * & \\ & & * & * & * & * & \bullet \\ & & & \ddots & \ddots & \ddots & \ddots \end{array} \right) & \xrightarrow{G_{56}} & \left(\begin{array}{cccccc} \bullet & \bullet & & & & & \\ \vdots & \vdots & \bullet & & & & \\ \bullet & \bullet & \bullet & & & & \\ & \ddots & \bullet & \bullet & * & & \\ & & \bullet & \bullet & \bullet & * & \\ & & & \ddots & \ddots & \ddots & \ddots \\ & & & & * & * & * \\ & & & & & * & * \end{array} \right) \end{array}$$

Offenbar ist der Aufwand (Anzahl der Multiplikationen bzw. „flops“) zur Berechnung von $V^T A V$ aus A im wesentlichen proportional zu n .

Nun geben wir noch in Pseudo-Code ein Programm zur Durchführung eines QR-Schrittes für eine symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ bei gegebenem Shift-Parameter $\sigma \in \mathbb{R}$ an (siehe auch G. W. STEWART (1973, S. 372) und H. R. SCHWARZ (1988, S. 280)). Hierbei wird die in 5.2.2 eingeführte Funktion $\text{rot}(\alpha, \beta) = (c, s, \gamma)$

benutzt, die zu vorgegebenen $(\alpha, \beta) \in \mathbb{R}^2$ Konstanten c und s mit $c^2 + s^2 = 1$ sowie γ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

bestimmt.

Im ersten Schritt wird $\alpha := \delta_1 - \sigma$, $\beta := \gamma_1$ gesetzt, δ_1 und γ_1 durch $x := \delta_1$, $y := \gamma_1$ abgespeichert, $\text{rot}(\alpha, \beta) = (c, s, \nu)$ aufgerufen und

$$\begin{pmatrix} \delta_1 & \alpha \\ \alpha & x \end{pmatrix} := \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x & y \\ y & \delta_2 \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

mit

$$\begin{aligned} \delta_1 &:= c^2x + 2csy + s^2\delta_2, \\ \alpha &:= (c^2 - s^2)y + cs(\delta_2 - x), \\ x &:= s^2x - 2csy + c^2\delta_2 \end{aligned}$$

berechnet. Da das erste Diagonalelement im Verlauf der Rechnung nicht mehr verändert wird, ist es mit dem neuen Wert überspeichert worden. Nach dem ersten Schritt ist der obere 3×3 -Block von $G_{12}AG_{12}^T$ durch

$$\begin{pmatrix} \delta_1 & \alpha & \beta \\ \alpha & x & y \\ \beta & y & \delta_3 \end{pmatrix} \quad \text{mit } (\beta \ y) := (0 \ \gamma_2) \begin{pmatrix} c & -s \\ s & c \end{pmatrix} = (s\gamma_2 \ c\gamma_2)$$

gegeben, alle anderen Elemente haben sich nicht verändert. Im nächsten Schritt kann nach dem Aufruf von $\text{rot}(\alpha, \beta) = (c, s, \nu)$ das erste neue Nebendiagonalelement $\gamma_1 := \nu$ gesetzt und entsprechend fortgefahren werden. Damit erhält man das folgende Programm.

- Input: Sei eine unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ mit den Haupt- bzw. Nebendiagonalelementen $\delta_1, \dots, \delta_n$ bzw. $\gamma_1, \dots, \gamma_{n-1}$ gegeben. Ferner ist ein Shift-Parameter $\sigma \in \mathbb{R}$ vorgegeben. Es wird die in 5.2.2 definierte Funktion "rot" benutzt.

- $\alpha := \delta_1 - \sigma$, $\beta := \gamma_1$, $x := \delta_1$, $y := \gamma_1$.

Für $k = 1, \dots, n-1$:

$$(c, s, \nu) := \text{rot}(\alpha, \beta)$$

Falls $k > 1$, dann: $\gamma_{k-1} := \nu$

$$\delta_k := c^2x + 2csy + s^2\delta_{k+1},$$

$$\alpha := (c^2 - s^2)y + cs(\delta_{k+1} - x),$$

$$x := s^2x - 2csy + c^2\delta_{k+1}$$

Falls $k < n-1$, dann: $\beta := s\gamma_{k+1}$, $y := c\gamma_{k+1}$

$$\gamma_{n-1} := \alpha, \quad \delta_n := x$$

- Output: In $\delta_1, \dots, \delta_n$ stehen die Hauptdiagonalelemente und in $\gamma_1, \dots, \gamma_{n-1}$ die Nebendiagonalelemente einer symmetrischen Tridiagonalmatrix, die im wesentlichen mit $A_+ = Q^T A Q$ übereinstimmt, wobei Q der orthogonale Anteil einer QR -Zerlegung von $A - \sigma I$ ist.

Nachdem die Durchführung eines QR -Schrittes für eine symmetrische Tridiagonalmatrix geschildert wurde, muß nun etwas zur Wahl des Shift-Parameters σ gesagt werden. Es werden im wesentlichen zwei Strategien benutzt.

(a) $\sigma := \delta_n$.

(b) σ wird als derjenige Eigenwert von

$$\begin{pmatrix} \delta_{n-1} & \gamma_{n-1} \\ \gamma_{n-1} & \delta_n \end{pmatrix}$$

bestimmt, der näher bei δ_n liegt. Diese Wahl wird als *Wilkinson-Shift* bezeichnet.

Von B. N. PARLETT (1980, S. 149) (siehe auch G. H. GOLUB, C. F. VAN LOAN (1989, S. 423)) wird empfohlen, den Wilkinson-Shift folgendermaßen zu berechnen:

$$d := \frac{\delta_{n-1} - \delta_n}{2}, \quad \sigma := \delta_n - \frac{\operatorname{sign}(d) \gamma_{n-1}^2}{|d| + \sqrt{d^2 + \gamma_{n-1}^2}}.$$

Nun sollen noch einige Bemerkungen zur Konvergenz des QR -Verfahrens bei einer unreduzierten, symmetrischen Tridiagonalmatrix A gemacht werden.

Will man Satz 2.8 anwenden, um hinreichende Bedingungen für die Konvergenz des einfachen QR -Verfahrens (alle Shift-Parameter $\sigma_k = 0$) zu erhalten, so stellt man fest, daß die Voraussetzung (***) automatisch erfüllt ist (siehe Bemerkung im Anschluß an Satz 2.8). Falls A singulär sein sollte, so findet schon nach einem Schritt des ungeshifteten QR -Verfahrens eine Deflation statt, so daß auch diese Voraussetzung überflüssig ist. Beachtet man noch, daß die Eigenwerte von A wegen Satz 3.3 alle (reell und) einfach sind, so erkennt man, daß auch die Voraussetzung (*) erfüllt ist, wenn es kein Paar $(\lambda, -\lambda)$ von Eigenwerten gibt. Unter schwachen Voraussetzungen konvergiert die durch das einfache QR -Verfahren erzeugte Folge $\{A_k\}$ symmetrischer Tridiagonalmatrizen also gegen eine Diagonalmatrix, in der die Eigenwerte der Ausgangsmatrix dem Betrag nach geordnet erscheinen. Wegen seiner i. allg. schlechten Konvergenzeigenschaften sollte man das einfache QR -Verfahren trotzdem nicht anwenden.

Interessanter sind Aussagen über die Konvergenz und die Konvergenzgeschwindigkeit des mit der Strategie (a) oder (b) geshifteten QR -Verfahrens bei symmetrischen Tridiagonalmatrizen. Die grundlegende Arbeit hierzu stammt von J. H. WILKINSON (1968) (siehe auch C. L. LAWSON, R. J. HANSON (1974, S. 240 ff.) für eine ausführlichere Darstellung). Wir wollen hier nur die wesentlichen Ergebnisse zitieren und müssen dabei auf Beweise verzichten.

Das QR -Verfahren erzeugt eine Folge $\{A_k\}$ symmetrischer Tridiagonalmatrizen A_k mit den Hauptdiagonalelementen $\delta_1^{(k)}, \dots, \delta_n^{(k)}$ und den Nebendiagonalelementen $\gamma_1^{(k)}, \dots, \gamma_{n-1}^{(k)}$. Man spricht von *globaler Konvergenz* des QR -Verfahrens, wenn $\lim_{k \rightarrow \infty} \gamma_{n-1}^{(k)} = 0$, und sagt, es sei *quadratisch bzw. kubisch konvergent*, wenn eine Konstante $c > 0$ existiert derart, daß $|\gamma_{n-1}^{(k+1)}| \leq c |\gamma_{n-1}^{(k)}|^2$ bzw. $|\gamma_{n-1}^{(k+1)}| \leq c |\gamma_{n-1}^{(k)}|^3$ für alle hinreichend großen k . Es liegt nahe, diesen modifizierten Konvergenzbegriff zu benutzen, da man ja an (schneller) Konvergenz der Folge $\{\gamma_{n-1}^{(k)}\}$ interessiert ist, um (möglichst schnell) durch Streichen der letzten Zeile und letzten Spalte (setze $n := n - 1$) zu einem in der Dimension reduzierten Problem zu gelangen.

Für die Shift-Strategie (a) zeigt Wilkinson, daß $|\gamma_{n-1}^{(k+1)}| \leq |\gamma_{n-1}^{(k)}|$ für alle k . Hieraus folgt die Konvergenz der Folge $\{|\gamma_{n-1}^{(k)}|\}$. Ist dieser Limes L gleich Null, liegt also globale Konvergenz vor, so ist die Konvergenz sogar kubisch. Ist dagegen $L > 0$ (dann liegt keine Konvergenz des Verfahrens im obigen Sinne vor), so konvergiert wenigstens die Folge $\{\gamma_{n-2}^{(k)}\}$ (eventuell langsam) gegen Null. In diesem Falle kann man daher schließlich die letzten beiden Zeilen und Spalten streichen.

Für das QR -Verfahren mit der Shift-Strategie (b) zeigt Wilkinson zunächst die globale und dann die quadratische Konvergenz. Von Ausnahmen abgesehen kann sogar die kubische Konvergenz nachgewiesen werden.

5.3.4 Die Berechnung der Singulärwertzerlegung

Auf die Singulärwertzerlegung einer Matrix sind wir in Unterabschnitt 1.6.2 eingegangen. Wir betrachten im folgenden Matrizen $A \in \mathbb{R}^{m \times n}$, deren Zeilenzahl m nicht kleiner als die Spaltenzahl n ist, was der bei linearen Ausgleichsproblemen (siehe Abschnitt 1.6) interessanter Fall ist. Andernfalls ersetze man A durch A^T . Eine Definition der Singulärwertzerlegung sowie eine Existenzaussage sind im nächsten Satz zusammengestellt.

Satz 3.6 Sei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $r := \text{Rang}(A)$. Dann existieren Matrizen $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ mit $U^T U = I$ (die Spalten von U sind orthonormiert) und $V^T V = I$ (die Matrix V ist orthogonal) sowie reelle Zahlen $\sigma_1, \dots, \sigma_n$ mit

$$\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$$

derart, daß die sogenannte *Singulärwertzerlegung*

$$A = U \operatorname{diag}(\sigma_1, \dots, \sigma_n) V^T$$

gilt. Hierbei sind die singulären Werte $\sigma_1, \dots, \sigma_r$ positive Quadratwurzeln aus den positiven Eigenwerten von $A^T A$ bzw. AA^T (diese stimmen überein!). Die Spalten von U bilden ein Orthonormalsystem von Eigenvektoren zu den n größten Eigenwerten von $AA^T \in \mathbb{R}^{m \times m}$ und die Spalten von V bilden ein Orthonormalsystem von Eigenvektoren zu den Eigenwerten von $A^T A \in \mathbb{R}^{n \times n}$.

Bemerkung: Gegenüber Satz 6.2 in Abschnitt 1.6 ist die Aussage bzw. sind die Bezeichnungen in Satz 3.6 geringfügig geändert worden, die Gleichwertigkeit der Aussagen ist aber offensichtlich. \square

Natürlich könnte man die singulären Werte von $A \in \mathbb{R}^{m \times n}$ dadurch erhalten, daß man die symmetrische Matrix $A^T A \in \mathbb{R}^{n \times n}$ bildet, mit einem Verfahren zur Eigenwertbestimmung symmetrischer Matrizen (QR -Verfahren oder Jacobi-Verfahren) die positiven Eigenwerte dieser Matrix berechnet und hieraus die positiven Quadratwurzeln nimmt. Ähnlich wie bei einem linearen Ausgleichsproblem

$$(LA) \quad \text{Minimiere } \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

bei dem es i. allg. keine gute Idee ist, die Koeffizientenmatrix $A^T A$ der Normalgleichungen explizit zu bilden, um dann die Lösung bzw. die Lösungen von (LA) aus $A^T A x = A^T b$ zu erhalten, sollte man auch hier auf die Berechnung von $A^T A$ verzichten.

Wir werden in diesem Unterabschnitt das Verfahren von Golub-Reinsch (siehe J. H. WILKINSON, C. REINSCH (1971, S. 134 ff.), dort findet man auch ein Algol-Programm) schildern.

Beim QR -Verfahren zur Berechnung der Eigenwerte (und Eigenvektoren) einer Matrix macht man zu Beginn einen Reduktionsschritt und transformiert die gegebene Matrix in eine ähnliche Hessenberg-Matrix (bzw. Tridiagonalmatrix). Die Eigenwerte werden hierbei nicht verändert, ferner lassen sich die Eigenvektoren der gegebenen Matrix aus denen der reduzierten Matrix mit Hilfe der Transformationsmatrix leicht berechnen. Ganz ähnlich geht man auch bei der Berechnung einer Singulärwertzerlegung vor. Grundlage ist das folgende Lemma.

Lemma 3.7 Sei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ gegeben, ferner seien

$$P = (p_1 \ \cdots \ p_m) \in \mathbb{R}^{m \times m}, \quad Q = (q_1 \ \cdots \ q_n) \in \mathbb{R}^{n \times n}$$

orthogonal und

$$P^T A Q = \begin{pmatrix} B \\ 0 \end{pmatrix} \quad \left. \begin{array}{c} \} & n \\ \} & m-n \end{array} \right.$$

Dann gilt:

1. $A \in \mathbb{R}^{m \times n}$ und $B \in \mathbb{R}^{n \times n}$ haben dieselben singulären Werte.
2. Sei $B = U_B \text{diag}(\sigma_1, \dots, \sigma_n) V_B^T$ mit (orthogonalen) $U_B \in \mathbb{R}^{n \times n}$, $V_B \in \mathbb{R}^{n \times n}$ eine Singulärwertzerlegung von B . Dann ist durch

$$A = U \text{diag}(\sigma_1, \dots, \sigma_n) V^T \quad \text{mit} \quad U := (p_1 \ \cdots \ p_n) U_B, \quad V := Q V_B$$

eine Singulärwertzerlegung von A gegeben.

Beweis: Der Beweis erfolgt durch einfaches Nachrechnen. So ist z. B.

$$B^T B = \begin{pmatrix} B \\ 0 \end{pmatrix}^T \begin{pmatrix} B \\ 0 \end{pmatrix} = (P^T A Q)^T (P^T A Q) = Q^T A^T P P^T A Q = Q^T (A^T A) Q,$$

also ist $B^T B$ orthogonal ähnlich zu $A^T A$. Daher stimmen die Eigenwerte von $A^T A$ und $B^T B$ und damit auch die singulären Werte von A und B überein. Wegen

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T = (p_1 \ \cdots \ p_n) B Q^T = (p_1 \ \cdots \ p_n) U_B \text{diag}(\sigma_1, \dots, \sigma_n) (Q V_B)^T$$

folgt der Rest der Behauptungen. \square

Genauer sieht der Reduktionsschritt bei der Berechnung einer Singulärwertzerlegung folgendermaßen aus:

- Man bestimme Householder-Matrizen

$$P_1, \dots, P_n \in \mathbb{R}^{m \times m}, \quad Q_1, \dots, Q_{n-2} \in \mathbb{R}^{n \times n}$$

derart, daß $P_n \cdots P_1 A Q_1 \cdots Q_{n-2}$ eine *obere Bidiagonalmatrix* ist, also die Form

$$P_n \cdots P_1 A Q_1 \cdots Q_{n-2} = \begin{pmatrix} d_1 & f_2 & 0 & \cdots & 0 \\ d_2 & f_3 & \ddots & & \vdots \\ \ddots & \ddots & 0 & & \\ & & \ddots & f_n & \\ & & & & d_n \\ \hline 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & \cdots & 0 \end{pmatrix} = \begin{pmatrix} B \\ 0 \end{pmatrix} \quad \begin{array}{l} \} n \\ \} m-n \end{array}$$

hat.

Die Idee zu diesem Reduktionsschritt und seine Durchführung stammen von G. H. GOLUB, W. KAHAN (1965). Grundlegend ist wieder, daß man einen vom Nullvektor verschiedenen Vektor durch Multiplikation mit einer geeigneten Householder-Matrix in ein Vielfaches des ersten Einheitsvektors überführen kann (siehe Lemma 2.2). Im Prinzip sieht dieses *Bidiagonalisierungsverfahren* folgendermaßen aus:

- Input: Gegeben $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$.
- Für $k = 1, \dots, n$:

Falls $(a_{kk}, \dots, a_{mk})^T \neq 0$, dann:

Bestimme Householder-Matrix $\bar{P}_k \in \mathbb{R}^{(m-k+1) \times (m-k+1)}$ mit

$$\bar{P}_k (a_{kk}, \dots, a_{mk})^T = (*, 0, \dots, 0)^T$$

Setze $P_k := \text{diag}(I_{k-1}, \bar{P}_k)$, berechne $A := P_k A$

Andernfalls: Setze $P_k := I$

Falls $k \leq n-2$, dann:

Falls $(a_{k,k+1}, \dots, a_{kn})^T \neq 0$, dann:

Bestimme Householder-Matrix $\bar{Q}_k \in \mathbb{R}^{(n-k+1) \times (n-k+1)}$ mit

$$(a_{k,k+1}, \dots, a_{kn}) \bar{Q}_k = (*, 0, \dots, 0)$$

Setze $Q_k := \text{diag}(I_k, \bar{Q}_k)$, berechne $A := A Q_k$

Andernfalls: Setze $Q_k := I$

- Output: A wird überschrieben mit einer oberen Bidiagonalmatrix.

In einer Implementation des Verfahrens wird man das $m \times n$ -Feld A dazu benutzen, um die relevanten Informationen über die Matrizen P_1, \dots, P_n und Q_1, \dots, Q_{n-2} aufzunehmen, während die Haupt- und Superdiagonalelemente der Bidiagonalmatrix B in Feldern d und f gespeichert werden. Anschließend kann man die ersten n Spalten von $P_1 \cdots P_n$ und das Produkt $Q_1 \cdots Q_{n-2}$ bilden (jedenfalls dann, wenn nicht nur die singulären Werte, sondern die volle Singulärwertzerlegung berechnet werden soll) und diese in $m \times n$ - bzw. $n \times n$ -Feldern U bzw. V speichern. Ist dies geschehen, so hat man zu der Ausgangsmatrix $A \in \mathbb{R}^{m \times n}$ Matrizen $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ mit $U^T U = I$, $V^T V = I$ sowie eine obere Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ mit $A = UBV^T$ bestimmt. Wäre B sogar eine Diagonalmatrix, so hätten wir die gesuchte Singulärwertzerlegung von A schon gewonnen.

An einer 4×3 -Matrix veranschaulichen wir uns den Prozeß, eine gegebene Matrix durch Multiplikation mit Householder-Matrizen von links und rechts auf obere Bidiagonal-Gestalt zu transformieren. Festbleibende Elemente werden durch \bullet , sich verändernde durch $*$ gekennzeichnet.

$$\begin{pmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{P_1} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \xrightarrow{Q_1} \begin{pmatrix} \bullet & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \xrightarrow{P_2} \begin{pmatrix} \bullet & \bullet & * \\ * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \xrightarrow{P_3} \begin{pmatrix} \bullet & \bullet & * \\ * & \bullet & \bullet \\ * & * & * \\ 0 & 0 & 0 \end{pmatrix}$$

Wir haben jetzt das Problem darauf reduziert, eine Singulärwertzerlegung der oberen Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ zu berechnen. Die Hauptdiagonalelemente von B seien d_1, \dots, d_n , die Superdiagonalelemente seien f_2, \dots, f_n . Die Matrix $C := B^T B$ ist eine (symmetrische) $n \times n$ -Tridiagonalmatrix, ihre Diagonal- bzw. Nebendiagonalelemente sind gegeben durch

$$c_{11} = d_1^2, \quad c_{ii} = d_i^2 + f_i^2 \quad (i = 2, \dots, n), \quad c_{i,i+1} = c_{i+1,i} = d_i f_{i+1} \quad (i = 1, \dots, n-1).$$

Daher ist $B^T B$ genau dann unreduziert, wenn $d_1 \cdots d_{n-1} \neq 0$ und $f_2 \cdots f_n \neq 0$. Ist dies nicht der Fall, so kann die Berechnung einer Singulärwertzerlegung der oberen Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ auf die von zwei niederdimensionalen oberen Bidiagonalmatrizen B_1 und B_2 zurückgeführt werden. Grundlage hierfür ist das folgende Lemma.

Lemma 3.8 *Die Matrix $B \in \mathbb{R}^{n \times n}$ sei eine obere Bidiagonalmatrix mit Hauptdiagonalelementen d_1, \dots, d_n sowie Superdiagonalelementen f_2, \dots, f_n .*

1. Sei $f_{k+1} = 0$ für ein $k \in \{1, \dots, n-1\}$, so daß B zerfällt:

$$B = \left(\begin{array}{c|c} B_1 & 0 \\ \hline 0 & B_2 \end{array} \right) \quad \underbrace{\quad}_{k} \quad \underbrace{\quad}_{n-k}$$

Ist dann

$$B_1 = U_1 \operatorname{diag}(\sigma_1, \dots, \sigma_k) V_1^T, \quad B_2 = U_2 \operatorname{diag}(\sigma_{k+1}, \dots, \sigma_n) V_2^T$$

eine Singulärwertzerlegung von B_1 bzw. B_2 , so ist

$$B = \left(\begin{array}{c|c} U_1 & 0 \\ \hline 0 & U_2 \end{array} \right) \text{diag}(\sigma_1, \dots, \sigma_n) \left(\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array} \right)^T$$

eine Singulärwertzerlegung von B .

2. Ist $d_k = 0$ und $f_{k+1} \neq 0$ für ein $k \in \{1, \dots, n-1\}$, so kann man $n-k$ Givens-Rotationen $G_{k,k+1}, \dots, G_{k,n}$ bestimmen derart, daß $B_+ := G_{k,n} \cdots G_{k,k+1} B$ eine obere Bidiagonalmatrix mit $(B_+)_{k,k+1} = 0$ ist.

Beweis: Den ersten Teil des Lemmas beweist man durch einfaches Nachrechnen. Die Idee für den Beweis des zweiten Teiles besteht darin, das störende Superdiagonalelement in der k -ten Zeile sukzessive in der Zeile nach rechts zu schieben und schließlich aus der Matrix zu drängen. Statt eines formalen Beweises machen wir uns die Strategie anhand einer 5×5 -Matrix klar, bei der das zweite Diagonalelement verschwindet. Wieder werden bei einer Transformation festbleibende Elemente durch \bullet , sich verändernde durch $*$ gekennzeichnet.

$$B = \begin{pmatrix} \bullet & \bullet & & & \\ 0 & \bullet & & & \\ \bullet & \bullet & & & \\ \bullet & \bullet & & & \\ \bullet & \bullet & & & \end{pmatrix} \xrightarrow{G_{23}} \begin{pmatrix} \bullet & \bullet & & & \\ 0 & 0 & * & & \\ * & * & & & \\ \bullet & \bullet & & & \\ \bullet & \bullet & & & \end{pmatrix} \xrightarrow{G_{24}} \begin{pmatrix} \bullet & \bullet & 0 & 0 & * \\ 0 & 0 & \bullet & \bullet & \\ \bullet & \bullet & * & * & \\ \bullet & \bullet & & & \end{pmatrix} \xrightarrow{G_{25}}$$

und man erhält schließlich

$$B_+ = \begin{pmatrix} \bullet & \bullet & & & \\ 0 & & & & \\ \hline & \bullet & \bullet & & \\ & \bullet & \bullet & & \\ & & & * & \end{pmatrix}.$$

Ein exakter Beweis ist offensichtlich, siehe auch den folgenden Algorithmus. \square

Jetzt geben wir den eben skizzierten Algorithmus genauer an.

- Input: Sei $B \in \mathbb{R}^{n \times n}$ eine obere Bidiagonalmatrix mit d_1, \dots, d_n und f_2, \dots, f_n als Haupt- bzw. Nebendiagonalelementen. Für ein $k \in \{1, \dots, n-1\}$ sei $d_k = 0$, $f_{k+1} \neq 0$. Gegeben sei ferner eine Matrix $U \in \mathbb{R}^{m \times n}$. Mit der Ausgangsmatrix $A \in \mathbb{R}^{m \times n}$ sei z. B. $A = UBV^T$.
- Setze $c := 0$, $s := 1$
- Für $j = k+1, \dots, n$:

$$f := sf_j, \quad f_j := cf_j$$

$$c := d_j/(f^2 + d_j^2)^{1/2}, \quad s := -f/(f^2 + d_j^2)^{1/2}, \quad d_j := (f^2 + d_j^2)^{1/2}$$
- Für $i = 1, \dots, m$:

$$(u_{ik}, u_{ij}) := (u_{ik}, u_{ij}) \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

- Output: Die Felder d und f enthalten die Haupt- bzw. Nebendiagonalelemente einer Bidiagonalmatrix $B_+ := G_{k,n} \cdots G_{k,k+1} B$, die also aus B durch Multiplikation von links mit Givens-Rotationen $G_{k,k+1}, \dots, G_{k,n}$ hervorgeht. Hierbei ist $f_{k+1} = 0$, das k -te Superdiagonalelement von B_+ verschwindet also, so daß B_+ zerfällt. Ferner wird $U \in \mathbb{R}^{m \times n}$ überschrieben mit $U_+ := U G_{k,k+1}^T \cdots G_{k,n}^T$. War vorher $A = UBV^T$, so ist nach Abschluß $A = U_+ B_+ V^T$.

Wir haben jetzt das Problem darauf reduziert, eine Singulärwertzerlegung der oberen Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ zu bestimmen, für die $C := B^T B$ unreduziert ist. Wir geben gleich einen Schritt des Verfahrens von Golub-Reinsch an, in dem eine obere Bidiagonalmatrix B_+ berechnet wird, von der man sich erhofft, daß das letzte Nebendiagonalelement wesentlich kleiner als das entsprechende in B ist. Nach wenigen Schritten sollte dann eine Deflation möglich sein, also der Übergang zu einer $(n-1)$ -dimensionalen Aufgabe. Grundlage ist wieder Satz 2.10, genau wie beim QR-Doppelschritt und dem QR-Verfahren für symmetrische Tridiagonalmatrizen.

- Input: Sei $B \in \mathbb{R}^{n \times n}$ eine obere Bidiagonalmatrix mit den Haupt- bzw. Nebendiagonalelementen d_1, \dots, d_n bzw. f_2, \dots, f_n . Es sei $d_1 \cdots d_{n-1} \neq 0$ und $f_2 \cdots f_n \neq 0$, d. h. $C := B^T B$ ist eine unreduzierte, symmetrische Tridiagonalmatrix. Weiter nehmen wir an, jedenfalls wenn eine volle Singulärwertzerlegung der Ausgangsmatrix A zu berechnen ist, daß Matrizen $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ mit $U^T U = I$, $V^T V = I$ und $A = UBV^T$ gegeben sind.
- Bestimme den Shift-Parameter $\sigma \in \mathbb{R}$ als den Eigenwert des unteren 2×2 -Blocks

$$\begin{pmatrix} d_{n-1}^2 + f_{n-1}^2 & d_{n-1}f_n \\ d_{n-1}f_n & d_n^2 + f_n^2 \end{pmatrix}$$

in C , der $d_n^2 + f_n^2$ am nächsten liegt.

- Bestimme Givens-Rotation $T_{12} = T_{12}(c, s)$ mit $\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} d_1^2 - \sigma \\ d_1 f_2 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$.

Berechne $B := BT_{12}^T$, $V := VT_{12}^T$.

Nach Konstruktion transformiert T_{12} die erste Spalte von $B^T B$ in ein Vielfaches des ersten Einheitsvektors. Daher stimmt die erste Spalte von T_{12}^T mit der ersten Spalte von Q , dem orthogonalen Anteil in einer QR-Zerlegung von $B^T B - \sigma I$, (eventuell bis auf den Faktor -1) überein.

In der transformierten Matrix BT_{12}^T verändert sich nur der obere 2×2 -Block, dieser ist nach der Transformation gegeben durch

$$\begin{pmatrix} d_1 & f_2 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} = \begin{pmatrix} d_1 c + f_2 s & -d_1 s + f_2 c \\ d_2 s & d_2 c \end{pmatrix}.$$

In BT_{12}^T wird die obere Bidiagonal-Gestalt also nur durch das Element in der Position $(2, 1)$ gestört. Die Idee besteht jetzt darin, dieses Element durch Multiplikation von links mit einer geeigneten Givens-Rotation S_{12} zu annullieren. Die neuen ersten

beiden Zeilen sind eine Linearkombination der alten, so daß nach der Transformation ein die Bidiagonal-Gestalt störendes Element in der Position (1, 3) steht. Dieses Element wiederum wird durch Multiplikation von *rechts* mit einer Givens-Rotation T_{23}^T annulliert und auf die Position (3, 2) gedrängt. Nun ist klar, daß man durch abwechselnde Multiplikation von links und rechts mit geeigneten Givens-Rotationen das die Bidiagonal-Gestalt störende Element „in Rössel-Sprüngen“ aus der Matrix drängen kann. Genauer sehen die weiteren Schritte folgendermaßen aus:

- Für $k = 1, \dots, n - 1$:

Bestimme Givens-Rotation $S_{k,k+1} = S_{k,k+1}(c, s)$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{kk} \\ b_{k+1,k} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

Berechne $B := S_{k,k+1}B$, $U := US_{k,k+1}^T$.

Falls $k < n - 1$, dann:

Bestimme Givens-Rotation $T_{k+1,k+2} = T_{k+1,k+2}(c, s)$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{k,k+1} \\ b_{k,k+2} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

Berechne $B := BT_{k+1,k+2}^T$, $V := VT_{k+1,k+2}^T$.

- Output: Die Ausgangsmatrix $B \in \mathbb{R}^{n \times n}$ ist überschrieben mit der oberen Bidiagonalmatrix

$$B_+ := S_{n-1,n} \cdots S_{12}BT_{12}^T \cdots T_{n-1,n}^T.$$

Ferner sind die Matrizen $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ überschrieben mit

$$U_+ := US_{12}^T \cdots S_{n-1,n}^T, \quad V_+ := VT_{12}^T \cdots T_{n-1,n}^T.$$

Daher ist $U_+B_+V_+^T = UBV^T$.

Mit den eben eingeführten Bezeichnungen ist

$$B_+^TB_+ = (T_{n-1,n} \cdots T_{12})(B^TB)(T_{n-1,n} \cdots T_{12})^T.$$

Wegen

$$(T_{n-1,n} \cdots T_{12})^Te_1 = T_{12}^T \cdots T_{n-1,n}^Te_1 = T_{12}^Te_1$$

ist die erste Spalte von $(T_{n-1,n} \cdots T_{12})^T$ genau die erste Spalte von T_{12}^T , die wiederum nach Konstruktion mit der ersten Spalte von Q , dem orthogonalen Anteil in einer QR -Zerlegung von $B^TB - \sigma I$, (eventuell bis auf einen Faktor -1) übereinstimmt. Nach Satz 2.10 stimmen daher $B_+^TB_+$ und Q^TB^TBQ sowie Q und $(T_{n-1,n} \cdots T_{12})^T$ im wesentlichen überein, wenn $B_+^TB_+$ unreduziert ist (andernfalls zerfällt B_+). Indem wir alleine mit der Bidiagonalmatrix B (und nicht etwa mit der Tridiagonalmatrix B^TB) gearbeitet haben, ist durch das obige Verfahren ein impliziter QR -Schritt mit Shift σ auf B^TB durchgeführt worden. Daher besteht Grund zur Hoffnung, daß zumindestens das letzte Superdiagonalelement in der transformierten Bidiagonalmatrix B_+ wesentlich kleiner ist als das entsprechende in der Ausgangsmatrix B .

Anhand einer 4×4 -Bidiagonalmatrix veranschaulichen wir uns obigen Algorithmus. Abwechselnd wird mit $T_{k,k+1}^T$ von rechts, mit $S_{k,k+1}$ von links multipliziert. Bei einer Transformation festbleibende Elemente werden mit \bullet , sich verändernde mit $*$ bezeichnet.

$$\begin{pmatrix} \bullet & \bullet & & \\ \bullet & \bullet & \bullet & \\ \bullet & & \bullet & \\ \bullet & & & \bullet \end{pmatrix} \xrightarrow{T_{12}} \begin{pmatrix} * & * & & \\ * & * & \bullet & \\ \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{S_{12}} \begin{pmatrix} * & * & * & \\ * & * & * & \\ \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{T_{23}} \begin{pmatrix} \bullet & * & & \\ * & * & * & \\ * & * & \bullet & \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \\ \xrightarrow{S_{23}} \begin{pmatrix} \bullet & * & * & * \\ * & * & * & * \\ * & * & * & * \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{T_{34}} \begin{pmatrix} \bullet & * & & \\ \bullet & * & * & \\ * & * & * & \\ * & * & * & \bullet \end{pmatrix} \xrightarrow{S_{34}} \begin{pmatrix} \bullet & * & & \\ \bullet & * & * & \\ \bullet & * & * & * \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix}$$

Einer Implementation des obigen Algorithmus sollte nun nichts mehr im Wege stehen. Hierbei kann die in Unterabschnitt 5.2.2 eingeführte Funktion ‘‘rot’’ benutzen, die zu gegebenen $\alpha, \beta \in \mathbb{R}$ Konstanten $(c, s, \gamma) := \text{rot}(\alpha, \beta)$ mit

$$c^2 + s^2 = 1, \quad \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

berechnet. Ferner berücksichtige man, daß die im Verlauf des Algorithmus transformierte Matrix B immer nur in (höchstens) einem Element von der Bidiagonal-Gestalt abweicht. Nach jedem Schritt $B \longrightarrow B_+$ sollte getestet werden, ob die neue Bidiagonalmatrix B_+ zerfällt, ob also z. B. das letzte Superdiagonalelement in B_+ betragsmäßig kleiner oder gleich einer vorgegebenen Toleranz ist.

Bemerkung: Es gibt verschiedene, gleichwertige Möglichkeiten, die *Pseudoinverse* $A^+ \in \mathbb{R}^{n \times m}$ einer Matrix $A \in \mathbb{R}^{m \times n}$ zu definieren (siehe Unterabschnitt 1.6.3 und dort insbesondere Satz 6.5). Eine Möglichkeit besteht darin, A^+ als eindeutige Lösung der Moore-Penrose-Gleichungen

$$(*) \quad AX = (AX)^T, \quad XA = (XA)^T, \quad AXA = A, \quad XAX = X$$

zu erklären. Ist $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $r := \text{Rang}(A)$, ist ferner

$$A = U \text{diag}(\sigma_1, \dots, \sigma_n) V^T$$

mit $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ und $U^T U = V^T V = I$ sowie

$$\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n$$

eine Singulärwertzerlegung von A , so genügt

$$A^+ := V \text{diag}(1/\sigma_1, \dots, 1/\sigma_r, 0, \dots, 0) U^T$$

den Moore-Penrose-Gleichungen (*), ist also in der Tat die Pseudoinverse von A . Obiges Verfahren zur Berechnung einer Singulärwertzerlegung von A liefert also auf

einfache Weise auch die Pseudoinverse A^+ von A . Der Zusammenhang lässt sich auch folgendermaßen ausdrücken: Sind $u_1, \dots, u_n \in \mathbb{R}^m$ die Spalten von U und entsprechend $v_1, \dots, v_n \in \mathbb{R}^n$ die Spalten von V , sind ferner $\sigma_1, \dots, \sigma_r$ die singulären Werte von A , so gelten die Beziehungen

$$A = \sum_{k=1}^r \sigma_k u_k v_k^T, \quad A^+ = \sum_{k=1}^r \frac{1}{\sigma_k} v_k u_k^T.$$

Ist ein lineares Ausgleichsproblem

$$(LA) \quad \text{Minimiere } \|Ax - b\|_2, \quad x \in \mathbb{R}^n$$

mit $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $m \geq n$ gegeben, so ist dieses bekanntlich (siehe z. B. Satz 6.1 in Abschnitt 1.6) stets lösbar und eindeutig lösbar genau dann, wenn A den vollen Rang besitzt bzw. $\text{Rang}(A) = n$ gilt. Ferner besitzt (LA) stets eine eindeutige Lösung x_{LA} minimaler euklidischer Norm, welche durch $x_{LA} = A^+b$ gegeben ist. Berücksichtigt man obige Darstellung der Pseudoinversen A^+ , so erhält man

$$x_{LA} = \sum_{k=1}^r \frac{u_k^T b}{\sigma_k} v_k.$$

Durch die Berechnung einer Singulärwertzerlegung können daher auch lineare Ausgleichsprobleme mit Rangdefizit (d. h. die Matrix $A \in \mathbb{R}^{m \times n}$ hat nicht den vollen Rang n) effizient gelöst werden. \square

Aufgaben

1. Man programmiere das zyklische Jacobi-Verfahren, kombiniert mit einer einfachen Schwellenmethode, zur Berechnung aller Eigenwerte und (eventuell) aller Eigenvektoren einer symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$.

Hinweis: Als Input-Parameter wähle man eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ (von der nur die obere Hälfte benutzt und verändert wird) und einen boolean-Ausdruck *eivec*, der für *eivec=true* aussagt, daß auch die Eigenvektoren zu berechnen sind. Ferner sei ein $\epsilon > 0$ vorgegeben. Man erzeuge durch das Jacobi-Verfahren eine Folge zu A orthogonal ähnlicher Matrizen A_k und breche das Verfahren ab, wenn $N(A_k) < \epsilon^2$, wobei $N(A)$ die Summe der quadrierten Außendiagonalelemente von A bedeutet. Im Output enthalte A in den Diagonalen die berechneten Eigenwerte, ferner seien (für *eivec=true*) in den Spalten von V ein zugehöriges System von Eigenvektoren enthalten. Hierzu setze man am Anfang $V := I$ (für *eivec=true*), definiere eine Unterroutine *rotate*(p, q), in der zu einem vorgegebenem Indexpaar (p, q) mit $1 \leq p < q \leq n$ die das Element a_{pq} annullierende Givens-Rotation G_{pq} sowie $A := G_{pq}^T A G_{pq}$ und (für *eivec=true*) $V := V G_{pq}$ berechnet werden. Man durchlaufe die Außendiagonalelemente (p, q) zeilenweise zyklisch und rotiere nur, wenn $|a_{pq}| \geq \epsilon^2$. Der wesentliche Teil des Programms könnte damit folgendermaßen aussehen:

- Solange $N(A) > \epsilon^2$:

Für $p = 1, \dots, n-1$:

Für $q = p+1, \dots, n$:

Falls $|a_{pq}| \geq \epsilon^2$, dann: *rotate*(p, q)

Beim Abbruch wäre dann $N(A) \leq \epsilon^2$, d. h. die Eigenwerte von A wären (bei exakter Rechnung) mit der absoluten Genauigkeit ϵ bestimmt. Für ein ausgefeiltes Programm sei auf J. H. WILKINSON, C. REINSCH (1971, S. 202 ff.) verwiesen.

Zur Kontrolle wende man das Programm auf die Matrix (siehe J. H. WILKINSON, C. REINSCH (1971, S. 223))

$$A = \begin{pmatrix} 10 & 1 & 2 & 3 & 4 \\ 1 & 9 & -1 & 2 & -3 \\ 2 & -1 & 7 & 3 & -5 \\ 3 & 2 & 3 & 12 & -1 \\ 4 & -3 & -5 & -1 & 15 \end{pmatrix} \quad \text{mit den Eigenwerten} \quad \begin{array}{ll} \lambda_1 & \approx 19.1754202773 \\ \lambda_2 & \approx 15.8089207645 \\ \lambda_3 & \approx 9.36555492016 \\ \lambda_4 & \approx 6.99483783064 \\ \lambda_5 & \approx 1.65526620775 \end{array}$$

an. Als weiteren Test berechne man die Eigenwerte und ein Orthonormalsystem von Eigenvektoren der Matrix (siehe R. T. GREGORY, D. L. KARNEY (1969, S. 55))

$$A = \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix} \quad \text{mit den Eigenwerten} \quad \begin{array}{ll} \lambda_1 & = 10 \\ \lambda_2 & = 5 \\ \lambda_3 & = 2 \\ \lambda_4 & = 1 \end{array}$$

und

$$u_1 = \frac{1}{\sqrt{10}} \begin{pmatrix} 2 \\ 2 \\ 1 \\ 1 \end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} -1 \\ -1 \\ 2 \\ 2 \end{pmatrix}, \quad u_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \quad u_4 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

als zugehörigen normierten Eigenvektoren. Als weiteren Test erprobe man das Programm schließlich noch an der folgenden Matrix, die drei Paare sehr dicht beieinanderliegender Eigenwerte besitzt (siehe R. T. GREGORY, D. L. KARNEY (1969, S. 60)):

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 & 1 & 2 \\ 2 & 4 & 5 & -1 & 0 & 3 \\ 3 & 5 & 6 & -2 & -3 & 0 \\ 0 & -1 & -2 & 1 & 2 & 3 \\ 1 & 0 & -3 & 2 & 4 & 5 \\ 2 & 3 & 0 & 3 & 5 & 6 \end{pmatrix} \quad \text{mit} \quad \begin{array}{ll} \lambda_1 & \approx 12.41133643 \\ \lambda_2 & \approx 12.41133642 \\ \lambda_3 & \approx 0.2849864395 \\ \lambda_4 & \approx 0.2849864365 \\ \lambda_5 & \approx -1.696322849 \\ \lambda_6 & \approx -1.696322851 \end{array}$$

2. Gegeben sei eine unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ mit den Haupt- bzw. Nebendiagonalelementen $\delta_1, \dots, \delta_n$ bzw. $\gamma_1, \dots, \gamma_{n-1}$. Wegen Satz 3.3 hat A nur einfache Eigenwerte, diese seien gemäß $\lambda_1 > \dots > \lambda_n$ angeordnet. Man programmiere das Bisektionsverfahren zur Berechnung des j -ten Eigenwertes der Matrix A .

Hinweis: Als Input-Parameter nehme man $\delta_1, \dots, \delta_n$ sowie $\gamma_1, \dots, \gamma_{n-1}$, $j \in \{1, \dots, n\}$ sowie ein $\epsilon > 0$. Man starte das Bisektionsverfahren mit dem durch den Satz von Gerschgorin gewonnenen Ausgangsintervall $[a_{\min}, b_{\max}]$ (siehe Bemerkung am Schluß von Unterabschnitt 5.3.2). Das Bisektionsverfahren liefert eine Folge von Intervallen $[a_k, b_k]$. Man stoppe das Verfahren und gebe $(a_k + b_k)/2$ als Näherung für den gesuchten Eigenwert aus, sobald $b_k - a_k \leq 2 \text{ macheps}(|a_k| + |b_k|) + \epsilon$, wobei macheps die kleinste positive Zahl auf dem verwendeten Rechner ist, für die $1 + \text{macheps} > 1$. Dieses

Abbruchkriterium entspricht dem bei J. H. WILKINSON, C. REINSCH (1971, S. 249 ff.) verwendeten. Hier wird auch vorgeschlagen, $q_1(\xi), \dots, q_n(\xi)$ rekursiv aus

$$q_1(\xi) := \delta_1 - \xi, \quad q_i(\xi) := (\delta_i - \xi) - \gamma_{i-1}^2 / q_{i-1}(\xi), \quad i = 2, \dots, n,$$

zu berechnen (es ist $q_i(\xi) = p_i(\xi)/p_{i-1}(\xi)$), wobei $q_i(\xi) := (\delta_i - \xi) - |\gamma_{i-1}| / \text{macheps}$ für $q_{i-1}(\xi) = 0$ gesetzt wird, und $N_n(\xi)$ als Anzahl der positiven $q_i(\xi)$ zu nehmen.

Man teste das Programm an den Matrizen aus Aufgabe 1 (nachdem diese auf symmetrische Tridiagonalgestalt transformiert sind) und berechne alle Eigenwerte dieser Matrizen.

3. Man programmiere das QR -Verfahren zur Berechnung aller Eigenwerte einer symmetrischen Tridiagonalmatrix mit den Hauptdiagonalelementen $\delta_1, \dots, \delta_n$ und den Nebendiagonalelementen $\gamma_1, \dots, \gamma_{n-1}$. Hierbei verwende man das in dem Unterabschnitt 5.3.3 angegebene Verfahren, einen QR -Schritt mit dem *impliziten* Shift-Parameter σ durchzuführen. Ferner benutze man zum Vergleich beide in 5.3.3 angegebenen Shift-Strategien. Man teste das Programm an den Matrizen aus Aufgabe 1 (nachdem diese auf symmetrische Tridiagonalgestalt transformiert sind) und berechne alle Eigenwerte dieser Matrizen.

Hinweis: In einem Programm muß entschieden werden, wann ein Subdiagonalelement $\gamma_i^{(k)}, i = 1, \dots, n-1$, „zu Null erklärt“ wird, so daß eine Deflation bzw. eine Reduktion auf zwei kleinere Probleme erfolgen kann. Der übliche Test hierfür ist

$$|\gamma_i^{(k)}| \leq \epsilon(|\delta_i^{(k)}| + |\delta_{i+1}^{(k)}|),$$

wobei $\epsilon > 0$ ein kleines Vielfaches von macheps ist, der kleinsten positiven Zahl, für die auf dem verwendeten Rechner $1 + \text{macheps} > 1$ gilt.

4. Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch mit Eigenwerten $\lambda_j(A), j = 1, \dots, n$, und $A = QR$ mit orthogonalem Q und oberer Dreiecksmatrix $R = (r_{ij})$, so ist $|r_{nn}| \geq \min_{j=1, \dots, n} |\lambda_j(A)|$.

Hinweis: Siehe J. H. WILKINSON (1968, Lemma 2).

5. Man schreibe ein Programm zur Berechnung der Eigenvektoren einer symmetrischen Tridiagonalmatrix mit Hilfe der inversen Iteration nach Wielandt (siehe 5.2.4). Hiermit berechne man Eigenvektoren zu den Matrizen aus Aufgabe 1, wobei die schon berechneten Eigenwerte zum Start benutzt werden können.

Hinweis: Natürlich sollte man davon Gebrauch machen, daß man die LR -Zerlegung einer Tridiagonalmatrix besonders einfach erhalten kann.

6. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Bei gegebenem $x \in \mathbb{R}^n \setminus \{0\}$ heißt

$$R(x) := \frac{x^T A x}{x^T x}$$

der zugehörige *Rayleigh-Quotient* (siehe Bemerkung im Anschluß an das Rayleighsche Maximumsprinzip, Satz 1.9). Seien $\lambda_1, \dots, \lambda_n$ die Eigenwerte von A . Man zeige:

- (a) Für jedes $x \in \mathbb{R}^n \setminus \{0\}$ ist

$$\min_{j=1, \dots, n} |R(x) - \lambda_j| \leq \sqrt{\left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 - R(x)^2}.$$

- (b) Sei $x \in \mathbb{R}^n \setminus \{0\}$ und $R(x)$ kein Eigenwert von A . Dann ist

$$\min_{j=1,\dots,n} |R(x) - \lambda_j| \leq \frac{\|x\|_2}{\|[A - R(x)I]^{-1}x\|_2}.$$

Hinweis: Man wende jeweils Korollar 1.8 an.

7. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch, $R(x)$ bezeichne wie in Aufgabe 6 den zu einem Vektor $x \in \mathbb{R}^n \setminus \{0\}$ gehörenden Rayleigh-Quotienten. Man betrachte das folgende Verfahren, das als *Rayleigh-Quotienten Iterationsverfahren* bekannt ist (siehe z. B. B. N. PARLETT (1980, S. 70 ff.) und G. H. GOLUB, C. F. VAN LOAN (1989, S. 440)).

- Wähle $x_0 \in \mathbb{R}^n$ mit $\|x_0\|_2 = 1$.
- Für $k = 0, 1, 2, \dots$:

Berechne $\mu_k := R(x_k)$.

Falls $A - \mu_k I$ singulär, dann:

Bestimme x_{k+1} mit $(A - \mu_k I)x_{k+1} = 0$, $\|x_{k+1}\|_2 = 1$, STOP

Andernfalls: Berechne $y_{k+1} := (A - \mu_k I)^{-1}x_k$.

Normiere y_{k+1} , d. h. berechne $x_{k+1} := y_{k+1}/\|y_{k+1}\|_2$.

Das Rayleigh-Quotienten Iterationsverfahren breche nicht vorzeitig ab und liefere Folgen $\{\mu_k\}$ und $\{x_k\}$. Dann gilt:

- (a) Sind $\lambda_1, \dots, \lambda_n$ die Eigenwerte von A , so ist

$$\min_{j=1,\dots,n} |\mu_k - \lambda_j| \leq \frac{1}{\|y_{k+1}\|_2}, \quad k = 0, 1, \dots$$

Man wird das Verfahren also abbrechen, wenn $\|y_{k+1}\|_2$ hinreichend groß ist.

- (b) Für $k = 0, 1, \dots$ ist $\|(A - \mu_{k+1} I)x_{k+1}\|_2 \leq \|(A - \mu_k I)x_k\|_2$, insbesondere konvergiert die Folge $\{\|(A - \mu_k) x_k\|_2\}$.
- (c) Ist $\tau := \lim_{k \rightarrow \infty} \|(A - \mu_k I)x_k\|_2 = 0$ und (μ, x) ein Häufungspunkt von $\{(\mu_k, x_k)\}$ (weshalb muß ein solcher existieren?), so ist μ ein Eigenwert und x ein zugehöriger Eigenvektor von A .

Hinweis: Bei B. N. PARLETT (1980, S. 72 ff.) findet man wesentlich subtilere Konvergenzaussagen. Diese zeigen, grob gesprochen, daß das Rayleigh-Quotienten Iterationsverfahren i. allg. global kubisch konvergiert, eine für ein Verfahren außerordentlich wünschenswerte Eigenschaft. Ist $A \in \mathbb{R}^{n \times n}$ eine (symmetrische) Tridiagonalmatrix, so ist der Aufwand pro Iteration im wesentlichen proportional zu n .

8. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Man betrachte das folgende Verfahren:

- Setze $\tilde{A}_1 := A$.
- Für $k = 1, 2, \dots$:

Berechne die Cholesky-Zerlegung von \tilde{A}_k , d. h. die untere Dreiecksmatrix L_k mit positiven Diagonalelementen, für die $\tilde{A}_k = L_k L_k^T$ (wegen Satz 4.1 in Abschnitt 1.4 ist die Cholesky-Zerlegung einer positiv definiten Matrix eindeutig bestimmt).

Berechne $\tilde{A}_{k+1} := L_k^T L_k$.

Das einfache QR-Verfahren ist dagegen gegeben durch

- Setze $A_1 := A$.
- Für $k = 1, 2, \dots$:

Berechne die QR-Zerlegung von A_k , d.h. die orthogonale Matrix Q_k und die obere Dreiecksmatrix R_k mit positiven Diagonalelementen, für die $A_k = Q_k R_k$ (unter dieser Zusatzbedingung ist die QR-Zerlegung einer nichtsingularen Matrix bekanntlich eindeutig bestimmt).

Berechne $A_{k+1} := R_k Q_k$.

Man zeige:

- (a) $\{\tilde{A}_k\}$ ist eine Folge symmetrischer, positiv definiter Matrizen, die zur Ausgangsmatrix A ähnlich sind.
- (b) Mit $\hat{L}_k := L_1 \cdots L_k$ ist $A^k = \hat{L}_k \hat{L}_k^T$.
- (c) Mit $\hat{R}_k := R_k \cdots R_1$ (siehe Satz 2.6) ist $A^{2k} = \hat{R}_k^T \hat{R}_k$.
- (d) Mit Hilfe der Eindeutigkeit der Cholesky-Zerlegung von A^{2k} schließe man auf $A_{k+1} = \tilde{A}_{2k+1}$.

Hinweis: Siehe J. H. WILKINSON (1965, S. 544 ff.). Das auf der Cholesky-Zerlegung basierende Verfahren ist nur von theoretischem, nicht aber von praktischem Interesse. Zwei Schritte dieses Verfahrens stimmen im wesentlichen mit einem Schritt des QR-Verfahrens überein. Es hat keine Bedeutung für die Praxis gewonnen, da es schwierig ist, einen Shift σ_k einzubauen, für den $A_k - \sigma_k I$ immer noch positiv definit ist. Andererseits ist eine geeignete Shift-Strategie notwendig, um die sonst unbefriedigende Konvergenzgeschwindigkeit zu verbessern.

9. Man schreibe eine Routine, die zu einer Matrix $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ Householder-Matrizen $P_1, \dots, P_n \in \mathbb{R}^{m \times m}$, $Q_1, \dots, Q_{n-2} \in \mathbb{R}^{n \times n}$ sowie eine obere Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ mit

$$A = (P_1 \cdots P_n) \begin{pmatrix} B \\ 0 \end{pmatrix} (Q_1 \cdots Q_{n-2})^T$$

berechnet. Anschließend schreibe man eine Routine, die aus den faktorisiert gespeicherten Householder-Matrizen die Produkte

$$P := P_1 \cdots P_n, \quad Q := Q_1 \cdots Q_{n-2}$$

bildet. Danach programmiere man einen Schritt des Verfahrens von Golub-Reinsch. Hier wird zu einer Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$, für die $B^T B$ unreduziert ist, und Matrizen $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ mit

$$A = U B V^T, \quad U^T U = I, \quad V^T V = I$$

durch einen impliziten Shift σ eine obere Bidiagonalmatrix $B_+ \in \mathbb{R}^{n \times n}$ sowie Matrizen $U_+ \in \mathbb{R}^{m \times n}$, $V_+ \in \mathbb{R}^{n \times n}$ mit

$$A = U_+ B_+ V_+^T, \quad U_+^T U_+ = I, \quad V_+^T V_+ = I$$

berechnet. Schließlich füge man das alles zum Verfahren von Golub-Reinsch zusammen und teste es an der Matrix

$$A := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}.$$

Hinweis: Bei der Transformation auf obere Bidiagonal-Gestalt haben wir die Darstellung

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T$$

mit

$$B := \begin{pmatrix} -12.88409873 & 21.87643283 \\ & 2.24623524 & -0.61328133 \\ & & 0.00000000 \end{pmatrix}$$

sowie

$$P := \begin{pmatrix} -0.07761505 & -0.83305216 & -0.32346650 & -0.44200614 \\ -0.31046021 & -0.45123659 & 0.76074059 & 0.34824383 \\ -0.54330537 & -0.06942101 & -0.55108168 & 0.62953076 \\ -0.77615053 & 0.31239456 & 0.11380759 & -0.53576845 \end{pmatrix},$$

$$Q := \begin{pmatrix} 1.00000000 & 0.00000000 & 0.00000000 \\ 0.00000000 & -0.66700225 & -0.74505570 \\ 0.00000000 & -0.74505570 & 0.66700225 \end{pmatrix}$$

erhalten. Mit

$$\Sigma := \text{diag}(25.46240744, 1.29066168, 0.00000000)$$

ist dann $A = U\Sigma V^T$ mit

$$U := \begin{pmatrix} -0.14087668 & -0.82471435 & -0.32346650 \\ -0.34394629 & -0.42626394 & 0.76074059 \\ -0.54701591 & -0.02781353 & -0.55108168 \\ -0.75008553 & 0.37063688 & 0.11380759 \end{pmatrix},$$

$$V := \begin{pmatrix} -0.50453315 & 0.76077568 & 0.40824829 \\ -0.57451570 & 0.05714052 & -0.81649658 \\ -0.64449826 & -0.64649464 & 0.40824829 \end{pmatrix}$$

die gesuchte Singulärwertzerlegung von A . Hierbei sind die Diagonalelemente von Σ nichtnegativ gemacht worden, indem man notfalls ein entsprechendes Element und die entsprechende Spalte in V mit -1 multipliziert.

10. Seien $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $b \in \mathbb{R}^m$ gegeben. Mit Hilfe einer Singulärwertzerlegung von A lässt sich die eindeutige Lösung minimaler euklidischer Norm des linearen Ausgleichsproblems

$$(P) \quad \text{Minimiere } \|Ax - b\|_2, \quad x \in \mathbb{R}^n$$

in einfacher Weise berechnen (siehe die Bemerkung am Schluß von 5.3.4). Gegeben seien nun die Matrizen

$$A := \begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}, \quad B := \begin{pmatrix} -1 & 1 & 0 \\ 2 & -1 & 1 \\ 1 & 10 & 11 \\ 4 & 0 & 4 \\ 0 & -6 & -6 \\ -3 & 6 & 3 \\ 1 & 11 & 12 \\ 0 & -5 & -5 \end{pmatrix}.$$

Zu A und den Spalten von B bestimme man die drei zugehörigen Lösungen minimaler euklidischer Norm von (P).

Hinweis: Es ist $\text{Rang}(A) = 3$, die singulären Werte von A sind

$$\begin{aligned}\sigma_1 &= \sqrt{1248} \approx 35.327043465311, \\ \sigma_2 &= 20 = 20.000000000000, \\ \sigma_3 &= \sqrt{384} \approx 19.595917942265,\end{aligned}$$

die Lösungen x_1, x_2, x_3 der drei Ausgleichsprobleme sind:

x_1	x_2	x_3
-0.0833333333	0.0000000000	-0.0833333333
0.0000000000	0.0000000000	0.0000000000
0.2500000000	0.0000000000	0.2500000000
-0.0833333333	0.0000000000	-0.0833333333
0.0833333333	0.0000000000	0.0833333333

(siehe J. H. WILKINSON, C. REINSCH (1971, S.149)).

11. Wie kann man mit dem in 5.3.4 beschriebenen Verfahren auf einfache Weise zu einer vorgegebenen Matrix $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ orthogonale Matrizen $U_0 \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ berechnen derart, daß

$$A = U_0 \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T, \quad \Sigma := \text{diag}(\sigma_1, \dots, \sigma_n)$$

gilt?

Hinweis: Man beachte, daß im Reduktionsschritt orthogonale Matrizen $P \in \mathbb{R}^{m \times m}$, $Q \in \mathbb{R}^{n \times n}$ sowie eine Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ mit

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T$$

berechnet sind.

12. Bei gegebenen $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$ und $\alpha > 0$ betrachte man das folgende, restriktive lineare Ausgleichsproblem:

$$(P) \quad \text{Minimiere } \|Ax - b\|_2, \quad \|x\|_2 \leq \alpha.$$

Man zeige: Ein $x^* \in \mathbb{R}^n$ mit $\|x^*\|_2 \leq \alpha$ ist eine Lösung von (P), wenn ein $\lambda^* \in \mathbb{R}$ mit

$$(*) \quad \lambda^* \geq 0, \quad (A^T A + \lambda^* I)x^* = A^T b, \quad \lambda^*(\alpha^2 - \|x^*\|_2^2) = 0$$

existiert.

Hinweis: Sei $x \in \mathbb{R}^n$ mit $\|x\|_2 \leq \alpha$ beliebig. Dann ist

$$\begin{aligned} \frac{1}{2} \|Ax - b\|_2^2 - \frac{1}{2} \|Ax^* - b\|_2^2 &\geq (A^T A x^* - A^T b)^T (x - x^*) \\ &= \lambda^* (x^* - x)^T x^* \\ &= \lambda^* (\alpha^2 - x^T x^*) \\ &\geq 0, \end{aligned}$$

also x^* eine Lösung von (P).

13. Zur numerischen Lösung des in Aufgabe 12 formulierten restringierten linearen Ausgleichsproblems (P) betrachte man den folgenden Algorithmus:

- Berechne eine Singulärwertzerlegung von A , also Matrizen $U \in \mathbb{R}^{m \times n}$, $V = (v_1 \ \dots \ v_n) \in \mathbb{R}^{n \times n}$ und $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ mit

$$A = U \Sigma V^T, \quad U^T U = V^T V = I, \quad \sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0.$$

- Berechne $b := U^T b$
- Falls $\sum_{i=1}^r (b_i/\sigma_i)^2 \leq \alpha^2$, dann:
Berechne $x^* := \sum_{i=1}^r (b_i/\sigma_i) v_i$

Andernfalls:

Bestimme $\lambda^* > 0$ mit $\sum_{i=1}^r [\sigma_i b_i / (\sigma_i^2 + \lambda^*)]^2 = \alpha^2$

Berechne $x^* := \sum_{i=1}^r [\sigma_i b_i / (\sigma_i^2 + \lambda^*)] v_i$

Mit Hilfe von Aufgabe 12 motiviere und analysiere man diesen Algorithmus. Anschließend teste man das Verfahren an den Daten aus Aufgabe 10 sowie $\alpha := 0.2$.

Hinweis: Offenbar wird zunächst getestet, ob $\|A^+ b\|_2^2 \leq \alpha^2$. Ist dies der Fall, so ist $x^* := A^+ b$ (setze $\lambda^* := 0$ in Aufgabe 12) eine Lösung von (P). Für $\lambda > 0$ besitzt

$$(A^T A + \lambda I)x = A^T b$$

die eindeutige Lösung $x(\lambda) := V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T b$. Daher bestimme man $\lambda^* > 0$ so, daß $\|x(\lambda^*)\|_2^2 = \alpha^2$. Wegen Aufgabe 12 ist $x^* := x(\lambda^*)$ eine Lösung von (P).

Kapitel 6

Lineare Optimierungsaufgaben

Lineare Optimierungsaufgaben treten bei vielen Anwendungen auf, ihre numerische Behandlung gehört zu den wichtigsten Problemen der numerischen Mathematik. Zahlreiche (zum Teil umfangreiche) Lehrbücher gibt es über dieses Gebiet. Wir nennen nur G. B. DANTZIG (1966), G. HADLEY (1962), D. GALE (1960), L. COLLATZ, W. WETTERLING (1971), C. H. PAPADIMITRIOU, K. STEIGLITZ (1982), V. CHVÁTAL (1983), G. R. WALSH (1985), A. SCHRIJVER (1986) und verweisen insbesondere auf einen neuere Entwicklungen berücksichtigenden Übersichtsartikel von D. GOLDFARB, M. J. TODD (1989).

Nach einer Einführung schildern wir in Abschnitt 6.2 ausführlich das inzwischen klassische Simplexverfahren, das 1947 von G. B. Dantzig entwickelt wurde. Hierbei werden wir allerdings auf die vielen Modifikationen für speziell strukturierte lineare Optimierungsaufgaben kaum eingehen, hierzu sei auf die angegebene Literatur verwiesen. Hieran anschließend wird in Abschnitt 6.3 verhältnismäßig kurz über die Dualitätstheorie der linearen Optimierung berichtet. In einem abschließenden Abschnitt 6.4 erläutern wir einige der wesentlichen Ideen des Karmarkar-Verfahrens, das 1984 vorgestellt wurde und, nur geringfügig übertrieben, für Furore sorgte.

6.1 Einführung, Beispiele

Unter einer *linearen Optimierungsaufgabe* (oft auch *lineares Programm* genannt) versteht man, grob gesagt, die Aufgabe, eine lineare Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ auf einer Menge $M \subset \mathbb{R}^n$, die durch endlich viele affin lineare Gleichungen und Ungleichungen gegeben ist, zu minimieren. Die zu minimierende Funktion f heißt *Ziel- oder Kostenfunktion*, die Menge M derjenigen Punkte des \mathbb{R}^n , die den geforderten Nebenbedingungen bzw. Restriktionen genügen, heißt die Menge der *zulässigen Lösungen*. Gesucht ist ein $x^* \in M$ mit $f(x^*) \leq f(x)$ für alle $x \in M$, bzw. eine zulässige Lösung mit minimalen Kosten. Man beachte, daß die Minimierung von f gleichwertig mit der Maximierung von $-f$ ist.

Ganz am Anfang wollen wir hier schon eine Bezeichnung vereinbaren. Und zwar wird die " \leq "- bzw. die " \geq "-Relation zwischen Vektoren stets komponentenweise verstanden. Für $x = (x_j), y = (y_j) \in \mathbb{R}^n$ ist daher $x \leq y$ genau dann, wenn $x_j \leq y_j$ für $j = 1, \dots, n$. Entsprechend ist $x \geq y$ definiert.

Beispiel: Beim *Produktionsplanungsproblem* wird davon ausgegangen, daß in einem Betrieb (z. B. einer Molkerei) n Produkte (z. B. Butter, Buttermilch, Milchpulver, verschiedene Käsesorten und Molke) hergestellt werden, wozu m Hilfsmittel (z. B. Rohmilch, Maschinen, Lagerraum) benötigt werden. Unter der Nebenbedingung, daß die Hilfsmittel nur bis zu einer gewissen Maximalmenge zur Verfügung stehen, soll der Gewinn maximiert werden. Ein *Produktionsplan* besteht in der Angabe eines Vektors $x = (x_1, \dots, x_n)^T$. Dieser sagt aus, daß vom j -ten Produkt x_j Mengeneinheiten produziert werden sollen. Zur Herstellung einer Mengeneinheit des j -ten Produktes werden a_{ij} Mengeneinheiten des i -ten Hilfsmittels benötigt. Ferner steht das i -te Hilfsmittel nur bis zu einer endlichen Maximalmenge von b_i Einheiten zur Verfügung. Ein Produktionsplan $x = (x_1, \dots, x_n)^T$ ist daher zulässig, wenn

$$\sum_{j=1}^n a_{ij} x_j \leq b_i \quad (i = 1, \dots, m), \quad x_j \geq 0 \quad (j = 1, \dots, n).$$

Ist p_j der Reingewinn bei der Herstellung einer Mengeneinheit des j -ten Produkts, so ist der Gesamtgewinn bei der Anwendung des Produktionsplanes $x = (x_1, \dots, x_n)^T$ durch $\sum_{j=1}^n p_j x_j$ gegeben. Mit

$$A := (a_{ij}) \in \mathbb{R}^{m \times n}, \quad b := (b_i) \in \mathbb{R}^m, \quad p := (p_j) \in \mathbb{R}^n$$

ist das Produktionsplanungsproblem daher durch

$$\text{Maximiere } p^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax \leq b\}$$

gegeben. □

Beispiel: Es seien n Nahrungsmittel gegeben, die m Grundsubstanzen (etwa Eiweiß, Fett, Vitamine, Kohlehydrate) enthalten. Eine Einheit des j -ten Nahrungsmittels enthalte a_{ij} Einheiten der i -ten Grundsubstanz. Ein *Diätplan* besteht in der Angabe eines Vektors $x = (x_1, \dots, x_n)^T$ und dieser sagt aus, daß vom j -ten Nahrungsmittel x_j Einheiten zu nehmen sind. Er ist aber nur dann zulässig, wenn das zugehörige „Menü“ von der i -ten Grundsubstanz mindestens b_i Einheiten enthält, d. h. wenn

$$\sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m) \quad \text{bzw.} \quad Ax \geq b,$$

und die triviale Nebenbedingung

$$x_j \geq 0 \quad (j = 1, \dots, n) \quad \text{bzw.} \quad x \geq 0$$

erfüllt sind. Eine Einheit des j -ten Nahrungsmittels koste c_j Geldeinheiten. Dann kostet der Diätplan $x = (x_1, \dots, x_n)^T$ insgesamt $c^T x = \sum_{j=1}^n c_j x_j$ Geldeinheiten, so daß wir beim *Diätproblem* die lineare Optimierungsaufgabe

$$\text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax \geq b\}$$

zu lösen haben. □

Beispiel: Ein überbestimmtes lineares Gleichungssystem $Ax = b$ mit einer Koeffizientenmatrix $A \in \mathbb{R}^{m \times n}$, einer rechten Seite $b \in \mathbb{R}^m$ und $m \geq n$ ist i. allg. nicht lösbar. Bei einem linearen Ausgleichsproblem wird der Defekt bezüglich der euklidischen Norm im \mathbb{R}^m minimiert. Es kann aber auch sinnvoll sein, den Defekt bezüglich der Maximumsnorm im \mathbb{R}^m zu minimieren, also die Aufgabe

$$\text{Minimiere } \|Ax - b\|_\infty, \quad x \in \mathbb{R}^n$$

zu betrachten. Man spricht hier von einer *diskreten, linearen Tschebyscheffschen Approximationsaufgabe*. Eine solche Aufgabe kann als lineare Optimierungsaufgabe formuliert werden. Denn ist (x^*, δ^*) eine Lösung des linearen Programms

$$\text{Minimiere } \delta \text{ auf } M := \{(x, \delta) \in \mathbb{R}^n \times \mathbb{R} : \delta \geq 0, -\delta e \leq Ax - b \leq \delta e\},$$

wobei $e := (1, \dots, 1)^T \in \mathbb{R}^m$, so ist x^* eine Lösung des diskreten, linearen Tschebyscheffschen Approximationproblems und $\delta^* = \|Ax^* - b\|_\infty$. Hierzu gilt offenbar auch die Umkehrung, so daß wir beide Probleme mit gutem Gewissen als äquivalent bezeichnen können. \square

Nach diesen Beispielen präzisieren wir die Form der linearen Optimierungsaufgaben, mit denen wir uns beschäftigen werden, und zeigen, wie man diese in eine äquivalente *Normalform* überführen kann.

Im folgenden sei m stets die Anzahl der Nebenbedingungen und n die Anzahl der Variablen. Von den m Nebenbedingungen mögen m_0 in Form von Ungleichungen vorliegen (mit $m_0 \in \mathbb{Z}$ und $0 \leq m_0 \leq m$). Wir können annehmen, daß die Nebenbedingungen so numeriert sind, daß zuerst die Ungleichungen und dann erst die Gleichungen kommen. Ferner können wir davon ausgehen, daß die Ungleichungsrestriktionen sozusagen gleichgerichtet sind, also sämtlich als \geq -Ungleichungen auftreten (notfalls multipliziere man \leq -Ungleichungen mit -1). Wie wir in obigen Beispielen (außer bei der diskreten Tschebyscheff-Approximation) gesehen haben, sind oft wenigstens gewisse Variablen x_j durch $x_j \geq 0$ vorzeichenbeschränkt. Notfalls kann durch eine Ummumerierung erreicht werden, daß die ersten n_0 Variablen (mit $n_0 \in \mathbb{Z}$ und $0 \leq n_0 \leq n$) vorzeichenbeschränkt, die restlichen $n - n_0$ frei sind. Als Input-Daten haben wir daher:

- Gegeben $m \in \mathbb{N}$ (Anzahl der Nebenbedingungen), $n \in \mathbb{N}$ (Anzahl der Variablen), $m_0 \in \mathbb{Z}$ mit $0 \leq m_0 \leq m$ (Anzahl der Ungleichungsrestriktionen) und $n_0 \in \mathbb{Z}$ mit $0 \leq n_0 \leq n$ (Anzahl der vorzeichenbeschränkten Variablen), ferner die Koeffizientenmatrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, die rechte Seite $b = (b_i) \in \mathbb{R}^m$ und der Kostenvektor $c = (c_j) \in \mathbb{R}^n$.

Die zu diesen Daten gehörende lineare Optimierungsaufgabe lautet:

$$\begin{aligned} & \text{Minimiere } \sum_{j=1}^n c_j x_j \quad \text{auf} \\ M := & \left\{ x \in \mathbb{R}^n : x_j \geq 0 \quad (j = 1, \dots, n_0), \quad \begin{array}{l} \sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m_0), \\ \sum_{j=1}^n a_{ij} x_j = b_i \quad (i = m_0 + 1, \dots, m) \end{array} \right\}. \end{aligned}$$

Man sagt, das gegebene lineare Programm sei in *Normalform*, wenn $n_0 = n$, also alle Variablen vorzeichenbeschränkt sind, und $m_0 = 0$ ist, bzw. nur Gleichungen als Restriktionen auftreten. Ein lineares Programm in Normalform hat damit die Form

$$\text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}.$$

Nur *scheinbar* ist die Normalform ein Spezialfall eines linearen Programms. Denn jede lineare Optimierungsaufgabe läßt sich in eine äquivalente Aufgabe in Normalform überführen, wobei allerdings die Anzahl der Variablen vergrößert wird. Unter *äquivalent* verstehen wir hierbei, etwas lax formuliert, daß man aus einer Lösung der einen Aufgabe eine der anderen sofort erhalten kann (und umgekehrt).

Durch m_0 nichtnegative *Schlupfvariable* y_i , $i = 1, \dots, m_0$, können die m_0 Ungleichungsrestriktionen in Gleichungen überführt werden:

$$\sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m_0) \iff \left. \begin{array}{l} \sum_{j=1}^n a_{ij} x_j - y_i = b_i \\ y_i \geq 0 \end{array} \right\} \quad (i = 1, \dots, m_0).$$

Im (erweiterten) Zielfunktionsvektor bekommen die Schlupfvariablen natürlich das Gewicht Null.

Die $n - n_0$ freien Variablen x_j können jeweils durch ein Paar (x_j^+, x_j^-) nichtnegativer Variabler ersetzt werden:

$$x_j = x_j^+ - x_j^- \quad \text{mit } x_j^+ \geq 0, x_j^- \geq 0 \quad (j = n_0 + 1, \dots, n).$$

Den Prozeß der Überführung eines linearen Programms in äquivalente Normalform wollen wir uns durch Beispiele klar machen.

Beispiel: Die „natürliche“ Formulierung des Produktionsplanungsproblems ist, wie wir oben gesehen haben, durch

$$\text{Maximiere } p^T x \text{ unter den Nebenbedingungen } x \geq 0, \quad Ax \leq b$$

gegeben. Hier ist es sinnvoll, die \leq -Ungleichungsrestriktionen durch *Addition* von nichtnegativen Schlupfvariablen in Gleichungen zu überführen. Als äquivalentes Programm in Normalform erhält man damit:

$$\begin{aligned} \text{Minimiere } & \left(\begin{array}{c} -p \\ 0 \end{array} \right)^T \left(\begin{array}{c} x \\ y \end{array} \right) \text{ unter den Nebenbedingungen} \\ & \left(\begin{array}{c} x \\ y \end{array} \right) \geq \left(\begin{array}{c} 0 \\ 0 \end{array} \right), \quad (A \quad I) \left(\begin{array}{c} x \\ y \end{array} \right) = b. \end{aligned}$$

Man beachte, daß hier der Vektor $y \in \mathbb{R}^{m_0}$ der Schlupfvariablen eine naheliegende ökonomische Interpretation zuläßt. Denn ist $x = (x_1, \dots, x_n)^T$ ein zulässiger Produktionsplan, so gibt $y_i = b_i - \sum_{j=1}^n a_{ij} x_j$ an, wieviele Einheiten des i -ten Hilfsmittels von dem Produktionsplan nicht benutzt werden. \square

Beispiel: Das Diätproblem führte auf ein lineares Programm der Form

$$\text{Minimiere } c^T x \text{ unter den Nebenbedingungen } x \geq 0, \quad Ax \geq b.$$

Ein äquivalentes Programm in Normalform ist durch

$$\begin{array}{l} \text{Minimiere } \begin{pmatrix} c \\ 0 \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} \text{ unter den Nebenbedingungen} \\ \begin{pmatrix} x \\ y \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (A \quad -I) \begin{pmatrix} x \\ y \end{pmatrix} = b \end{array}$$

gegeben. \square

Beispiel: Gegeben sei die lineare Optimierungsaufgabe

$$(P_0) \quad \left\{ \begin{array}{l} \text{Minimiere } -x_1 - x_2 \text{ unter den Nebenbedingungen} \\ x_1 + 3x_2 \leq 13, \\ x_1 \geq 0, \quad x_2 \geq 0, \quad 3x_1 + x_2 \leq 15, \\ -x_1 + x_2 \leq 3. \end{array} \right.$$

In Abbildung 6.1 skizzieren wir die zugehörige Menge M_0 der zulässigen Lösungen in einer (x_1, x_2) -Ebene. Mit \bullet sind die Ecken von M_0 gekennzeichnet. Nach Einführung

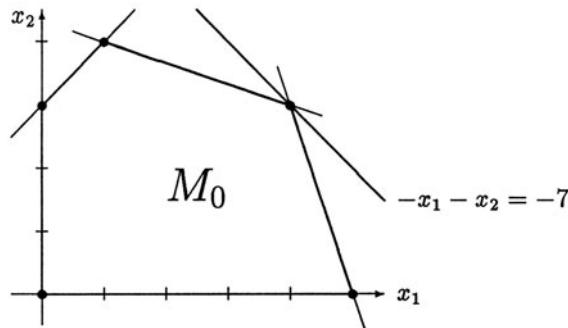


Abbildung 6.1: Veranschaulichung des obigen linearen Programms

von Schlupfvariablen x_3, x_4, x_5 erkennt man, daß (P_0) äquivalent dazu ist, eine Lösung von

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

zu bestimmen, wobei

$$m := 3, \quad n := 5, \quad \begin{array}{|c|c|} \hline c^T & \boxed{} \\ \hline A & \boxed{b} \\ \hline \end{array} := \begin{array}{|c|c|c|c|c|c|} \hline -1 & -1 & 0 & 0 & 0 & | \\ \hline 1 & 3 & 1 & 0 & 0 & | 13 \\ \hline 3 & 1 & 0 & 1 & 0 & | 15 \\ \hline -1 & 1 & 0 & 0 & 1 & | 3 \\ \hline \end{array}$$

Ganz offensichtlich ist die Ecke $(4, 3)$ die eindeutige Lösung der gestellten Optimierungsaufgabe (P_0) . \square

Nachdem einige Beispiele von linearen Optimierungsaufgaben angegeben worden sind, die allgemeine Form eines linearen Programms beschrieben und die Überführung in äquivalente Normalform erläutert wurde, sollen jetzt noch einige grundlegende Begriffe eingeführt werden. Hierbei wird *nicht* die spezielle Struktur linearer Programme benutzt, so daß wir von einer nicht notwendig linearen Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) \text{ auf } M$$

ausgehen, wobei $f: \mathbb{R}^n \rightarrow \mathbb{R}$ und $M \subset \mathbb{R}^n$. Dann heißt (P) *zulässig*, wenn die Menge M der *zulässigen Lösungen* nicht leer ist. Eine zulässige Lösung $x^* \in M$ heißt eine (globale) *Lösung* von (P), wenn $f(x^*) \leq f(x)$ für alle $x \in M$. Besitzt (P) eine Lösung, so heißt (P) *lösbar*. Mit

$$\inf(P) := \begin{cases} \inf_{x \in M} f(x) & \text{für } M \neq \emptyset, \\ +\infty & \text{für } M = \emptyset \end{cases}$$

wird der *Wert* von (P) bezeichnet. Ist (P) lösbar, so schreiben wir $\min(P)$ statt $\inf(P)$.

Die wichtigsten Fragen, die wir für lineare Optimierungsaufgaben in den folgenden Abschnitten beantworten wollen, sind dann:

- Wann ist die Lösbarkeit von (P) gesichert?
- Was sind notwendige, was hinreichende Bedingungen dafür, daß ein $x^* \in M$ eine Lösung von (P) ist?
- Wie berechnet man eine Lösung von (P)?

Aufgaben

1. Ein Landwirt bewirtschaftet ein Grundstück von 40 ha Größe mit Zuckerrüben und Weizen. Er kann hierzu 2400 DM und 312 Arbeitstage einsetzen. Pro ha betragen seine Anbaukosten bei Rüben 40 DM und bei Weizen 120 DM. Für Rüben benötigt er 7 Arbeitstage, für Weizen 12 Arbeitstage pro ha. Der Reingewinn bei Rüben sei 100 DM pro ha, bei Weizen sei er 250 DM pro ha.

Man stelle das zugehörige lineare Programm auf, stelle in einer (x_1, x_2) -Ebene (bzw. einer Rüben-Weizen-Ebene) die Menge der zulässigen Lösungen dar und ermittle graphisch eine Lösung.

2. Die Aufgabe, bei gegebenen $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ mit $m \geq n$ den Defekt $Ax - b$ bezüglich der Maximumnorm zu minimieren, führt auf das lineare Programm

$$\text{Minimiere } \delta \text{ unter den Nebenbedingungen } \delta \geq 0, \quad -\delta e \leq Ax - b \leq \delta e,$$

wobei $e = (1, \dots, 1)^T \in \mathbb{R}^m$. Man führe dieses lineare Programm in eine äquivalente Normalform über.

3. Sei $A \in \mathbb{R}^{m \times n}$ und

$$X := \left\{ x \in \mathbb{R}^n : x \geq 0, \sum_{j=1}^n x_j = 1 \right\}, \quad Y := \left\{ y \in \mathbb{R}^m : y \geq 0, \sum_{i=1}^m y_i = 1 \right\}.$$

Man zeige, daß die Optimierungsaufgabe

$$\text{Minimiere } f(x) := \max_{y \in Y} y^T Ax \text{ auf } X$$

äquivalent dem linearen Programm

$$\text{Minimiere } \alpha \text{ unter den Nebenbedingungen } Ax \leq \alpha e, \quad x \geq 0, \quad e^T x = 1$$

ist. Hierbei bezeichne e den Vektor im \mathbb{R}^m bzw. \mathbb{R}^n , der nur Einsen als Komponenten besitzt.

4. Man gebe Beispiele von zulässigen linearen Programmen an, die

- (a) nicht lösbar sind,
- (b) eindeutig lösbar sind,
- (c) unendlich viele Lösungen besitzen.

Ferner beweise man, daß das lineare Programm aus Aufgabe 2 zulässig und lösbar ist.

6.2 Das Simplexverfahren

6.2.1 Geometrische Grundlagen des Simplexverfahrens

Das Simplexverfahren zur Lösung linearer Optimierungsaufgaben besitzt eine einfache geometrische Interpretation. Um diese und ihre algebraische Umsetzung zu verstehen, ist die Einführung einiger Begriffe nützlich.

Eine Menge $C \subset \mathbb{R}^n$ heißt bekanntlich *konvex*, wenn mit je zwei Punkten aus C auch die gesamte Verbindungsstrecke zwischen diesen beiden Punkten zu C gehört, wenn also

$$x, y \in C \implies [x, y] := \{(1 - \lambda)x + \lambda y : \lambda \in [0, 1]\} \subset C.$$

So ist z. B. die Menge der zulässigen Lösungen eines linearen Programms eine konvexe Menge. Das gilt insbesondere für

$$M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

die Menge der zulässigen Lösungen eines linearen Programms in Normalform. Hier und im folgenden ist $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$.

Ein Punkt $x \in \mathbb{R}^n$ heißt *Konvekzkombination* von Punkten $x_1, \dots, x_N \in \mathbb{R}^n$, wenn sich x in der Form

$$x = \sum_{i=1}^N \lambda_i x_i \quad \text{mit} \quad \lambda_i \geq 0 \quad (i = 1, \dots, N) \quad \text{und} \quad \sum_{i=1}^N \lambda_i = 1$$

darstellen läßt.

Ist $C \subset \mathbb{R}^n$ konvex, so heißt ein Punkt $x \in C$ eine *Ecke* von C , wenn sich x nicht als Konvexitätskombination von zwei anderen Punkten aus C darstellen läßt.

Ist $a \in \mathbb{R}^n$, $a \neq 0$ und $\beta \in \mathbb{R}$, so heißt $H := \{x \in \mathbb{R}^n : a^T x = \beta\}$ eine *Hyperebene* im \mathbb{R}^n und $H^- := \{x \in \mathbb{R}^n : a^T x \leq \beta\}$ ein zugehöriger *abgeschlossener Halbraum*.

Ein *Polyeder* ist der Durchschnitt von endlich vielen abgeschlossenen Halbraümen. Als Durchschnitt konvexer Mengen ist ein Polyeder konvex. Ein nichtleeres, beschränktes Polyeder heißt ein *Polytop*. So ist z. B. die Menge der zulässigen Lösungen eines linearen Programms ein Polyeder. Dies gilt insbesondere für die Menge M der zulässigen Lösungen eines linearen Programms in Normalform. In dem folgenden einfachen Lemma werden notwendige und hinreichende Bedingungen dafür angegeben, daß M sogar ein Polytop ist.

Lemma 2.1 Die Menge $M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$ sei nichtleer. Dann ist M genau dann beschränkt bzw. ein Polytop, wenn es kein $d \in \mathbb{R}^n \setminus \{0\}$ mit $d \geq 0$ und $Ad = 0$ gibt.

Beweis: Gibt es ein $d \neq 0$ mit $d \geq 0$ und $Ad = 0$, so ist mit beliebigem $x \in M$ auch $x + td \in M$ für alle $t \geq 0$. Ein von x ausgehender Strahl in Richtung d würde also in M liegen, so daß M notwendig nicht beschränkt ist. Ist umgekehrt M nicht beschränkt, so gibt es eine Folge $\{x^k\} \subset M$ mit $\|x^k\|_2 \rightarrow +\infty$. Aus $d^k := x^k/\|x^k\|_2$ läßt sich eine gegen ein $d \neq 0$ konvergente Teilfolge auswählen. Offenbar ist $d \geq 0$ und $Ad = 0$. \square

In Abbildung 6.2 werden einige der eingeführten Begriffe veranschaulicht. Die hier angegebene Menge C ist ein Polytop. Offensichtlich läßt sich jeder Punkt von C als Konvexitätskombination der endlich vielen Ecken darstellen.

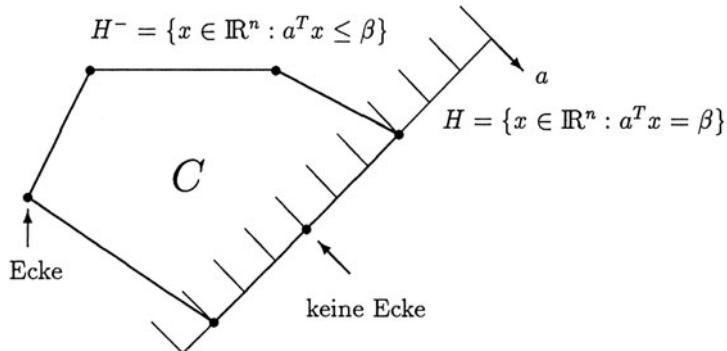


Abbildung 6.2: Veranschaulichung einiger grundlegender Begriffe

Ecken der Menge M der zulässigen Lösungen eines linearen Programms in Normalform werden in dem folgenden Satz algebraisch charakterisiert.

Satz 2.2 Sei $M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$ mit $A = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Dann ist ein $x \in M$ genau dann eine Ecke von M , wenn die zu positiven Komponenten von x gehörenden Spalten von A linear unabhängig sind, wenn also die Spalten $\{a_j\}_{j \in B(x)}$ mit $B(x) := \{j \in \{1, \dots, n\} : x_j > 0\}$ linear unabhängig sind. Ist ferner $M \neq \emptyset$, so besitzt M mindestens eine, aber höchstens endlich viele Ecken.

Beweis: Angenommen, die Vektoren $\{a_j\}_{j \in B(x)}$ seien linear abhängig. Dann existieren reelle Zahlen β_j , $j \in B(x)$, die nicht alle verschwinden, mit $\sum_{j \in B(x)} \beta_j a_j = 0$. Wegen $x_j > 0$ für $j \in B(x)$ existiert ein $\delta > 0$ mit $x_j \pm \delta \beta_j \geq 0$ für $j \in B(x)$. Definiert man $x^1, x^2 \in \mathbb{R}^n$ durch

$$x_j^1 := \begin{cases} x_j + \delta \beta_j & \text{falls } j \in B(x), \\ 0 & \text{falls } j \notin B(x), \end{cases} \quad x_j^2 := \begin{cases} x_j - \delta \beta_j & \text{falls } j \in B(x), \\ 0 & \text{falls } j \notin B(x), \end{cases}$$

so sind x^1 und x^2 zwei von x verschiedene Punkte aus M , die x als Mittelpunkt besitzen, so daß x keine Ecke von M ist. Ist also $x \in M$ eine Ecke von M , so sind die Vektoren $\{a_j\}_{j \in B(x)}$ linear unabhängig.

Umgekehrt seien die Vektoren $\{a_j\}_{j \in B(x)}$ linear unabhängig. Ist $x = (1-\lambda)x^1 + \lambda x^2$ mit $x^1, x^2 \in M$ und $\lambda \in (0, 1)$, so folgt zunächst aus $x_j = 0$ für $j \notin B(x)$, daß auch $x_j^1 = x_j^2 = 0$ für $j \notin B(x)$. Also ist

$$0 = b - b = Ax^1 - Ax^2 = A(x^1 - x^2) = \sum_{j \in B(x)} (x_j^1 - x_j^2) a_j.$$

Wegen der linearen Unabhängigkeit der $\{a_j\}_{j \in B(x)}$ folgt $x_j^1 = x_j^2$ für alle $j \in B(x)$, insgesamt $x^1 = x^2 = x$. Der Punkt $x \in M$ läßt sich also nicht als Konvexitätskombination zweier anderer Punkte aus M darstellen, ist also eine Ecke von M .

Sei nun $M \neq \emptyset$. Dann existiert ein Punkt x^* in M mit einer minimalen Anzahl positiver Komponenten. Deren Indizes seien mit $B(x^*)$ bezeichnet. O. B. d. A. ist $B(x^*) \neq \emptyset$, da andernfalls $x^* = 0$ eine Ecke von M ist. Die Spalten $\{a_j\}_{j \in B(x^*)}$ sind linear unabhängig, da man sonst (siehe den ersten Teil des obigen Beweises) ein $\hat{x} \in M$ mit weniger positiven Komponenten als x^* finden könnte, was der Definition von x^* widersprechen würde. Also ist x^* eine Ecke von M , daher besitzt M wenigstens eine Ecke. Ferner gibt es nur endlich viele Ecken von M , da es nur eine endliche Zahl von Möglichkeiten gibt, aus den n Spalten von A linear unabhängige Spalten auszuwählen. \square

Anschaulich (siehe Abbildung 6.2) ist ziemlich klar, daß sich jeder Punkt eines Polytops als Konvexitätskombination seiner endlich vielen Ecken darstellen läßt. Der folgende Satz präzisiert diese Aussage für die Menge M der zulässigen Lösungen eines linearen Programms in Normalform.

Satz 2.3 Sei $M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\} \neq \emptyset$, wobei $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Mit $\{v_i : i \in I\}$ werde die nichtleere, endliche Menge der Ecken von M bezeichnet. Dann läßt sich jeder Punkt $x \in M$ in der Form

$$x = \sum_{i \in I} \lambda_i v_i + d$$

darstellen, wobei

$$\lambda_i \geq 0 \quad (i \in I), \quad \sum_{i \in I} \lambda_i = 1$$

und $d \geq 0$, $Ad = 0$. Insbesondere gilt: Ist M beschränkt, also ein Polytop, so läßt sich jeder Punkt aus M als Konvexitätskombination der endlich vielen Ecken von M darstellen.

Beweis: Der Beweis wird durch vollständige Induktion nach der Anzahl p positiver Komponenten von $x \in M$ erbracht. Ist $p = 0$, so ist $x = 0$ selber schon eine Ecke und die behauptete Darstellung gilt trivialerweise. Es wird nun angenommen, daß man jeden Punkt aus M mit weniger als p positiven Komponenten in der behaupteten Weise darstellen kann. Ferner besitze $x \in M$ genau p positive Komponenten, so daß $B(x) := \{j \in \{1, \dots, n\} : x_j > 0\}$ aus genau p Elementen besteht.

O. B. d. A. ist x keine Ecke von M . Nach Satz 2.2 sind die Spalten $\{a_j\}_{j \in B(x)}$ linear abhängig. Daher existiert ein $w \in \mathbb{R}^n \setminus \{0\}$ mit $w_j = 0$ für alle $j \notin B(x)$ und $Aw = 0$. Es wird nun unterschieden zwischen den Fällen, daß w Komponenten von beiderlei Vorzeichen hat, daß $w \geq 0$ oder daß $w \leq 0$.

Der Vektor w habe positive *und* negative Komponenten. Die Idee des Beweises besteht darin, x als Konvexitätskombination von zwei Punkten aus M mit weniger als p positiven Komponenten darzustellen, auf diese die Induktionsvoraussetzung anzuwenden und dadurch auf die behauptete Darstellung von x zu schließen. Hierzu definiere man positive Zahlen δ_1 und δ_2 durch

$$\delta_1 := \min \left\{ \frac{x_j}{w_j} : j \in B(x), w_j > 0 \right\}, \quad \delta_2 := \min \left\{ -\frac{x_j}{w_j} : j \in B(x), w_j < 0 \right\}$$

und anschließend

$$x^1 := x - \delta_1 w, \quad x^2 := x + \delta_2 w.$$

Nach Konstruktion sind x^1 und x^2 zwei Punkte aus M , die weniger als p positive Komponenten besitzen. Ferner ist

$$x = (1 - \mu)x^1 + \mu x^2 \quad \text{mit} \quad \mu := \frac{\delta_1}{\delta_1 + \delta_2} \in (0, 1)$$

ein Punkt auf der Verbindungsstrecke zwischen x^1 und x^2 . Nach Induktionsvoraussetzung besitzen x^1 und x^2 Darstellungen

$$x^1 = \sum_{i \in I} \lambda_i^1 v_i + d^1, \quad x^2 = \sum_{i \in I} \lambda_i^2 v_i + d^2$$

der behaupteten Art. Mit

$$\lambda_i := (1 - \mu)\lambda_i^1 + \mu\lambda_i^2 \quad (i \in I), \quad d := (1 - \mu)d^1 + \mu d^2$$

ist die gewünschte Darstellung

$$x = \sum_{i \in I} \lambda_i v_i + d$$

gefunden.

Nun sei $w \geq 0$. Wie eben definiere man $\delta_1 > 0$ und $x^1 \in M$ durch

$$\delta_1 := \min \left\{ \frac{x_j}{w_j} : j \in B(x), w_j > 0 \right\}, \quad x^1 := x - \delta_1 w.$$

Auf x^1 kann die Induktionsvoraussetzung angewandt werden. Aus $x = x^1 + \delta_1 w$ erhält man die behauptete Darstellung von x . Der Fall $w \leq 0$ kann entsprechend behandelt werden. Damit ist der Induktionsbeweis abgeschlossen.

Ist M beschränkt, so existiert kein $d \neq 0$ mit $d \geq 0$ und $Ad = 0$. Die eben bewiesene Darstellung eines beliebigen Punktes aus M besagt, daß sich jeder Punkt aus M als Konvexitätskombination der endlich vielen Ecken von M darstellen läßt. \square

Nun haben wir die Hilfsmittel beisammen, um den folgenden für das Simplexverfahren grundlegenden Satz zu beweisen.

Satz 2.4 Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $M \neq \emptyset$. Dann gilt: Entweder besitzt (P) eine der endlich vielen Ecken von M als Lösung, oder es ist $\inf(P) = -\infty$, die Zielfunktion von (P) also auf der Menge der zulässigen Lösungen nicht nach unten beschränkt.

Beweis: Existiert ein $d \in \mathbb{R}^n$ mit $d \geq 0$, $Ad = 0$ und $c^T d < 0$, so ist die Zielfunktion von (P) auf der Menge der zulässigen Lösungen nicht nach unten beschränkt. Denn für ein beliebiges $\hat{x} \in M$ ist einerseits $\hat{x} + t d \in M$ für alle $t \geq 0$, andererseits $c^T(\hat{x} + t d) \rightarrow -\infty$ für $t \rightarrow +\infty$.

Daher nehmen wir nun an, für jedes $d \in \mathbb{R}^n$ mit $d \geq 0$ und $Ad = 0$ sei $c^T d \geq 0$. Mit $\{v_i : i \in I\}$ werde die nichtleere, endliche Menge der Ecken von M bezeichnet. Der Darstellungssatz 2.3 impliziert, daß sich ein beliebiges $x \in M$ darstellen läßt als

$$x = \sum_{i \in I} \lambda_i v_i + d$$

mit

$$\lambda_i \geq 0 \quad (i \in I), \quad \sum_{i \in I} \lambda_i = 1 \quad \text{und} \quad d \geq 0, \quad Ad = 0.$$

Dann ist aber

$$c^T x = c^T \left(\sum_{i \in I} \lambda_i v_i + d \right) = \sum_{i \in I} \lambda_i c^T v_i + \underbrace{c^T d}_{\geq 0} \geq \min_{i \in I} c^T v_i,$$

woraus die Behauptung folgt. \square

Zwei wichtige Erkenntnisse erhalten wir aus der Aussage von Satz 2.4. Zum einen wissen wir nun: Besitzt das lineare Programm in Normalform (P) eine Lösung, so besitzt (P) auch eine Ecke von M als Lösung. Da es hiervon nur endlich viele gibt, könnte man im Prinzip alle Ecken von M berechnen und von ihnen eine mit minimalen Kosten bestimmen. Diese naive Vorgehensweise verbietet sich, da M i. allg. sehr viele Ecken besitzt. Zum anderen liefert Satz 2.4 eine *Existenzaussage* für lineare Programme:

- Ist (P) zulässig, d. h. $M \neq \emptyset$, und $\inf(P) := \inf\{c^T x : x \in M\} > -\infty$, so besitzt (P) eine Lösung.

Im folgenden werden wir bei der Untersuchung des linearen Programms in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

bzw. dessen Menge M zulässiger Lösungen o. B. d. A. annehmen, daß

$$(V) \quad \text{Rang}(A) = m,$$

die Zeilen von $A \in \mathbb{R}^{m \times n}$ also linear unabhängig sind. Dies ist *theoretisch* keine Einschränkung. Denn ist die Rang-Voraussetzung (V) nicht erfüllt, so ist entweder das Gleichungssystem $Ax = b$ nicht lösbar und damit $M = \emptyset$, oder es treten redundante Gleichungen unter den Nebenbedingungen $Ax = b$ auf, die entfernt werden können. Man beachte, daß bei dem Produktionsplanungsproblem und dem Diätproblem nach deren Überführung auf Normalform die Rang-Voraussetzung (V) automatisch erfüllt ist, da die resultierende Koeffizientenmatrix durch die Einführung von Schlupfvariablen m Einheitsvektoren (bzw. deren Negatives) enthält.

Die nächste Definition ist grundlegend für das Simplexverfahren.

Definition 2.5 Seien $A = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$ mit $\text{Rang}(A) = m$ und $b \in \mathbb{R}^m$ gegeben. Sei $B \subset \{1, \dots, n\}$ eine Indexmenge mit m Elementen und der Eigenschaft, daß die zu Indizes aus B gehörenden Spalten von A , also $\{a_j\}_{j \in B}$, linear unabhängig sind. Definiert man $x \in \mathbb{R}^n$ durch

$$x_j := 0 \quad \text{für } j \notin B, \quad \sum_{j \in B} x_j a_j = b,$$

so heißt x Basislösung von $Ax = b$ mit den Basisindizes B (oder zur Basis B). Die Basislösung x zur Basis B heißt eine zulässige Basislösung, wenn $x \geq 0$. Eine zulässige Basislösung x zur Basis B heißt nichtentartet, wenn $x_j > 0$ für alle $j \in B$, andernfalls heißt sie entartet.

Der folgende Satz gibt die Verbindung zwischen dem geometrischen Begriff einer Ecke von M und dem algebraischen Begriff einer zulässigen Basislösung.

Satz 2.6 Sei $M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$ mit $A = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$, $\text{Rang}(A) = m$ und $b \in \mathbb{R}^m$. Dann ist $x \in M$ genau dann Ecke von M , wenn x zulässige Basislösung von $Ax = b$ zu einer geeigneten Basis $B \subset \{1, \dots, n\}$ ist.

Beweis: Ist $x \in M$ eine Ecke von M , so sind die zu positiven Komponenten von x gehörenden Spalten von A nach Satz 2.2 linear unabhängig. Sind es weniger als m , so ergänze man sie durch weitere Spalten von A zu m linear unabhängigen Spalten von A , mit B werde die zugehörige Indexmenge bezeichnet. Dann ist x eine zulässige Basislösung von $Ax = b$ zur Basis B . Die Umkehrung ist trivial und folgt direkt aus Satz 2.2. \square

Es ist wichtig, sich den Unterschied zwischen einer nichtentarteten und einer entarteten zulässigen Basislösung bzw. Ecke von M klarzumachen. Ist x eine nichtentartete Ecke, so ist die zugehörige Basis B eindeutig bestimmt, nämlich als Menge derjenigen Indizes, für die die entsprechenden Komponenten von x positiv sind. Dagegen kann es zu einer entarteten Ecke x mit $p < m$ positiven Komponenten bis zu

$$\binom{n-p}{n-m} = \frac{(n-p)!}{(n-m)!(m-p)!}$$

verschiedene Basisdarstellungen ein und derselben entarteten Ecke x geben.

6.2.2 Die Phase II des Simplexverfahrens

Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A = (a_1 \ \cdots \ a_n) \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und die Rang-Voraussetzung

$$(V) \quad \text{Rang}(A) = m$$

erfüllt ist. Wir beschreiben in diesem Unterabschnitt die *Phase II* des Simplexverfahrens. In dieser wird vorausgesetzt, eine Ecke x des Polyeders M bzw. eine zulässige Basislösung zur Basis B sei bekannt. Dagegen dient die *Phase I* dazu, Widersprüche in den Nebenbedingungen bzw. $M = \emptyset$ zu entdecken, die Rang-Voraussetzung (V) zu überprüfen, gegebenenfalls redundante Gleichungen zu entfernen und eine zulässige Ausgangsbasislösung zu bestimmen. Da die Phase I darin besteht, die Phase II auf ein geeignetes Hilfsproblem anzuwenden, für welches trivialerweise die Rang-Voraussetzung (V) erfüllt und eine zulässige Basislösung bekannt ist, beginnen wir mit der Beschreibung der Phase II.

Um weitgehend eine Vektor-Matrix-Schreibweise benutzen zu können, wird es zweckmäßig sein, die folgenden Bezeichnungen zu benutzen.

Ist $B = \{j(1), \dots, j(m)\} \subset \{1, \dots, n\}$, so sei $A_B := (a_{j(1)} \ \cdots \ a_{j(m)})$, d. h. die Spalten von A_B werden aus den zur Indexmenge B gehörenden Spalten von A (in einer festen Reihenfolge) gebildet. Entsprechend sei für einen Vektor $z = (z_j) \in \mathbb{R}^n$ der m -Vektor $z_B := (z_{j(1)}, \dots, z_{j(m)})^T$ definiert. Mit $N := \{1, \dots, n\} \setminus B$ werde stets die Menge der Nichtbasisindizes bezeichnet, $A_N \in \mathbb{R}^{m \times (n-m)}$ und $z_N \in \mathbb{R}^{n-m}$ seien entsprechend zu A_B und z_B definiert.

Nun beschreiben wir die einfache geometrische Idee des Simplexverfahrens.

1. Sei x eine Ecke des Polyeders M .
2. Man bestimme eine von x ausgehende *Abstiegskante*, d. h. eine *Kante* des Polyeders M , längs der die Zielfunktion des linearen Programms (P) echt kleiner wird. Wenn eine solche Abstiegskante nicht existiert, so ist x eine Lösung von (P), STOP.

3. Ist die gefundene Abstiegskante unbeschränkt, so ist die Zielfunktion von (P) auf dem Polyeder M nicht nach unten beschränkt und damit (P) nicht lösbar, STOP.
4. Da die Zielfunktion von (P) längs der gefundenen, von x ausgehenden, Abstiegskante fällt, diese aber beschränkt ist, bestimme man die neue Näherung x^+ als Endpunkt dieser Abstiegskante. Es wird sich herausstellen, daß x^+ eine Ecke des Polyeders M ist.
5. Setze $x := x^+$ und gehe nach 2.

Eine Präzisierung dieser geometrischen Beschreibung¹ des Simplexverfahrens erfolgt im nächsten Satz.

Satz 2.7 Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}.$$

Hierbei sei $c \in \mathbb{R}^n$, $A = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$ mit $\text{Rang}(A) = m$ und $b \in \mathbb{R}^m$. Sei $x \in M$ eine zulässige Basislösung zur Basis $B = \{j(1), \dots, j(m)\}$, der Basisanteil von x also $x_B := A_B^{-1}b$, mit zugehörigen Kosten $c_0 := c_B^T A_B^{-1}b$. Ferner sei $y := A_B^{-T} c_B$ und $N := \{1, \dots, n\} \setminus B$. Dann gilt:

1. Ist $c_N - A_N^T y \geq 0$, so ist x eine Lösung von (P). Ist $c_N - A_N^T y > 0$, so ist x eindeutige Lösung von (P).
2. Ist $c_s - a_s^T y < 0$ und $w := A_B^{-1} a_s \leq 0$ mit einem $s \in N$, so ist $\inf(P) = -\infty$, die Zielfunktion von (P) also auf der Menge der zulässigen Lösungen nicht nach unten beschränkt.
3. Sei $c_s - a_s^T y < 0$ und $w := A_B^{-1} a_s \not\leq 0$ mit einem $s \in N$. Man bestimme ein $r \in \{1, \dots, m\}$ mit $w_r > 0$ und

$$\frac{(A_B^{-1}b)_r}{w_r} = \min_{i=1, \dots, m} \left\{ \frac{(A_B^{-1}b)_i}{w_i} : w_i > 0 \right\} =: \theta^*.$$

Definiert man $x^+ \in \mathbb{R}^n$ durch

$$x_j^+ := \begin{cases} (A_B^{-1}b)_i - \theta^* w_i & \text{für } j = j(i), i \neq r, \\ \theta^* & \text{für } j = s, \\ 0 & \text{für } j \neq s, j \notin B, \end{cases}$$

¹Diese hätte auch etwas blumiger sein können. So findet man z. B. im SPIEGEL (Heft 49, Jahrgang 1984) in einem Artikel über das Karmarkar-Verfahren (siehe Abschnitt 6.4) die folgende Beschreibung des Simplexverfahrens.

Dantzs Methode: Die Milliarden von möglichen Lösungen aus dem gewaltigen Gleichungswust werden gleichsam als Eckpunkte außerordentlich facettenreicher Körper (eigentlich: vieldimensionaler Körper in vieldimensionalen Räumen) betrachtet. Der Weg zur optimalen Lösung nimmt sich wie eine Wanderung über den Monsterkörper aus, wobei jeweils große Gruppen von Eckpunkten dank der Simplex-Methode als für die Lösung untauglich ausgemustert werden—nur mehr für die stark reduzierte Zahl der als lösungsträchtig erkannten Punkte werden Berechnungen ausgeführt.

so ist x^+ eine zulässige Basislösung zur Basis

$$B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\} =: \{j^+(1), \dots, j^+(m)\}$$

mit den Kosten

$$c_0^+ := c^T x^+ = c^T x + \frac{(A_B^{-1} b)_r}{w_r} (c_s - a_s^T y) \leq c^T x = c_0.$$

Insbesondere gilt: Ist x eine nichtentartete Basislösung zur Basis B , so ist x^+ eine zulässige Basislösung zur Basis B^+ mit echt kleineren Kosten, d.h. es ist $c^T x^+ < c^T x$. Ferner ist

$$A_{B^+}^{-1} = \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}.$$

Beweis: Sei $c_N - A_N^T y \geq 0$ und $z \in M$ beliebig. Dann ist $Az = A_B z_B + A_N z_N = b$ und daher $z_B = x_B - A_B^{-1} A_N z_N$. Folglich ist

$$c^T z = c_B^T z_B + c_N^T z_N = c^T x + [c_N - A_N^T y]^T z_N \geq c^T x,$$

also x eine Lösung von (P). Auch die Eindeutigkeitsaussage liest man hieraus ab.

Sei nun $c_s - a_s^T y < 0$ und $w := A_B^{-1} a_s \leq 0$ für ein $s \in N$. Definiert man $x(\theta) \in \mathbb{R}^n$ für $\theta \geq 0$ durch

$$x_j(\theta) := \begin{cases} (A_B^{-1} b)_i - \theta w_i & \text{für } j = j(i) \in B, \\ \theta & \text{für } j = s, \\ 0 & \text{für } j \neq s, j \notin B, \end{cases}$$

so ist $x(\theta) \geq 0$ und

$$Ax(\theta) = A_B(x_B - \theta A_B^{-1} a_s) + \theta a_s = A_B x_B = b,$$

also $x(\theta) \in M$ für alle $\theta \geq 0$. Wegen

$$c^T x(\theta) = c_B^T(x_B - \theta A_B^{-1} a_s) + \theta c_s = c^T x + \theta \underbrace{(c_s - a_s^T y)}_{< 0}$$

folgt mit $\theta \rightarrow +\infty$, daß $\inf(P) = -\infty$.

Sei nun $c_s - a_s^T y < 0$ und $w := A_B^{-1} a_s \not\leq 0$ mit einem $s \in N$. Ferner sei ein $r \in \{1, \dots, m\}$ mit $w_r > 0$ und

$$\frac{(A_B^{-1} b)_r}{w_r} = \min_{i=1, \dots, m} \left\{ \frac{(A_B^{-1} b)_i}{w_i} : w_i > 0 \right\} =: \theta^*$$

bestimmt. Dann ist $x^+ = x(\theta^*) \geq 0$ und $x_{j(r)}^+ = 0$ nach Wahl von r . Ferner ist $Ax^+ = b$ und

$$c_0^+ := c^T x^+ = c^T x + \theta^* (c_s - a_s^T y) = c^T x + \frac{(A_B^{-1} b)_r}{w_r} (c_s - a_s^T y) \leq c^T x = c_0.$$

Erhält man B^+ , wie angegeben, aus B dadurch, daß man $j(r)$ gegen s austauscht, so ist

$$A_{B^+} = \begin{pmatrix} a_{j(1)} & \cdots & a_{j(r-1)} & a_s & a_{j(r+1)} & \cdots & a_{j(m)} \end{pmatrix} = A_B + (a_s - a_{j(r)}) e_r^T.$$

Wegen

$$\sigma := 1 + e_r^T A_B^{-1} (a_s - a_{j(r)}) = 1 + e_r^T (w - e_r) = w_r \neq 0$$

folgt aus der Sherman-Morrison-Formel (siehe Lemma 2.14 in Abschnitt 1.2), daß A_{B^+} nichtsingulär ist und

$$A_{B^+}^{-1} = A_B^{-1} - \frac{1}{w_r} A_B^{-1} (a_s - a_{j(r)}) e_r^T A_B^{-1} = \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}$$

gilt. Wegen $x_j^+ = 0$ für alle $j \notin B^+$ und $x^+ \in M$ ist daher x^+ eine zulässige Basislösung zur Basis B^+ . Insgesamt ist der Satz bewiesen. \square

Wir fassen die Schritte beim *revidierten Simplexverfahren* zusammen.

1. Sei $x \in M$ eine zulässige Basislösung zur Basis $B = \{j(1), \dots, j(m)\}$. Der Basisanteil von x ist $x_B := A_B^{-1} b$, die zugehörigen Kosten sind $c_0 := c_B^T A_B^{-1} b$. Mit $N := \{1, \dots, n\} \setminus B$ sei die Menge der Nichtbasisindizes bezeichnet.
2. Berechne $y := A_B^{-T} c_B$. Anschließend berechne $\bar{c}_j := c_j - a_j^T y$ für $j \in N$.
3. Falls $\bar{c}_j \geq 0$ für alle $j \in N$, dann STOP: x ist Lösung von (P).
4. Wähle $s \in N$ mit $\bar{c}_s < 0$.
- Oft wählt man $s \in N$ so, daß $\bar{c}_s = \min_{j \in N} \{\bar{c}_j : \bar{c}_j < 0\}$, obwohl diese Wahl nicht unbedingt den größtmöglichen Abfall der Zielfunktion gewährleistet.
5. Berechne $w := A_B^{-1} a_s$. Falls $w \leq 0$, dann STOP: Es ist $\inf(P) = -\infty$, die Zielfunktion auf der Menge der zulässigen Lösungen also nicht nach unten beschränkt.
6. Bestimme bzw. wähle $r \in \{1, \dots, m\}$ mit $w_r > 0$ so, daß

$$\frac{(A_B^{-1} b)_r}{w_r} = \min_{i=1, \dots, m} \left\{ \frac{(A_B^{-1} b)_i}{w_i} : w_i > 0 \right\} =: \theta^*.$$

7. Setze $B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\}$, $N^+ := \{1, \dots, n\} \setminus B^+$. Anschließend berechne den Basisanteil $x_{B^+}^+$ der neuen zulässigen Basislösung x^+ durch

$$x_j^+ := \begin{cases} (A_B^{-1} b)_i - \theta^* w_i & \text{für } j = j(i), i \neq r, \\ \theta^* & \text{für } j = s. \end{cases}$$

Dann ist

$$x_{B^+}^+ = A_{B^+}^{-1} b \quad \text{mit} \quad A_{B^+}^{-1} := \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}.$$

Die Kosten von x^+ sind $c_0^+ := c_0 + \theta^* \bar{c}_s$.

8. Vertausche $j(r)$ und s , setze $(x, B, N, c_0) := (x^+, B^+, N^+, c_0^+)$ und gehe zu 2.

Die Hauptarbeit beim revidierten Simplexverfahren besteht darin, die m -Vektoren $y := A_B^{-T} c_B$ und $w := A_B^{-1} a_s$ zu berechnen. Es wäre nicht gescheit, diese Berechnung jedesmal „ad hoc“ zu machen und nicht zu berücksichtigen, daß sich von Schritt zu Schritt die Koeffizientenmatrix A_B nur in einer Spalte verändert. Daher gewinnt man $A_{B^+}^{-1}$ dadurch, daß man A_B^{-1} von links mit der *Gauß-Jordan-Matrix* (hierunter versteht man eine Matrix, die nur in einer Spalte von der Identität abweicht) $E := I - (w - e_r) e_r^T / w_r$ multipliziert. Dies erreicht man durch den folgenden einfachen Prozeß:

- $e_r^T A_{B^+}^{-1} := (1/w_r) e_r^T A_B^{-1}$.

Die r -te Zeile von $A_{B^+}^{-1}$ erhält man also dadurch, daß man die r -te Zeile von A_B^{-1} durch w_r dividiert.

- Für $i = 1, \dots, m, i \neq r$:

$$e_i^T A_{B^+}^{-1} := e_i^T A_B^{-1} - w_i e_r^T A_B^{-1}.$$

Für $i \neq r$ gewinnt man daher die i -te Zeile von $A_{B^+}^{-1}$, indem man von der i -ten Zeile von A_B^{-1} das w_i -fache der r -ten Zeile von A_B^{-1} subtrahiert.

Hat man daher für die Ausgangsbasis B_0 die Inverse $A_{B_0}^{-1}$ berechnet (häufig wird $A_{B_0} = I$ sein), so erhält man

$$A_{B_k}^{-1} = E_k E_{k-1} \cdots E_1 A_{B_0}^{-1}$$

mit Gauß-Jordan-Matrizen E_1, \dots, E_k . Um das Aufsummieren von Rundungsfehlern zu vermeiden, wird man nach einer gewissen Zahl von Schritten die Matrix A_B^{-1} neu berechnen.

Beispiel: Wir betrachten die lineare Optimierungsaufgabe (siehe das Beispiel auf Seite 85)

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei

$$m := 3, \quad n := 5, \quad \begin{array}{c|c} c^T & \\ \hline A & b \end{array} := \begin{array}{ccccc|c} -1 & -1 & 0 & 0 & 0 & \\ \hline 1 & 3 & 1 & 0 & 0 & 13 \\ 3 & 1 & 0 & 1 & 0 & 15 \\ -1 & 1 & 0 & 0 & 1 & 3 \end{array}$$

Die Daten eines Schrittes des revidierten Simplexverfahrens denken wir uns in ein Tableau der Form

s	y^T	c_0
B	A_B^{-1}	x_B

geschrieben. Man erhält die folgenden Tableaus, wobei der zu entfernende Basisindex jeweils eingerahmt ist.

1	0	0	0	0	2	0	$-\frac{1}{3}$	0	-5		$-\frac{1}{4}$	$-\frac{1}{4}$	0	-7
3	1	0	0	13	3	1	$-\frac{1}{3}$	0	8	2	$\frac{3}{8}$	$-\frac{1}{8}$	0	3
4	0	1	0	15	1	0	$\frac{1}{3}$	0	5	1	$-\frac{1}{8}$	$\frac{3}{8}$	0	4
5	0	0	1	3	5	0	$\frac{1}{3}$	1	8	5	$-\frac{1}{2}$	$\frac{1}{2}$	1	4

Das letzte Tableau ist optimal, da $(\bar{c}_3, \bar{c}_4) = (\frac{1}{4}, \frac{1}{4})$. Als Lösung von (P) liest man $x^* := (4, 3, 0, 0, 4)^T$ ab, die zugehörigen Kosten sind $c_0^* := -7$. \square

Nur erwähnen wollen wir eine Implementation des Simplexverfahrens, die insgesamt vor allem aus Stabilitätsgründen vorzuziehen ist und von R. H. BARTELS, G. H. GOLUB (1969) stammt (siehe auch R. H. BARTELS (1971), D. GOLDFARB (1977)). Hierbei geht man davon aus, daß man nur eine *LR*-Zerlegung $PA_B = LR$ von A_B benötigt, um die linearen Gleichungssysteme $A_B^T y = c_B$ und $A_B w = a_s$ zu lösen. Die Idee besteht darin, aus einer *LR*-Zerlegung von A_B eine von A_{B+} zu berechnen. Weitere Literaturhinweise findet man bei D. GOLDFARB, M. J. TODD (1989).

Bemerkung: Wir haben hier eine Form des Simplexverfahrens beschrieben, welche man das *revidierte Simplexverfahren* nennt. In Aufgabe 4 wird dagegen ein Schritt des Simplexverfahrens geschildert, bei dem mit einem sogenannten *vollständigen Tableau* gearbeitet wird. Wegen der einfachen Transformationsregeln (siehe Aufgabe 5) ist diese Form des Simplexverfahrens besonders bei kleinen Beispielen, die man auch ohne Benutzung eines Computers lösen kann, beliebt. Ist aber $n \gg m$, die Zahl der Variablen also wesentlich größer als die Anzahl der Restriktionen, oder ist die Koeffizientenmatrix $A \in \mathbb{R}^{m \times n}$ dünn besetzt, so sollte das revidierte Simplexverfahren vorgezogen werden. \square

6.2.3 Die Vermeidung von Zyklen beim Simplexverfahren

Auf das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

werde die im letzten Unterabschnitt beschriebene Phase II des (revidierten) Simplexverfahrens angewandt.

Ist x eine *nichtentartete* zulässige Basislösung zur Basis B , so werden die Kosten der neuen Basislösung x^+ zur Basis B^+ echt vermindert. Da in jedem Schritt die Kosten zumindestens nicht vergrößert werden, kann man im Verlauf des Simplexverfahrens nicht zu x zurückkehren. Insbesondere ergibt sich hieraus, daß das Simplexverfahren nach endlich vielen Schritten abbrechen muß (mit einer Lösung oder der Information, daß die Zielfunktion auf M nicht nach unten beschränkt ist), wenn alle berechneten zulässigen Basislösungen nichtentartet sind, da es ja nur endlich viele zulässige Basislösungen gibt.

Ist dagegen x eine *entartete* zulässige Basislösung zur Basis $B = \{j(1), \dots, j(m)\}$ und ist

$$\{i \in \{1, \dots, m\} : w_i > 0, (A_B^{-1}b)_i = 0\} \neq \emptyset,$$

so erhält man in Schritt 6 des Simplexverfahrens

$$0 = \frac{(A_B^{-1}b)_r}{w_r} = \min_{i=1, \dots, m} \left\{ \frac{(A_B^{-1}b)_i}{w_i} : w_i > 0 \right\}.$$

In diesem Falle ist $x^+ = x$, man bleibt also in der Ecke x stehen, lediglich die Basissdarstellung von x^+ ist eine andere. Denn B^+ entsteht aus B dadurch, daß $s \notin B$ aufgenommen und $j(r) \in B$ entfernt wird. Dies wäre noch nicht schlimm, kritisch wird es aber dann, wenn man in ein und derselben Ecke stehen bleibt, lediglich von Schritt zu Schritt die Basis austauscht, und nach endlich vielen Schritten zur Ausgangsbasis zurückkehrt. Man spricht dann von einem *Zyklus* im Simplexverfahren. Ein solcher Zyklus kann *theoretisch* auftreten. Dieses Phänomen, das man bisher nur bei eigens hierzu konstruierten Beispielen (siehe E. M. L. BEALE (1955) und z. B. C. H. PAPADIMITRIOU, K. STEIGLITZ (1982, S. 51) sowie Aufgabe 8) beobachtet hat, verhindert, daß das Simplexverfahren ohne *Zusatzregel* ein endliches Verfahren ist. Eine „Anti-Zyklen-Regel“ präzisiert, wie der aufzunehmende Index $s \notin B$ und der zu entfernende Index $j(r) \in B$ zu wählen sind, um die Endlichkeit des Simplexverfahrens zu sichern. Die einfachste Zusatzregel zur Vermeidung von Zyklen stammt von R. G. BLAND (1977) (siehe z. B. auch A. SCHRIJVER (1986, S. 129 ff.) und C. H. PAPADIMITRIOU, K. STEIGLITZ (1982, S. 53 ff.)). Diese Regel besagt, daß man s und $j(r)$ stets als kleinstmöglichen Index wählen sollte. Auf die Bland-Regel, obwohl sie leicht implementierbar ist, wollen wir nicht näher eingehen. Eine Zusatzregel zur Vermeidung von Zyklen spielt für die Praxis keine Rolle und wird i. allg. bei einer Implementation des Simplexverfahrens nicht berücksichtigt. Trotzdem formulieren wir im folgenden Satz eine Zusatzregel für die Phase II des revidierten Simplexverfahrens, welche einen Abbruch nach endlich vielen Schritten sichert. Dieses Ergebnis wird vor allem von theoretischem Interesse sein, da die wichtigen Aussagen der Dualitätstheorie für lineare Optimierungsaufgaben hieraus leicht folgen werden.

Definition 2.8 Ein reeller Vektor $v \neq 0$ heißt *lexikographisch positiv*, wofür wir $v \succ 0$ schreiben, falls die erste von Null verschiedene Komponente von v positiv ist. Sind v und w reelle Vektoren gleicher Länge, so heißt w *lexikographisch größer* als v (bzw. v *lexikographisch kleiner* als w), wofür wir $v \prec w$ schreiben werden, wenn $w - v$ lexikographisch positiv ist.

Eine endliche Menge paarweise verschiedener Vektoren gleicher Länge besitzt offenbar ein eindeutiges lexikographisch kleinstes Element.

Satz 2.9 Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Es sei $M \neq \emptyset$ und $\text{Rang}(A) = m$. Das revidierte Simplexverfahren werde gestartet mit einer zulässigen Basislösung zur Basis

B_0 . Die Zeilen der Matrix $\begin{pmatrix} A_{B_0}^{-1}b & A_{B_0}^{-1} \end{pmatrix} \in \mathbb{R}^{m \times (m+1)}$ seien lexikographisch positiv. (Das ist keine Einschränkung, da o. B. d. A. $A_{B_0} = I$ angenommen werden kann.) Bei der Auswahl der in die aktuelle Basis $B = \{j(1), \dots, j(m)\}$ aufzunehmenden bzw. zu entfernenden Indizes $s \notin B$ bzw. $j(r) \in B$ beachte man die folgende Zusatzregel:

1. Wähle $s \notin B$ beliebig mit $\bar{c}_s := c_s - c_B^T A_B^{-1} a_s < 0$ und $w := A_B^{-1} a_s \not\leq 0$. Es wird also angenommen, daß bei der aktuellen zulässigen Basislösung noch kein Abbruch erfolgt.
2. Sei $B_s := \{i \in \{1, \dots, m\} : w_i > 0\}$, bestimme $r \in B_s$ so, daß

$$\frac{1}{w_r} \begin{pmatrix} (A_B^{-1}b)_r \\ A_B^{-T} e_r \end{pmatrix} \prec \frac{1}{w_i} \begin{pmatrix} (A_B^{-1}b)_i \\ A_B^{-T} e_i \end{pmatrix} \quad \text{für alle } i \in B_s \setminus \{r\}.$$

Durch diese Forderung ist r eindeutig bestimmt, da die Vektoren, von denen der lexikographisch kleinste zu bestimmen ist, offensichtlich paarweise verschieden sind.

Unter dieser Zusatzregel bricht das (revidierte) Simplexverfahren nach endlich vielen Schritten ab, entweder mit einer optimalen zulässigen Basislösung oder der Information, daß die Zielfunktion auf der Menge der zulässigen Lösungen nicht nach unten beschränkt ist.

Beweis: Wir zeigen, daß von einem Simplexschritt zum nächsten die Zeilen der Matrix $\begin{pmatrix} A_B^{-1}b & A_B^{-1} \end{pmatrix}$ lexikographisch positiv bleiben, daß also

$$\begin{pmatrix} (A_B^{-1}b)_i \\ A_B^{-T} e_i \end{pmatrix} \succ 0, \quad i = 1, \dots, m, \implies \begin{pmatrix} (A_{B+}^{-1}b)_i \\ A_{B+}^{-T} e_i \end{pmatrix} \succ 0, \quad i = 1, \dots, m,$$

während

$$\begin{pmatrix} c_{B+}^T A_{B+}^{-1} b \\ A_{B+}^{-T} c_{B+} \end{pmatrix} \prec \begin{pmatrix} c_B^T A_B^{-1} b \\ A_B^{-T} c_B \end{pmatrix}.$$

Hierdurch ist garantiert, daß man nicht zu einer schon berechneten zulässigen Basislösung zurückkehrt, und die Behauptung wird bewiesen sein.

Wegen

$$A_{B+}^{-1} = \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}$$

ist

$$\begin{aligned} (A_{B+}^{-1} b)_i &= \begin{cases} (A_B^{-1} b)_i - w_i \frac{(A_B^{-1} b)_r}{w_r} & \text{für } i \neq r, \\ \frac{(A_B^{-1} b)_r}{w_r} & \text{für } i = r, \end{cases} \\ A_{B+}^{-T} e_i &= \begin{cases} A_B^{-T} e_i - \frac{w_i}{w_r} A_B^{-T} e_r & \text{für } i \neq r, \\ \frac{1}{w_r} A_B^{-T} e_r & \text{für } i = r. \end{cases} \end{aligned}$$

Hieraus liest man ab:

$$\begin{aligned} \begin{pmatrix} (A_{B+}^{-1} b)_r \\ A_{B+}^{-T} e_r \end{pmatrix} &= \underbrace{\frac{1}{w_r}}_{>0} \underbrace{\begin{pmatrix} (A_B^{-1} b)_r \\ A_B^{-T} e_r \end{pmatrix}}_{\succ 0} \succ 0, \\ \begin{pmatrix} (A_{B+}^{-1} b)_i \\ A_{B+}^{-T} e_i \end{pmatrix} &= \underbrace{\frac{w_i}{w_r}}_{>0} \left[\frac{1}{w_i} \begin{pmatrix} (A_B^{-1} b)_i \\ A_B^{-T} e_i \end{pmatrix} - \frac{1}{w_r} \begin{pmatrix} (A_B^{-1} b)_r \\ A_B^{-T} e_r \end{pmatrix} \right] \succ 0 \quad \text{für } i \in B_s \setminus \{r\}, \\ \begin{pmatrix} (A_{B+}^{-1} b)_i \\ A_{B+}^{-T} e_i \end{pmatrix} &= \underbrace{\begin{pmatrix} (A_B^{-1} b)_i \\ A_B^{-T} e_i \end{pmatrix}}_{\succ 0} - \underbrace{\frac{w_i}{w_r} \begin{pmatrix} (A_B^{-1} b)_r \\ A_B^{-T} e_r \end{pmatrix}}_{\geq 0} \succ 0 \quad \text{für } i \in \{1, \dots, m\} \setminus B_s. \end{aligned}$$

Wegen

$$\begin{pmatrix} c_{B+}^T A_{B+}^{-1} b \\ A_{B+}^{-T} c_{B+} \end{pmatrix} = \begin{pmatrix} c_B^T A_B^{-1} b \\ A_B^{-T} c_B \end{pmatrix} + \underbrace{\frac{\bar{c}_s}{w_r} \begin{pmatrix} (A_B^{-1} b)_r \\ A_B^{-T} e_r \end{pmatrix}}_{\leq 0} \prec \begin{pmatrix} c_B^T A_B^{-1} b \\ A_B^{-T} c_B \end{pmatrix}$$

ist der Satz bewiesen. \square

6.2.4 Die Phase I des Simplexverfahrens

Gegeben sei wiederum das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A = (a_{ij}) = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$, $b = (b_i) \in \mathbb{R}^m$ und $c = (c_j) \in \mathbb{R}^n$. Da man notfalls eine Gleichungsrestriktion mit -1 multiplizieren kann, ist o. B. d. A. $b \geq 0$. Ziel der Phase I des Simplexverfahrens ist es, eine zulässige Basislösung für (P) zu berechnen (mit der dann die Phase II gestartet werden kann) bzw. zu entdecken, daß (P) nicht zulässig, also $M = \emptyset$ ist, oder A nicht vollen Rang hat und in diesem Falle redundante Gleichungen zu entfernen. Dieses Ziel wird durch die Anwendung der Phase II des Simplexverfahrens auf ein geeignetes Hilfsproblem erreicht.

Enthält A in den Spalten schon die m Einheitsvektoren des \mathbb{R}^m , so kann die Phase II sofort gestartet werden. Wir nehmen an, das sei nicht der Fall, führen einen Vektor $y \in \mathbb{R}^n$ von sogenannten *künstlichen Variablen* ein, betrachten das lineare Programm in Normalform

$$\text{Minimiere } e^T y \quad \text{auf } \hat{M} := \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{n+m} : \begin{pmatrix} x \\ y \end{pmatrix} \geq 0, \begin{pmatrix} A & I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = b \right\}$$

und nennen dieses (\hat{P}) . Hierbei ist $e := (1, \dots, 1)^T \in \mathbb{R}^m$. Zur Abkürzung setzen wir

$$\hat{A} := \begin{pmatrix} A & I \end{pmatrix} = (a_1 \ \dots \ a_n \ e_1 \ \dots \ e_m) \in \mathbb{R}^{m \times (n+m)}.$$

Mit der Basis $B := \{n+1, \dots, n+m\}$ kann die Phase II des Simplexverfahrens gestartet werden. Da die Zielfunktion von (\hat{P}) auf der Menge der zulässigen Lösungen durch 0 nach unten beschränkt ist, besitzt (\hat{P}) nach Satz 2.4 eine optimale Basislösung zu einer Basis $B = \{j(1), \dots, j(m)\} \subset \{1, \dots, n+m\}$, die wir mit der Phase II des Simplexverfahrens berechnen können. Wir unterscheiden zwei Fälle.

Fall 1: Es ist $\min(\hat{P}) > 0$.

Dann besitzt (P) keine zulässige Lösung, man breche mit einer entsprechenden Meldung ab.

Fall 2: Es ist $\min(\hat{P}) = 0$.

Dann ist gesichert, daß (P) eine zulässige Lösung besitzt. Nun berechne man $r \in \{1, \dots, m\}$ mit $j(r) = \max_{i=1, \dots, m} j(i)$. Ein angenehmer Fall liegt vor, wenn $j(r) \leq n$ bzw. die optimale Basis B keinen zu einer künstlichen Variablen gehörenden Index enthält. Dann ist durch die Anwendung der Phase II des Simplexverfahrens auf (\hat{P}) eine zulässige Basislösung von $Ax = b$ mit dem Basisanteil $x_B = \hat{A}_B^{-1}b = A_B^{-1}b$ sowie die Matrix $\hat{A}_B^{-1} = A_B^{-1}$ berechnet worden. Hiermit kann die Phase II des (revidierten) Simplexverfahrens zur Lösung des eigentlich interessierenden Problems (P) gestartet werden. Ist dagegen $j(r) \in B$ ein künstlicher Basisindex, also $j(r) > n$, so ist die gewonnene optimale Lösung zu (\hat{P}) notwendig entartet, da alle Komponenten der gewonnenen Basislösung zu künstlichen Indizes wegen $\min(\hat{P}) = 0$ verschwinden müssen. Insbesondere ist $(\hat{A}_B^{-1}b)_r = 0$. Die Idee besteht nun darin, den künstlichen Basisindex $j(r)$ gegen ein $s \in \{1, \dots, n\} \setminus B$ auszutauschen oder festzustellen, daß eine Gleichung in $Ax = b$ redundant ist und folglich gestrichen werden kann. Genauer sind die folgenden beiden Fälle möglich.

- (a) Es existiert ein $s \in \{1, \dots, n\} \setminus B$ mit $e_r^T \hat{A}_B^{-1} a_s \neq 0$.

Man setze $w := \hat{A}_B^{-1}a_s$ und $B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\}$. Anschließend berechne man

$$\hat{A}_{B^+}^{-1} := \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) \hat{A}_B^{-1}.$$

An der gewonnenen Basislösung ändert sich hierbei natürlich nichts, d. h. es ist $\hat{A}_{B^+}^{-1}b = \hat{A}_B^{-1}b$, da $(\hat{A}_B^{-1}b)_r = 0$. Zum Schluß setze man $B := B^+$ und prüfe erneut, ob $B \subset \{1, \dots, n\}$ gilt, ob also die künstlichen Indizes aus der Basis vertrieben sind.

- (b) Für alle $s \in \{1, \dots, n\} \setminus B$ ist $e_r^T \hat{A}_B^{-1} a_s = 0$.

Für $j = j(i) \in \{1, \dots, n\} \cap B$ ist $\hat{A}_B^{-1}a_j = e_i$ und daher $e_r^T \hat{A}_B^{-1}a_j = 0$. Insgesamt ist daher die r -te Zeile von $\hat{A}_B^{-1}A$ eine Nullzeile. Folglich ist

$$A^T(\hat{A}_B^{-T}e_r) = 0 \quad \text{bzw.} \quad \sum_{i=1}^m (e_i^T \hat{A}_B^{-T}e_r) A^T e_i = 0,$$

die Zeilen von A sind also linear abhängig. Der künstliche Basisindex $j(r)$ sei durch $j(r) = n + q$ mit $q \in \{1, \dots, m\}$ gegeben. Der Koeffizient von $A^T e_q$, also der q -ten Spalte von A^T bzw. der q -ten Zeile von A , ist

$$e_q^T \hat{A}_B^{-T}e_r = e_r^T \hat{A}_B^{-1}e_q = e_r^T e_r = 1,$$

so daß

$$A^T e_q = - \sum_{\substack{i=1 \\ i \neq q}}^m (e_i^T \hat{A}_B^{-T} e_r) A^T e_i.$$

Daher ist die q -te Zeile von A eine Linearkombination der übrigen Zeilen. Die q -te Gleichung in $Ax = b$ ist also redundant und wird daher gestrichen. Außerdem streicht man in dem Basisanteil $\hat{A}_B^{-1}b$ der aktuellen Basislösung die r -te Komponente (hier stand eine Null) und in \hat{A}_B^{-1} die r -te Zeile und die q -te Spalte (siehe auch Aufgabe 9). Anschließend setzt man

$$B := \{j(1), \dots, j(r-1), j(r+1), \dots, j(m)\}, \quad m := m-1.$$

Da $j(r)$ der *größte* künstliche Basisindex war, ist auch nach dieser Reduktion $B \subset \{1, \dots, n+m\}$. Dann wird erneut geprüft, ob $B \subset \{1, \dots, n\}$, ob also die künstlichen Indizes aus der Basis vertrieben sind.

Auf diese Weise können nach endlich vielen Schritten alle künstlichen Variablen aus der Basis vertrieben und redundante Gleichungen gestrichen werden, so daß man mit einer zulässigen Basislösung zu einer Basis $B \subset \{1, \dots, n\}$ eines eventuell reduzierten linearen Gleichungssystems $Ax = b$ endet. Genauer ist neben dem Basisanteil $x_B = A_B^{-1}b$ auch A_B^{-1} berechnet, dem Übergang zur Phase II des (revidierten) Simplexverfahrens steht nichts mehr im Wege.

Bemerkung: Sind in der Koeffizientenmatrix A von $Ax = b$ mit $b \geq 0$ schon ein oder mehrere Einheitsvektoren enthalten, so kommt man mit entsprechend weniger künstlichen Variablen aus. Eine Modifikation der obigen Vorgehensweise ist ziemlich offensichtlich, siehe auch das folgende Beispiel. \square

Beispiel: Sie wollen Ihrer Tante (vielleicht eine reiche Erbtante?) zum Geburtstag eine Freude machen. Ihre Tante trinkt gerne einen süßen Wein und da Ihnen eine Beerenauslese zu teuer ist, kommen Sie auf die Idee, ihr einen Liter Wein zukommen zu lassen, den Sie selbst zusammengestellt haben.

Hierzu können Sie einen Landwein für 1.00 DM pro Liter, zur Anhebung der Süße Diäthylenglykol-haltiges Frostschutzmittel für 1.20 DM pro Liter und für eine Verbesserung der Lagerungsfähigkeit eine Natriumacid-Lösung für 1.80 DM pro Liter kaufen. Verständlicherweise wollen Sie eine möglichst billige Mischung herstellen, wobei aber folgende Nebenbedingungen zu beachten sind: Um eine hinreichende Süße zu garantieren, muß die Mischung mindestens 1/3 Frostschutzmittel enthalten. Andererseits muß (z. B. wegen gesetzlicher Bestimmungen) mindestens halb so viel Wein wie Frostschutzmittel enthalten sein. Der Natriumacid-Anteil muß mindestens halb so groß, darf aber andererseits höchstens so groß wie der Glykol-Anteil sein und darf die Hälfte des Weinanteils nicht unterschreiten.

Die gesuchte Mischung bestehe aus x_1 Liter Frostschutzmittel, x_2 Liter Natriumacid-Lösung und x_3 Liter Wein. Für eine „zulässige“ Mischung erhalten wir die Nebenbedingungen

$$x_1 + x_2 + x_3 = 1, \quad x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0$$

sowie

$$x_1 \geq \frac{1}{3}, \quad x_3 \geq \frac{1}{2}x_1, \quad \frac{1}{2}x_1 \leq x_2 \leq x_1, \quad \frac{1}{2}x_3 \leq x_2.$$

Die Kosten der Mischung ergeben sich zu $1.2x_1 + 1.8x_2 + x_3$ DM, diese sind zu minimieren. Zu lösen ist also die Optimierungsaufgabe

$$\begin{array}{ll} \text{Minimiere} & \frac{6}{5}x_1 + \frac{9}{5}x_2 + x_3 \text{ unter den Nebenbedingungen} \\ x_1 & - 2x_3 \leq 0 \\ x_1 - 2x_2 & \leq 0 \\ -x_1 + x_2 & \leq 0 \\ -2x_2 + x_3 & \leq 0 \\ 3x_1 & \geq 1 \\ x_1 + x_2 + x_3 & = 1 \end{array}$$

Nach Einführung von Schlupfvariablen hat man die Aufgabe

$$\text{Minimiere } c^T x \text{ unter den Nebenbedingungen } x \geq 0, Ax = b$$

mit

c^T		$\frac{6}{5}$	$\frac{9}{5}$	1	0	0	0	0	0	
A	b	1	0	-2	1	0	0	0	0	0
		1	-2	0	0	1	0	0	0	0
		-1	1	0	0	0	1	0	0	0
		0	-2	1	0	0	0	1	0	0
		3	0	0	0	0	0	0	-1	1
		1	1	1	0	0	0	0	0	1

zu lösen. Die ersten vier Einheitsvektoren sind als Spalten in der Koeffizientenmatrix A schon enthalten, so daß es genügt, zwei künstliche Variable einzuführen. Die Daten des in der Phase I zu lösenden Hilfsproblems sind daher durch

0	0	0	0	0	0	0	0	0	1	1	
1	0	-2	1	0	0	0	0	0	0	0	0
1	-2	0	0	1	0	0	0	0	0	0	0
-1	1	0	0	0	1	0	0	0	0	0	0
0	-2	1	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	-1	1	0	1	
1	1	1	0	0	0	0	0	0	1	1	

gegeben. Schreibt man wie im Beispiel auf Seite 97 die Ergebnisse der (revidierten) Simplexschritte in Tableaus der Form

s	y^T	c_0
B	A_B^{-1}	x_B

so erhält man in der Phase I die folgenden Tableaus:

1	0	0	0	0	1	1	2
4	1	0	0	0	0	0	0
5	0	1	0	0	0	0	0
6	0	0	1	0	0	0	0
7	0	0	0	1	0	0	0
9	0	0	0	0	1	0	1
10	0	0	0	0	0	1	1

3	-4	0	0	0	1	1	2
1	1	0	0	0	0	0	0
5	-1	1	0	0	0	0	0
6	1	0	1	0	0	0	0
7	0	0	0	1	0	0	0
9	-3	0	0	0	1	0	1
10	-1	0	0	0	0	1	1

2	$\frac{1}{2}$	$-\frac{9}{2}$	0	0	1	1	2
1	0	1	0	0	0	0	0
3	$-\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0
6	0	1	1	0	0	0	0
7	$\frac{1}{2}$	$-\frac{1}{2}$	0	1	0	0	0
9	0	-3	0	0	1	0	1
10	$\frac{1}{2}$	$-\frac{3}{2}$	0	0	0	1	1

8	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$-\frac{2}{3}$	1	$\frac{1}{3}$
1	0	0	0	0	$\frac{1}{3}$	0	$\frac{1}{3}$
3	$-\frac{1}{2}$	0	0	0	$\frac{1}{6}$	0	$\frac{1}{6}$
6	0	$\frac{1}{2}$	1	0	$\frac{1}{6}$	0	$\frac{1}{6}$
7	$\frac{1}{2}$	-1	0	1	$\frac{1}{6}$	0	$\frac{1}{6}$
2	0	$-\frac{1}{2}$	0	0	$\frac{1}{6}$	0	$\frac{1}{6}$
10	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$-\frac{2}{3}$	1	$\frac{1}{3}$

Anschließend erhält man das für die Phase I optimale Tableau

	0	0	0	0	0	0	0
1	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
3	$-\frac{3}{8}$	$\frac{1}{8}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
6	$\frac{1}{8}$	$\frac{5}{8}$	1	0	0	$\frac{1}{4}$	$\frac{1}{4}$
7	$\frac{5}{8}$	$-\frac{7}{8}$	0	1	0	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{8}$	$-\frac{3}{8}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
8	$\frac{3}{4}$	$\frac{3}{4}$	0	0	-1	$\frac{3}{2}$	$\frac{1}{2}$

Es ist eine zulässige Basislösung zur keine künstlichen Indizes enthaltenden Basis $B := \{1, 3, 6, 7, 2, 8\}$ berechnet worden. Der Basisanteil ist $x_B := (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2})^T$, ferner ist die Matrix A_B^{-1} ausgegeben worden. Hiermit kann die Phase II des Simplexverfahrens gestartet werden. In zwei Schritten erhält man die optimale Lösung:

4	$\frac{3}{20}$	$-\frac{1}{4}$	0	0	0	$\frac{13}{10}$	$\frac{13}{10}$
1	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
3	$-\frac{3}{8}$	$\frac{1}{8}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
6	$\frac{1}{8}$	$\frac{5}{8}$	1	0	0	$\frac{1}{4}$	$\frac{1}{4}$
7	$\frac{5}{8}$	$-\frac{7}{8}$	0	1	0	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{8}$	$-\frac{3}{8}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
8	$\frac{3}{4}$	$\frac{3}{4}$	0	0	-1	$\frac{3}{2}$	$\frac{1}{2}$

0	$-\frac{1}{25}$	0	$-\frac{6}{25}$	0	$\frac{31}{25}$	$\frac{31}{25}$
1	0	$\frac{3}{5}$	0	$-\frac{2}{5}$	0	$\frac{2}{5}$
3	0	$-\frac{2}{5}$	0	$\frac{3}{5}$	0	$\frac{2}{5}$
6	0	$\frac{4}{5}$	1	$-\frac{1}{5}$	0	$\frac{1}{5}$
4	1	$-\frac{7}{5}$	0	$\frac{8}{5}$	0	$\frac{2}{5}$
2	0	$-\frac{1}{5}$	0	$-\frac{1}{5}$	0	$\frac{1}{5}$
8	0	$\frac{9}{5}$	0	$-\frac{6}{5}$	-1	$\frac{6}{5}$

Zur optimalen Basis $B^* := \{1, 3, 6, 4, 2, 8\}$ ist die Lösung $x^* := (\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{2}{5}, 0, \frac{1}{5}, 0, \frac{1}{5})^T$ mit den zugehörigen Kosten $c_0^* := \frac{31}{25}$ berechnet worden. Die kostenminimale Mischung enthält daher $\frac{2}{5}$ Liter Frostschutzmittel, $\frac{1}{5}$ Liter Natriumacid-Lösung und $\frac{2}{5}$ Liter Wein, das Geburtstagspräsent kostet 1.24 DM. \square

Aufgaben

1. Sei

$$M := \left\{ x \in \mathbb{R}^n : x_j \geq 0 \quad (j = 1, \dots, n_0), \begin{array}{l} \sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m_0), \\ \sum_{j=1}^n a_{ij} x_j = b_i \quad (i = m_0 + 1, \dots, m) \end{array} \right\}$$

nichtleer. Man zeige, daß M genau dann ein Polytop bzw. beschränkt ist, wenn

$$\begin{aligned} d_j &\geq 0 \quad (j = 1, \dots, n_0), & \sum_{j=1}^n a_{ij} d_j &\geq 0 \quad (i = 1, \dots, m_0), \\ && \sum_{j=1}^n a_{ij} d_j &= 0 \quad (i = m_0 + 1, \dots, m) \end{aligned}$$

nur die triviale Lösung $d = 0$ besitzt.

2. Sei $M := \{x \in \mathbb{R}^n : Ax \leq b\}$ mit $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ nichtleer. Man überlege sich, daß M nicht notwendig Ecken besitzt. Anschließend gebe man, ähnlich wie in Satz 2.2, eine Charakterisierung von Ecken der Menge M an.
3. Man zeige, daß ein nichtleeres Polyeder $P \subset \mathbb{R}^n$ genau dann keine Ecke besitzt, wenn es ein $x \in P$ und ein $d \in \mathbb{R}^n \setminus \{0\}$ mit $x + td \in P$ für alle $t \in \mathbb{R}$ gibt, wenn es also eine ganz in P verlaufende Gerade gibt.
4. Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } f(z) := c^T z + c_0 \quad \text{auf } M := \{z \in \mathbb{R}^n : z \geq 0, Az = b\}$$

mit

$$c = (c_j) \in \mathbb{R}^n, \quad c_0 \in \mathbb{R}, \quad A = (a_{ij}) = (\begin{matrix} a_1 & \cdots & a_n \end{matrix}), \quad b = (b_i) \in \mathbb{R}^m.$$

Es wird vorausgesetzt, daß $b \geq 0$. Mit einer Indexmenge $B := \{j(1), \dots, j(m)\}$ sei ferner $A_B = I$ und $c_B = 0$, so daß durch

$$x_j := \begin{cases} b_i & \text{für } j = j(i) \in B, \\ 0 & \text{für } j \notin B \end{cases}$$

eine Ecke von M mit $f(x) = c_0$ gegeben ist. Man zeige:

- (a) Ist $c_N \geq 0$, so ist x eine Lösung von (P) und $\min(P) = c_0$. Ist sogar $c_N > 0$, so ist x eindeutige Lösung von (P).
- (b) Gibt es ein $s \in N$ mit $c_s < 0$ und $a_s \leq 0$, so ist $\inf(P) = -\infty$.

(c) Sei $c_s < 0$ und $a_s \not\leq 0$. Sei ein $r \in \{1, \dots, m\}$ mit $a_{rs} > 0$ und

$$\frac{b_r}{a_{rs}} = \min_{i=1, \dots, m} \left\{ \frac{b_i}{a_{is}} : a_{is} > 0 \right\}$$

bestimmt. Hiermit sei die Gauß-Jordan-Matrix $J_{rs} := I - (a_s - e_r) e_r^T / a_{rs}$ definiert, wobei e_r der r -te Einheitsvektor im \mathbb{R}^m ist, und das transformierte Tableau

$$\left(\begin{array}{c|c} (c^+)^T & -c_0^+ \\ \hline A^+ & b^+ \end{array} \right) := \left(\begin{array}{c|c} 1 & -(c_s/a_{rs})e_r^T \\ \hline 0 & J_{rs} \end{array} \right) \left(\begin{array}{c|c} c^T & -c_0 \\ \hline A & b \end{array} \right)$$

berechnet. Dann ist das lineare Programm in Normalform

$$(P^+) \quad \begin{cases} \text{Minimiere } f^+(z) := (c^+)^T z + c_0^+ \text{ auf} \\ M^+ := \{z \in \mathbb{R}^n : z \geq 0, A^+ z = b^+\} \end{cases}$$

äquivalent zu (P). Genauer ist $M = M^+$ und $f(z) = f^+(z)$ für alle $z \in M = M^+$. Ferner ist $b^+ \geq 0$, mit der Indexmenge

$$B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\} =: \{j^+(1), \dots, j^+(m)\}$$

ist $A_{B^+}^+ = I$ und $c_{B^+}^+ = 0$. Daher ist durch

$$x_j^+ := \begin{cases} b_i^+ & \text{für } j = j^+(i) \in B^+, \\ 0 & \text{für } j \notin B^+ \end{cases}$$

eine Ecke von $M = M^+$ mit

$$f(x^+) = f^+(x^+) = c_0^+ = c_0 + c_s b_r^+ \leq c_0 = f(x)$$

gegeben.

5. Man zeige, daß die Berechnung des transformierten Tableaus

$$\left(\begin{array}{c|c} (c^+)^T & -c_0^+ \\ \hline A^+ & b^+ \end{array} \right) := \left(\begin{array}{c|c} 1 & -(c_s/a_{rs})e_r^T \\ \hline 0 & J_{rs} \end{array} \right) \left(\begin{array}{c|c} c^T & -c_0 \\ \hline A & b \end{array} \right)$$

in Aufgabe 4 durch die folgenden einfachen Transformationsregeln erfolgen kann:

- $(a_{r1}^+ \ \dots \ a_{rn}^+ \ | \ b_r^+) := \frac{1}{a_{rs}} (a_{r1} \ \dots \ a_{rn} \ | \ b_r)$.

Die r -te Zeile in $(A^+ \ | \ b^+)$ erhält man also dadurch, daß man die r -te Zeile von $(A \ | \ b)$ durch das Pivotelement a_{rs} dividiert.

- $(c_1^+ \ \dots \ c_n^+ \ | \ -c_0^+) := (c_1 \ \dots \ c_n \ | \ -c_0) - c_s (a_{r1}^+ \ \dots \ a_{rn}^+ \ | \ b_r^+)$.

Die neue Kostenzeile gewinnt man daher dadurch, daß man von der alten das c_s -fache der r -ten Zeile von $(A^+ \ | \ b^+)$ subtrahiert.

- Für $i = 1, \dots, m, i \neq r$:

$$(a_{i1}^+ \cdots a_{in}^+ \mid b_i^+) := (a_{i1} \cdots a_{in} \mid b_i) - a_{is} (a_{r1}^+ \cdots a_{rn}^+ \mid b_r^+).$$

Für $i \neq r$ erhält man also die i -te Zeile von $(A^+ \mid b^+)$, indem man von der i -ten Zeile von $(A \mid b)$ das a_{is} -fache der r -ten Zeile von $(A^+ \mid b^+)$ subtrahiert.

6. Man wende das auf Aufgabe 4 basierende Simplexverfahren mit vollständigen Tableaus (unter Berücksichtigung der in Aufgabe 5 beschriebenen Transformationsregeln) auf das lineare Programm in Normalform an, dessen Daten durch

c^T		$\begin{array}{ccccc c} -1 & -1 & 0 & 0 & 0 & \\ \hline 1 & 3 & 1 & 0 & 0 & 13 \\ 3 & 1 & 0 & 1 & 0 & 15 \\ -1 & 1 & 0 & 0 & 1 & 3 \end{array}$
A	b	

gegeben sind (siehe das Beispiel auf Seite 85).

Hinweis: Man erhält die folgenden Tableaus, wobei die Pivotelemente eingerahmt sind.

$\begin{array}{ccccc c} -1 & -1 & 0 & 0 & 0 & 0 \end{array}$	$\begin{array}{ccccc c} 0 & -\frac{2}{3} & 0 & \frac{1}{3} & 0 & 5 \end{array}$	$\begin{array}{ccccc c} 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 7 \end{array}$
$\begin{array}{ccccc c} 1 & 3 & 1 & 0 & 0 & 13 \end{array}$	$\begin{array}{ccccc c} 0 & \boxed{\frac{8}{3}} & 1 & -\frac{1}{3} & 0 & 8 \end{array}$	$\begin{array}{ccccc c} 0 & 1 & \frac{3}{8} & -\frac{1}{8} & 0 & 3 \end{array}$
$\boxed{3}$ $\begin{array}{ccccc c} 1 & 0 & 1 & 0 & 0 & 15 \end{array}$	$\begin{array}{ccccc c} 1 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 5 \end{array}$	$\begin{array}{ccccc c} 1 & 0 & -\frac{1}{8} & \frac{3}{8} & 0 & 4 \end{array}$
$\begin{array}{ccccc c} -1 & 1 & 0 & 0 & 1 & 3 \end{array}$	$\begin{array}{ccccc c} 0 & \frac{4}{3} & 0 & \frac{1}{3} & 1 & 8 \end{array}$	$\begin{array}{ccccc c} 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 1 & 4 \end{array}$

Im letzten Tableau ist die Kostenzeile nichtnegativ. Die Komponenten zu den Nichtbasisindizes sind sogar positiv. Daher ist $x^* := (4, 3, 0, 0, 4)^T$ die eindeutige Lösung von (P) mit zugehörigen Kosten $c_0^* := -7$. Die optimalen Basisindizes sind $B^* := \{2, 1, 5\}$. Natürlich hätten wir im ersten Schritt auch die zweite Spalte als Pivotspalte nehmen können.

7. Man beweise die folgende (teilweise) Umkehrung der hinreichenden Optimalitätsbedingung in Satz 2.7: Ist x eine optimale, nichtentartete zulässige Basislösung zur Basis B des linearen Programms in Normalform

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

so ist $c_N - (A_B^{-1} A_N)^T c_B \geq 0$.

8. Man programmiere die Phase II des revidierten Simplexverfahrens zur Lösung eines linearen Programms in Normalform und teste das Programm an folgenden Beispielen,

wobei wir alle relevanten Daten in Tableauform $\begin{array}{c|c} c^T & \\ \hline A & b \end{array}$ angeben.

(a) Das Beispiel von Beale für das Auftreten eines Zyklus:

$-\frac{3}{4}$	20	$-\frac{1}{2}$	6	0	0	0	
$\frac{1}{4}$	-8	-1	9	1	0	0	0
$\frac{1}{2}$	-12	$-\frac{1}{2}$	3	0	1	0	0
0	0	1	0	0	0	1	1

Wählt man den in die Basis aufzunehmenden Index $s \notin B$ so, daß $\bar{c}_s = \min_{j \notin B} \bar{c}_j$ und den zu entfernenden Index $j(r)$ kleinstmöglich, so wird man den folgenden Zyklus von Basisindizes erhalten:

$$\{5, 6, 7\}, \{1, 6, 7\}, \{1, 2, 7\}, \{3, 2, 7\}, \{3, 4, 7\}, \{5, 4, 7\}, \{5, 6, 7\}, \dots$$

(b) Für das erste der beiden folgenden Beispiele kann man zur Kontrolle C. H. PADIMITRIOU, K. STEIGLITZ (1982, S. 30 und S. 59) konsultieren.

0	2	0	1	0	0	1	
1	1	1	1	0	0	0	4
1	0	0	0	1	0	0	2
0	0	1	0	0	1	0	3
0	3	1	0	0	0	1	6

und

-2	-1	0	0	0	
1	1	1	0	0	5
-1	1	0	1	0	0
6	2	0	0	1	21

9. $A \in \mathbb{R}^{m \times m}$ sei eine nichtsinguläre Matrix, deren r -te Spalte der q -te Einheitsvektor im \mathbb{R}^m ist. $A^{qr} \in \mathbb{R}^{(m-1) \times (m-1)}$ entstehe aus A durch Streichen der q -ten Zeile und der r -ten Spalte. Entsprechend entstehe $b^q \in \mathbb{R}^{m-1}$ aus $b \in \mathbb{R}^m$ durch Streichen der q -ten Komponente. Dann gilt:

(a) A^{qr} ist nichtsingulär und $(A^{qr})^{-1} = (A^{-1})^{rq}$.

Die Inverse von A^{qr} erhält man also dadurch, daß man in der Inversen A^{-1} von A die r -te Zeile und die q -te Spalte streicht.

(b) $(A^{qr})^{-1}b^q = (A^{-1}b)^r$.

Die Lösung y des linearen Gleichungssystems $A^{qr}y = b^q$ erhält man also dadurch, daß man in der Lösung x von $Ax = b$ die r -te Komponente streicht.

10. Man programmiere das revidierte Simplexverfahren (Phase I und Phase II) für ein lineares Programm der Form

$$\begin{aligned} \text{Minimiere } c^T x &= \sum_{j=1}^n c_j x_j \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : x_j \geq 0 \quad (j = 1, \dots, n_0), \quad \begin{array}{l} \sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m_0), \\ \sum_{j=1}^n a_{ij} x_j = b_i \quad (i = m_0 + 1, \dots, m) \end{array} \right\}. \end{aligned}$$

Input-Parameter seien also $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, $b = (b_i) \in \mathbb{R}^m$, $c = (c_j) \in \mathbb{R}^n$, ferner die Anzahl der Restriktionen $m \in \mathbb{N}$, die Anzahl der Variablen $n \in \mathbb{N}$ und schließlich $m_0 \in \mathbb{Z}$ mit $0 \leq m_0 \leq m$ und $n_0 \in \mathbb{Z}$ mit $0 \leq n_0 \leq n$. In der Phase I des

Simplexverfahrens kann man berücksichtigen, daß möglicherweise eine Schlupfvariable für eine Ungleichungsrestriktion schon die Rolle einer künstlichen Variablen spielen kann.

Man teste das Programm an der linearen Optimierungsaufgabe

$$\begin{array}{ll} \text{Minimiere} & -x_1 + 2x_2 + 3x_3 \quad \text{unter den Nebenbedingungen} \\ & -x_1 - 3x_2 + 2x_3 \geq -10 \\ & -3x_1 + 2x_2 + 3x_3 \geq 5, \quad x_1, x_2 \geq 0. \\ & 4x_1 - 2x_2 + x_3 = 4 \end{array}$$

6.3 Dualität bei linearen Programmen

Wir gehen wieder aus von einem linearen Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. Da redundante Gleichungen gestrichen werden können, ist o. B. d. A. $\text{Rang}(A) = m$. Wegen der Sätze 2.4 und 2.9 wissen wir: Ist (P) zulässig, d. h. $M \neq \emptyset$, und ist die Zielfunktion von (P) auf der Menge M der zulässigen Lösungen nach unten beschränkt, so bricht das Simplexverfahren (mit Zusatzregel) nach endlich vielen Schritten mit einer optimalen zulässigen Basislösung ab. Hieraus erhält man im nächsten Unterabschnitt sehr leicht die wichtigen Aussagen der Dualitätstheorie für lineare Optimierungsaufgaben. Hieran anschließend wird ein Teil dieser Ergebnisse ökonomisch interpretiert. Zum Schluß gehen wir noch kurz auf das duale Simplexverfahren ein.

6.3.1 Schwacher und starker Dualitätssatz

Dem linearen Programm (P) in Normalform wird ein *duales Programm* (D) mittels

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}$$

zugeordnet. Hierbei heißt N die Menge der *dual zulässigen* Lösungen. Eine Verwechslung der Menge N mit der gleichnamigen Menge von Nichtbasisvariablen kann eigentlich nicht eintreten.

Wir haben einem linearen Programm in Normalform ein duales Programm zugeordnet. Da wir *jedes* lineare Programm in äquivalente Normalform überführen können, ist damit für jedes lineare Programm ein duales Programm erklärt. Insbesondere können wir das duale Programm zu (D) bestimmen. Schreibt man das duale Programm (D) in äquivalenter Normalform, so lautet dieses (nach Einführung von Schlupfvariablen z zu $-A^T y \geq -c$ und nichtnegativer Variablen y^+ und y^- mittels

$y = y^+ - y^-$):

$$\begin{aligned} \text{Minimiere } & \left(\begin{array}{c} -b \\ b \\ 0 \end{array} \right)^T \left(\begin{array}{c} y^+ \\ y^- \\ z \end{array} \right) \quad \text{unter den Nebenbedingungen} \\ & \left(\begin{array}{c} y^+ \\ y^- \\ z \end{array} \right) \geq 0, \quad \left(\begin{array}{ccc} -A^T & A^T & -I \end{array} \right) \left(\begin{array}{c} y^+ \\ y^- \\ z \end{array} \right) = -c. \end{aligned}$$

Das hierzu duale Programm ist

$$\text{Maximiere } (-c)^T x \quad \text{unter den Nebenbedingungen} \quad \left(\begin{array}{c} -A \\ A \\ -I \end{array} \right) x \leq \left(\begin{array}{c} -b \\ b \\ 0 \end{array} \right)$$

und dieses Programm stimmt offenbar genau mit dem Ausgangsprogramm (P) überein. Wir halten fest:

- Das duale Programm zu (D) ist das Ausgangsprogramm (P).

Wir erinnern an Bezeichnungen, die in Abschnitt 6.1 eingeführt wurden, und definieren den *Wert* von (P) bzw. (D) durch

$$\inf(P) := \begin{cases} \inf_{x \in M} c^T x & \text{für } M \neq \emptyset, \\ +\infty & \text{für } M = \emptyset \end{cases} \quad \text{bzw.} \quad \sup(D) := \begin{cases} \sup_{y \in N} b^T y & \text{für } N \neq \emptyset, \\ -\infty & \text{für } N = \emptyset \end{cases}$$

und schreiben $\min(P)$ statt $\inf(P)$ bzw. $\max(D)$ statt $\sup(D)$, wenn (P) bzw. (D) eine Lösung besitzt. Schließlich sagen wir, (P) bzw. (D) sei *zulässig*, wenn $M \neq \emptyset$ bzw. $N \neq \emptyset$.

Eine erste Motivation für die Einführung eines dualen Programms ist durch den folgenden, fast trivialen Satz gegeben.

Satz 3.1 (Schwacher Dualitätssatz) Gegeben seien das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

und das hierzu duale Programm

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Dann gilt: Sind $x \in M$ und $y \in N$, so ist $b^T y \leq c^T x$. Gilt hier Gleichheit, sind also $x^* \in M$, $y^* \in N$ und ist $b^T y^* = c^T x^*$, so ist x^* eine Lösung von (P) und y^* eine Lösung von (D).

Beweis: Sind $x \in M$ und $y \in N$, so ist

$$b^T y = (Ax)^T y = x^T A^T y \leq x^T c = c^T x.$$

Der Rest der Behauptung ist trivial. Denn sind $x^* \in M$, $y^* \in N$ und ist $b^T y^* = c^T x^*$, so ist

$$b^T y \leq c^T x^* = b^T y^* \leq c^T x$$

für beliebige $x \in M$, $y \in N$, so daß x^* eine Lösung von (P) bzw. y^* eine Lösung von (D) ist. \square

Aus dem schwachen Dualitätssatz erhalten wir zwei interessante Aussagen: Zum einen kann man die minimalen Kosten des Ausgangsproblems (P), das wir in Zukunft häufig auch *primales Programm* nennen werden, dadurch nach unten abschätzen, daß man die Zielfunktion des dualen Programms in einem dual zulässigen Punkt $y \in N$ auswertet. Insbesondere ist die Zielfunktion von (P) auf M nach unten beschränkt, wenn $N \neq \emptyset$. Entsprechendes gilt für die Zielfunktion von (D). Wegen Satz 2.4 wissen wir daher:

- Sind (P) und (D) zulässig, so besitzen (P) und (D) jeweils eine Lösung und es gilt

$$-\infty < \max(D) \leq \min(P) < +\infty.$$

Zum anderen liefert der schwache Dualitätssatz eine *hinreichende Optimalitätsbedingung*. Ist nämlich $x^* \in M$ ein primal zulässiger Punkt, zu dem es ein $y^* \in N$ mit $b^T y^* = c^T x^*$ gibt, so ist x^* eine Lösung von (P) (und y^* eine Lösung von (D)). Tiefliegender ist die Aussage, daß hierdurch auch eine *notwendige Optimalitätsbedingung* gegeben ist. Für den Beweis des folgenden Satzes ist es nützlich, sich die zu Beginn von Unterabschnitt 6.2.2 eingeführten Bezeichnungen ins Gedächtnis zurückzurufen, und sich die Aussage von Satz 2.7 noch einmal anzusehen.

Satz 3.2 (Kuhn-Tucker) Gegeben seien das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

und das hierzu duale Programm

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Dann ist $x^* \in M$ genau dann eine Lösung von (P), wenn ein $y^* \in N$ mit $b^T y^* = c^T x^*$ existiert. Dieses y^* ist eine Lösung des dualen Programms (D).

Beweis: Wegen des schwachen Dualitätssatzes bleibt zu zeigen: Ist $x^* \in M$ eine Lösung von (P), so existiert ein $y^* \in N$ mit $b^T y^* = c^T x^*$.

Wir nehmen zunächst an, es sei $\text{Rang}(A) = m$. Da (P) nach Voraussetzung eine Lösung x^* besitzt, bricht das Simplexverfahren (mit Zusatzregel) nach endlich vielen Schritten mit einer optimalen, zulässigen Basislösung $x \in M$ zur Basis B ab, der Basisanteil von x ist $x_B = A_B^{-1}b$. Mit $y^* := A_B^{-T}c_B$ ist daher die Abbruchbedingung $c_N - A_N^T y^* \geq 0$ erfüllt. Wegen

$$A^T y^* = \begin{pmatrix} A_B^T \\ A_N^T \end{pmatrix} y^* = \begin{pmatrix} c_B \\ A_N^T y^* \end{pmatrix} \leq \begin{pmatrix} c_B \\ c_N \end{pmatrix} = c$$

ist $y^* \in N$, also y^* dual zulässig. Ferner ist

$$b^T y^* = c_B^T A_B^{-1} b = c^T x = c^T x^*.$$

Aus dem schwachen Dualitätssatz folgt, daß y^* eine Lösung von (D) ist.

Ist $\text{Rang}(A) = r < m$, so streiche man $m - r$ redundante Gleichungen im Gleichungssystem $Ax = b$, wende auf das so reduzierte (mit (P) aber noch äquivalente) Problem den ersten Teil des Beweises an und setze $y_i^* = 0$ für alle i , die einer gestrichenen Gleichung entsprechen. \square

Bemerkung: Aus dem Beweis des letzten Satzes erkennen wir, daß das revidierte Simplexverfahren bei erfolgreichem Abbruch in einer optimalen, zulässigen Basislösung zur Basis B gleichzeitig durch $y^* := A_B^{-T} c_B$ eine Lösung des dualen Programms (D) berechnet. \square

Beispiel: In einem Beispiel auf Seite 97 hatten wir ein lineares Programm in Normalform mit den Daten

c^T						
A						
b						

$$:= \begin{array}{|c|c|c|c|c|c|c|} \hline & -1 & -1 & 0 & 0 & 0 & | \\ \hline 1 & 3 & 1 & 0 & 0 & 0 & | 13 \\ \hline 3 & 1 & 0 & 1 & 0 & 0 & | 15 \\ \hline -1 & 1 & 0 & 0 & 1 & 0 & | 3 \\ \hline \end{array}$$

mit dem revidierten Simplexverfahren gelöst und das optimale Tableau

	$-\frac{1}{4}$	$-\frac{1}{4}$	0	-7	
2	$\frac{3}{8}$	$-\frac{1}{8}$	0	3	
1	$-\frac{1}{8}$	$\frac{3}{8}$	0	4	
5	$-\frac{1}{2}$	$\frac{1}{2}$	1	4	

erhalten. Hieraus liest man in der ersten Zeile die optimale Lösung $y^* := (-\frac{1}{4}, -\frac{1}{4}, 0)^T$ des dualen Programms ab. \square

Beispiel: Bei gegebenen $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ betrachte man die diskrete, lineare Tschebyscheffsche Approximationsaufgabe, $f(x) := \|Ax - b\|_\infty$ auf dem \mathbb{R}^n zu minimieren. Diese Aufgabe ist (siehe Aufgabe 2 in Abschnitt 6.1) äquivalent dem linearen Programm

(P) Minimiere δ unter den Nebenbedingungen $\delta \geq 0$, $-\delta e \leq Ax - b \leq \delta e$,

wobei $e := (1, \dots, 1)^T \in \mathbb{R}^m$. Nach Einführung von Schlupfvariablen und einer Darstellung der freien Variablen x als Differenz von vorzeichenbeschränkten Variablen erhält man das äquivalente Problem in Normalform mit den Daten

0^T	0^T	1	0^T	0^T		
A	-A	e	-I	0	b	
-A	A	e	0	-I	-b	

Das hierzu duale lineare Programm besitzt, wiederum in Normalform geschrieben, die Daten

(*)	<table border="1" style="border-collapse: collapse; width: 100%; text-align: center;"> <tr><td>$-b^T$</td><td>b^T</td><td>0</td><td></td></tr> <tr><td>A^T</td><td>$-A^T$</td><td>0</td><td>0</td></tr> <tr><td>e^T</td><td>e^T</td><td>1</td><td>1</td></tr> </table>	$-b^T$	b^T	0		A^T	$-A^T$	0	0	e^T	e^T	1	1
$-b^T$	b^T	0											
A^T	$-A^T$	0	0										
e^T	e^T	1	1										

Zur Lösung der diskreten, linearen Tschebyscheffschen Approximationssaufgabe liegt es nahe (siehe M. R. OSBORNE, G. A. WATSON (1967)), das Simplexverfahren auf das duale Programm mit den Daten (*) anzuwenden und aus der ersten Zeile eines optimalen Tableaus eine Lösung des zu (*) dualen Programms abzulesen. Multipliziert man nämlich diese Zeile mit -1 , so erhält man eine Lösung des Ausgangsproblems.

Etwas genauer wollen wir die Aufgabe betrachten, zu vorgegebenen $(t_i, b_i) \in \mathbb{R}^2$, $i = 1, \dots, m$, ein Polynom $p^* \in \Pi_{n-1}$ vom Grad $\leq n - 1$ mit

$$\max_{i=1, \dots, m} |p^*(t_i) - b_i| \leq \max_{i=1, \dots, m} |p(t_i) - b_i| \quad \text{für alle } p \in \Pi_{n-1}$$

zu bestimmen. Mit dem Ansatz $p(t) := \sum_{j=0}^{n-1} x_j t^j$ sowie

$$A := (t_i^j)_{\substack{i=1, \dots, m \\ j=0, \dots, n-1}} \in \mathbb{R}^{m \times n}, \quad b := (b_i)_{i=1, \dots, m} \in \mathbb{R}^m$$

ist eine Einordnung in die oben angegebene diskrete, lineare Tschebyscheffsche Approximationssaufgabe gelungen. Da i. allg. $m \gg n$ sein wird, ist es sehr viel günstiger, das revidierte Simplexverfahren auf das duale Programm anzuwenden.

Wir rechnen ein numerisches Beispiel bei H. R. SCHWARZ (1988, S. 83) nach. Es sei $n := 2$, $m := 4$ und

i	1	2	3	4
t_i	1	2	3	4
b_i	1	3	$\frac{13}{4}$	4

Die Daten des zu lösenden dualen Programms sind dann durch

-1	-3	$-\frac{13}{4}$	-4	1	3	$\frac{13}{4}$	4	0	0
1	1	1	1	-1	-1	-1	-1	0	0
1	2	3	4	-1	-2	-3	-4	0	0
1	1	1	1	1	1	1	1	1	1

gegeben. Zunächst werden zwei künstliche Variable eingeführt und die Phase I des Simplexverfahrens gestartet. Es ergeben sich die folgenden Tableaus.

4	1	1	0	0	5	-4	1	0	0	0	0	0	0	
10	1	0	0	0	4	1	0	0	0	4	$-\frac{1}{3}$	$\frac{1}{3}$	0	0
11	0	1	0	0	11	-4	1	0	0	5	$-\frac{4}{3}$	$\frac{1}{3}$	0	0
9	0	0	1	1	9	-1	0	1	1	9	$\frac{5}{3}$	$-\frac{2}{3}$	1	1

Die beiden künstlichen Variablen sind aus der Basis vertrieben, die Phase II des Simplexverfahrens kann gestartet werden und liefert die folgenden Tableaus.

2	0	-1	0	0		8	1	-2	0	0		$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$-\frac{1}{2}$	
4	$-\frac{1}{3}$	$\frac{1}{3}$	0	0		2	-1	1	0	0		2	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
5	$-\frac{4}{3}$	$\frac{1}{3}$	0	0		5	-2	1	0	0		5	-1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
9	$\frac{5}{3}$	$-\frac{2}{3}$	1	1		9	3	-2	1	1		8	$\frac{1}{2}$	$-\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

Daher ist

$$p^*(t) := \frac{1}{2} + t, \quad \max_{i=1, \dots, 4} |p^*(t_i) - b_i| = \frac{1}{2},$$

das gesuchte lineare Polynom ist gefunden. \square

Aus dem Satz von Kuhn-Tucker folgt sehr einfach die Aussage des *starken Dualitätssatzes* der linearen Optimierung.

Satz 3.3 (Starker Dualitätssatz) Gegeben seien das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

und das hierzu duale Programm

$$(D) \quad \text{Maximiere } b^T y \text{ auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Dann gilt:

1. Sind (P) und (D) zulässig, so besitzen (P) und (D) jeweils eine optimale Lösung und es ist $\max(D) = \min(P)$.
2. Ist (P) zulässig, aber (D) nicht zulässig, so ist $\inf(P) = -\infty$.
3. Ist (D) zulässig, aber (P) nicht zulässig, so ist $\sup(D) = +\infty$.

Beweis: Durch den Existenzsatz 2.4, den schwachen Dualitätssatz und den Satz von Kuhn-Tucker sind alle benötigten Hilfsmittel für den Beweis des starken Dualitätssatzes bereitgestellt.

Sind (P) und (D) zulässig, so folgt aus dem schwachen Dualitätssatz, daß

$$-\infty < \sup(D) \leq \inf(P) < +\infty.$$

Also ist insbesondere (P) zulässig und $\inf(P) > -\infty$, aus Satz 2.4 folgt die Existenz einer Lösung $x^* \in M$ von (P). Wegen des Satzes von Kuhn-Tucker gibt es eine Lösung $y^* \in N$ von (D) mit

$$\max(D) = b^T y^* = c^T x^* = \min(P).$$

Sei (P) zulässig, aber (D) nicht zulässig. Wäre $\inf(P) > -\infty$, so hätte (P) eine Lösung und wegen des Satzes von Kuhn-Tucker wäre (D) insbesondere zulässig, ein Widerspruch.

Sei (D) zulässig, aber (P) nicht zulässig. Beachtet man, daß man (P) als duales Programm zu (D) auffassen kann, so folgt $\sup(D) = +\infty$ aus dem gerade eben bewiesenen Teil des Satzes. \square

Eine einfache Folgerung aus dem starken Dualitätssatz ist

Lemma 3.4 (Farkas) Das System

$$(I) \quad Ax = b, \quad x \geq 0$$

besitzt genau dann keine Lösung, wenn das System

$$(II) \quad A^T y \leq 0, \quad b^T y > 0$$

eine Lösung besitzt.

Beweis: Man betrachte die zueinander dualen linearen Programme

$$(P) \quad \text{Minimiere } 0^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

und

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq 0\}.$$

Wegen $0 \in N$ ist (D) zulässig. (I) hat genau dann keine Lösung, wenn (P) nicht zulässig ist. Dies ist wegen des starken Dualitätssatzes genau dann der Fall, wenn $\sup(D) = +\infty$, was wiederum äquivalent der Existenz einer Lösung von (II) ist. \square

Bemerkung: Das Farkas-Lemma hat eine schöne geometrische Interpretation. Hierzu beachte man, daß das System

$$(I) \quad Ax = b, \quad x \geq 0$$

genau dann keine Lösung besitzt, wenn $b \notin K := \{Ax : x \geq 0\}$, wenn b sich also nicht als nichtnegative Linearkombination der Spalten von A darstellen läßt. Das Farkas-Lemma sagt aus, daß aus der Unlösbarkeit von (I) die Existenz einer Lösung $y \in \mathbb{R}^m$ von

$$(II) \quad A^T y \leq 0, \quad b^T y > 0$$

folgt. Definiert man mit diesem y die Hyperebene $H := \{z \in \mathbb{R}^m : y^T z = 0\}$ durch den Nullpunkt, bezeichnet man ferner mit $H^- := \{z \in \mathbb{R}^m : y^T z \leq 0\}$ einen zugehörigen Halbraum, so ist $K \subset H^-$ und $b \notin H^-$. Abbildung 6.3 verdeutlicht diese Aussage. Bemerkt sei lediglich, daß auch der starke Dualitätssatz einleuchtend geometrisch interpretiert werden kann (siehe J. WERNER (1984)). \square

Bisher sind wir stets von einem linearen Programm (P) in Normalform ausgegangen und haben zunächst den Existenzsatz 2.4 und dann mit dem zugehörigen dualen Programm (D) der Reihe nach den schwachen Dualitätssatz, den Satz von Kuhn-Tucker und den starken Dualitätssatz formuliert und bewiesen. Ferner haben wir nachgewiesen, daß das zu (D) duale Programm wieder auf das primale Programm (P) führt. Da sich jedes lineare Programm auf äquivalente Normalform bringen läßt, gelten alle diese Aussagen für allgemeine lineare Optimierungsaufgaben. Ist etwa das

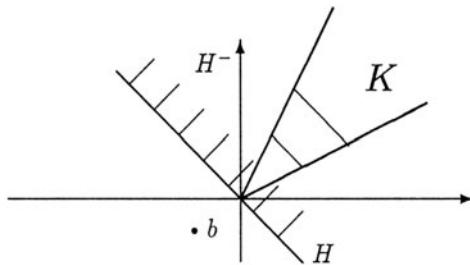


Abbildung 6.3: Geometrische Interpretation des Farkas-Lemmas

lineare Programm (P) durch

$$\text{Minimiere } \sum_{j=1}^n c_j x_j \text{ auf}$$

$$M := \left\{ x \in \mathbb{R}^n : x_j \geq 0 \quad (j = 1, \dots, n_0), \begin{array}{l} \sum_{j=1}^n a_{ij} x_j \geq b_i \quad (i = 1, \dots, m_0), \\ \sum_{j=1}^n a_{ij} x_j = b_i \quad (i = m_0 + 1, \dots, m) \end{array} \right\}$$

gegeben, so erhält man (nach Überführung von (P) in Normalform und leichter Rechnung) als zu (P) duales Programm (D) die Aufgabe

$$\text{Maximiere } \sum_{i=1}^m b_i y_i \text{ auf}$$

$$N := \left\{ y \in \mathbb{R}^m : y_i \geq 0 \quad (i = 1, \dots, m_0), \begin{array}{l} \sum_{i=1}^m a_{ji} y_i \leq c_j \quad (j = 1, \dots, n_0), \\ \sum_{i=1}^m a_{ji} y_i = c_j \quad (j = n_0 + 1, \dots, n) \end{array} \right\}.$$

Für den Fall $m_0 = m$ und $n_0 = n$ (d. h. im Ausgangsproblem (P) sind alle Nebenbedingungen Ungleichungen und alle Variablen vorzeichenbeschränkt) wollen wir den Satz von Kuhn-Tucker noch einmal formulieren.

Satz 3.5 Gegeben seien das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax \geq b\}$$

und das hierzu duale Programm

$$(D) \quad \text{Maximiere } b^T y \text{ auf } N := \{y \in \mathbb{R}^m : y \geq 0, A^T y \leq c\}.$$

Dann ist $x^* \in M$ genau dann eine Lösung von (P), wenn ein $y^* \in N$ mit $b^T y^* = c^T x^*$ bzw.

$$(c - A^T y^*)^T x^* = 0, \quad (Ax^* - b)^T y^* = 0$$

existiert. Dieses y^* ist Lösung des dualen Programms (D).

Beweis: Zu zeigen bleibt lediglich, daß für $x^* \in M$ und $y^* \in N$ die Gleichung $b^T y^* = c^T x^*$ äquivalent zu $(c - A^T y^*)^T x^* = 0$ und $(Ax^* - b)^T y^* = 0$ ist. Wegen

$$c^T x^* - b^T y^* = \underbrace{(c - A^T y^*)^T}_{\geq 0} \underbrace{x^*}_{\geq 0} + \underbrace{(Ax^* - b)^T}_{\geq 0} \underbrace{y^*}_{\geq 0}$$

ist dies aber trivial. \square

6.3.2 Ökonomische Interpretation der Dualität

In Abschnitt 6.1 wurde das Produktionsplanungsproblem auf Seite 82 angegeben: Unter Kapazitätsbeschränkungen an benötigte Hilfsmittel will ein Betrieb einen Produktionsplan aufstellen, der maximalen Gesamtgewinn liefert. Als lineares Programm lautet es:

$$\text{Maximiere } p^T x \text{ unter den Nebenbedingungen } x \geq 0, Ax \leq b.$$

Das hierzu duale Programm ist

$$\text{Minimiere } b^T y \text{ unter den Nebenbedingungen } y \geq 0, A^T y \geq p.$$

Dieses duale Programm kann man folgendermaßen interpretieren: Ein Konkurrent macht dem Betrieb das Angebot, alle m Hilfsmittel zu mieten bzw. zu kaufen und bietet hierzu für das i -te Hilfsmittel $y_i \geq 0$ Geldeinheiten pro Einheit. Insgesamt sind seine Kosten $\sum_{i=1}^m b_i y_i$ und diese wird er versuchen zu minimieren. Der Betrieb geht auf diesen Vorschlag allerdings nur ein, wenn $\sum_{i=1}^m a_{ij} y_i \geq p_j$ für $j = 1, \dots, n$, d. h. wenn der gezahlte Preis für sämtliche Hilfsmittel zur Produktion einer Einheit des j -ten Produktes nicht kleiner ist als der Reingewinn p_j , den der Betrieb erhalten hätte, wenn er die Produktion selbst durchführen würde. Der Konkurrent hat also genau das duale Problem zu lösen.

Der schwache Dualitätssatz kann so interpretiert werden: Ist $x \in \mathbb{R}^n$ ein zulässiger Produktionsplan für den Betrieb und $y \in \mathbb{R}^m$ ein akzeptables Angebot des Konkurrenten, so ist $p^T x \leq b^T y$. D. h. der Betrieb kann keinen größeren Reingewinn machen als den Betrag, den er bei einem akzeptablen Angebot vom Konkurrenten erhalten würde (er erspart sich sogar die Suche nach einem optimalen Produktionsplan). Dagegen sagt der starke Dualitätssatz aus, daß der maximale Reingewinn des Betriebes gleich den minimalen Kosten des Konkurrenten sind (wenn zulässige Produktionspläne und akzeptable Angebote existieren). Ferner sagt Satz 3.5 aus: Ist x^* ein optimaler, zulässiger Produktionsplan mit einem Reingewinn $p^T x^*$, so gibt es ein für den Konkurrenten optimales, zulässiges Angebot mit den Kosten $b^T y^* = p^T x^*$. Notwendigerweise gelten die sogenannten *Gleichgewichtsbedingungen*

$$(A^T y^* - p)^T x^* = 0, \quad (b - Ax^*)^T y^* = 0.$$

Wird in einem optimalen Produktionsplan x^* das j -te Produkt hergestellt, ist also $x_j^* > 0$, so ist notwendig $(A^T y^*)_j = p_j$. Wird die Kapazitätsbeschränkung für das i -te Hilfsmittel in einem optimalen Produktionsplan nicht voll ausgeschöpft, ist also

$(Ax^*)_i < b_i$, so ist notwendig $y_i^* = 0$, der Konkurrent wird daher in einem für ihn optimalen Angebot für das i -te Hilfsmittel nichts bezahlen.

Eine ähnliche Interpretation des dualen Programms und der Ergebnisse der Dualitätstheorie ist für sehr viele aus den Wirtschaftswissenschaften stammenden Problemstellungen, die auf lineare Optimierungsaufgaben führen, möglich.

Wir wollen nun eine weitere Interpretation der Lösung eines dualen Programms angeben. Wir gehen hierzu von einem linearen Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

aus, setzen Rang $(A) = m$ voraus und nehmen an, das Simplexverfahren breche mit einer *nichtentarteten* optimalen, zulässigen Basislösung x^* zur Basis B ab, der Basisanteil $x_B^* := A_B^{-1}b$ sei also ein positiver Vektor. Wir wissen (siehe Beweis des Satzes von Kuhn-Tucker), daß durch $y^* = A_B^{-T}c_B$ eine Lösung des dualen Problems

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}$$

gegeben ist. Ist $\Delta b \in \mathbb{R}^m$ eine so kleine Störung von b , daß noch $A_B^{-1}(b + \Delta b) \geq 0$, so ist

$$\hat{x} := \begin{pmatrix} A_B^{-1}(b + \Delta b) \\ 0 \end{pmatrix}$$

eine optimale, zulässige Basislösung des gestörten Problems

$$\text{Minimiere } c^T x \quad \text{unter den Nebenbedingungen } x \geq 0, Ax = b + \Delta b.$$

Das folgt einfach aus dem schwachen Dualitätssatz, wenn man

$$c^T \hat{x} = c_B^T A_B^{-1}(b + \Delta b) = (b + \Delta b)^T y^*$$

berücksichtigt. Insbesondere ist $c^T \hat{x} = c^T x^* + (y^*)^T \Delta b$ für alle hinreichend kleinen Δb . Die Lösung y^* des dualen Programms (D) zum ungestörten Programm (P) bestimmt daher die *Sensitivität* des Wertes des Ausgangsprogramms gegenüber Störungen der rechten Seite b (siehe auch Aufgabe 6). In den Wirtschaftswissenschaften nennt man y^* den Vektor der *Schattenpreise*.

6.3.3 Das duale Simplexverfahren

In diesem Unterabschnitt betrachten wir wieder das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\},$$

wobei $A = (a_1 \ \cdots \ a_n) \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Es wird Rang $(A) = m$ vorausgesetzt.

Beim (primalen) revidierten Simplexverfahren ging man von einer zulässigen Basislösung x zu der Basis B aus und berechnete $y := A_B^{-T}c_B$. Dann ist $b^T y = c^T x$. Ist also y zulässig für das duale Programm

$$(D) \quad \text{Maximiere } b^T y \quad \text{auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\},$$

bzw. der Optimalitätstest

$$\bar{c}_N := c_N - (A_B^{-1} A_N)^T c_B = c_N - A_N^T y \geq 0$$

erfüllt, so ist x eine Lösung von (P) und y eine Lösung von (D).

Dagegen geht man beim *dualen Simplexverfahren* von einer (i. allg. nicht zulässigen) Basislösung x zu einer Basis B aus (der Basisanteil $x_B := A_B^{-1}b$ ist also i. allg. kein nichtnegativer Vektor), für die $y := A_B^{-T} c_B$ dual zulässig ist, also $c_N - A_N^T y \geq 0$ genügt. Nach wie vor ist $c^T x = b^T y$. Genau diese Situation bleibt im Algorithmus erhalten und der duale Zielfunktionswert wird von Schritt zu Schritt zumindestens nicht verkleinert. Sobald man zu einer primal zulässigen Basislösung kommt, endet das Verfahren.

Beispiel: Das duale Simplexverfahren ist sofort anwendbar auf ein lineares Programm der Form

$$\text{Minimiere } c^T x \text{ unter den Nebenbedingungen } Ax \leq b, x \geq 0$$

mit einem *nichtnegativen* Kostenvektor c . Denn nach der Überführung auf Normalform können die Schlupfvariablen als Basisvariablen genommen werden. I. allg. wird hier b *kein* nichtnegativer Vektor sein (sonst wäre nämlich schon $x^* = 0$ eine Lösung). Durch die Anwendung des dualen Simplexverfahrens auf eine Aufgabe vom obigen Typ erspart man sich daher die Phase I des primalen Simplexverfahrens. \square

Satz 2.7 ist Grundlage für das (primale) revidierte Simplexverfahren. Entsprechend werden im folgenden Satz der Optimalitäts- und der Unlösbarkeitstest sowie ein Schritt beim revidierten dualen Simplexverfahren beschrieben.

Satz 3.6 Gegeben sei das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}.$$

Hierbei sei $c \in \mathbb{R}^n$, $A = (a_1 \ \dots \ a_n) \in \mathbb{R}^{m \times n}$ mit $\text{Rang}(A) = m$ und $b \in \mathbb{R}^m$. Sei x eine (i. allg. nicht zulässige) Basislösung zur Basis $B = \{j(1), \dots, j(m)\}$, der Basisanteil von x also $x_B := A_B^{-1}b$. Mit $N := \{1, \dots, n\} \setminus B$ werde die Menge der Nichtbasisindizes bezeichnet. Sei $y := A_B^{-T} c_B$ dual zulässig, also $\bar{c}_N := c_N - A_N^T y \geq 0$. Dann gilt:

1. Ist $A_B^{-1}b \geq 0$, also x eine zulässige Basislösung, so ist x eine Lösung von (P) (und y eine Lösung zum zu (P) dualen Programm (D)).
2. Ist $(A_B^{-1}b)_r < 0$ und $A_N^T A_B^{-T} e_r \geq 0$ mit einem $r \in \{1, \dots, m\}$, so ist $\sup(D) = +\infty$ und daher (P) nicht zulässig.
3. Sei $(A_B^{-1}b)_r < 0$ und $v_N := A_N^T A_B^{-T} e_r =: (v_j)_{j \in N} \not\geq 0$ mit einem $r \in \{1, \dots, m\}$. Man bestimme ein $s \in N$ mit $v_s < 0$ und

$$\frac{\bar{c}_s}{v_s} = \max_{j \in N} \left\{ \frac{\bar{c}_j}{v_j} : v_j < 0 \right\} =: \gamma^*,$$

anschließend berechne man $w := A_B^{-1}a_s$. Definiert man $x^+ \in \mathbb{R}^n$ durch

$$x_j^+ := \begin{cases} (A_B^{-1}b)_j - \frac{(A_B^{-1}b)_r}{w_r} w_i & \text{für } j = j(i), i \neq r, \\ \frac{(A_B^{-1}b)_r}{w_r} & \text{für } j = s, \\ 0 & \text{für } j \neq s, j \notin B, \end{cases}$$

so ist x^+ eine Basislösung zur Basis

$$B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\} =: \{j^+(1), \dots, j^+(m)\}.$$

Ferner ist

$$y^+ := A_{B^+}^{-T}c_{B^+} = y + \gamma^* A_B^{-T}e_r$$

dual zulässig und

$$c_0^+ := c^T x^+ = b^T y^+ = b^T y + \gamma^* (A_B^{-1}b)_r \geq b^T y = c^T x =: c_0.$$

Beweis: Ist $A_B^{-1}b \geq 0$, so ist die Basislösung x mit dem Basisanteil $x_B := A_B^{-1}b$ primal zulässig. Nach Voraussetzung ist $y := A_B^{-T}c_B$ dual zulässig. Da ferner

$$b^T y = c_B^T x_B = c^T x$$

gilt, folgt aus dem schwachen Dualitätssatz, daß x eine Lösung von (P) und y eine Lösung des dualen Programms (D) ist.

Besitzt $A_B^{-1}b$ eine negative Komponente $(A_B^{-1}b)_r < 0$ und ist $A_N^T A_B^{-T} e_r \geq 0$, so ist einerseits

$$y(\gamma) := y + \gamma A_B^{-T}e_r$$

für alle $\gamma \leq 0$ wegen

$$A^T y(\gamma) = \begin{pmatrix} c_B + \gamma e_r \\ A_N^T y + \gamma A_N^T A_B^{-T} e_r \end{pmatrix} \leq \begin{pmatrix} c_B \\ c_N \end{pmatrix} = c$$

dual zulässig, andererseits ist

$$b^T y(\gamma) = b^T y + \gamma b^T A_B^{-T} e_r = b^T y + \gamma \underbrace{(A_B^{-1}b)_r}_{< 0} \rightarrow +\infty \quad \text{für } \gamma \rightarrow -\infty,$$

also $\sup(D) = +\infty$ und daher (P) nicht zulässig.

Sei nun $(A_B^{-1}b)_r < 0$ und $v_N := A_N^T A_B^{-T} e_r =: (v_j)_{j \in N} \not\geq 0$ mit $r \in \{1, \dots, m\}$. Wählt bzw. bestimmt man $s \in N$ auf die angegebene Weise mit $v_s < 0$ und

$$\frac{\bar{c}_s}{v_s} = \max_{j \in N} \left\{ \frac{\bar{c}_j}{v_j} : v_j < 0 \right\} =: \gamma^*,$$

so ist die r -te Komponente des Vektors $w := A_B^{-1}a_s$ negativ wegen

$$w_r = e_r^T A_B^{-1} a_s = v_s < 0.$$

Die neuen Basisindizes B^+ gewinnt man aus B , indem man $j(r)$ durch s ersetzt. Daher erhält man A_{B^+} aus A_B dadurch, daß die r -te Spalte $a_{j(r)}$ durch a_s ersetzt wird. Genau wie im Beweis von Satz 2.7 folgt wegen $w_r \neq 0$, daß mit A_B auch A_{B^+} nichtsingulär ist, und die Inverse von A_{B^+} nach der Sherman-Morrison-Formel durch

$$A_{B^+}^{-1} = \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}$$

gegeben ist. Insbesondere ist x^+ mit dem Basisanteil $x_{B^+}^+ := A_{B^+}^{-1} b$ eine Basislösung. Den neuen Vektor $y^+ := A_{B^+}^{-T} c_{B^+}$ berechnet man aus

$$y^+ = A_B^{-T} \left(I - \frac{e_r (w - e_r)^T}{w_r} \right) [c_B + (c_s - c_{j(r)}) e_r] = y + \frac{c_s - a_s^T y}{v_s} A_B^{-T} e_r = y + \gamma^* A_B^{-T} e_r.$$

Die neuen Nichtbasisindizes sind $N^+ := (N \setminus \{s\}) \cup \{j(r)\}$. Um nachzuweisen, daß y^+ dual zulässig ist, hat man $c_j - a_j^T y^+ \geq 0$ für alle $j \in N^+$ zu beweisen. Nun ist

$$c_{j(r)} - a_{j(r)}^T y^+ = \underbrace{c_{j(r)} - a_{j(r)}^T y}_{=0} - \gamma^* \underbrace{a_{j(r)}^T A_B^{-T} e_r}_{=1} = -\gamma^* \geq 0$$

und

$$c_j - a_j^T y^+ = c_j - a_j^T y - \gamma^* a_j^T A_B^{-T} e_r = \bar{c}_j - \gamma^* v_j \geq 0 \quad \text{für } j \in N^+ \setminus \{j(r)\}$$

nach Definition von γ^* . Also ist y^+ dual zulässig. Schließlich ist offensichtlich

$$c_0 := c^T x^+ = b^T y^+ = b^T y + \gamma^* (A_B^{-1} b)_r \geq b^T y = c^T x =: c_0,$$

der Satz ist bewiesen. \square

Wir fassen die Schritte beim (revidierten) dualen Simplexverfahren zusammen.

- Sei x eine Basislösung zur Basis $B = \{j(1), \dots, j(m)\}$, der Basisanteil von x also $x_B := A_B^{-1} b$. Mit $N := \{1, \dots, n\} \setminus B$ werde die Menge der Nichtbasisindizes bezeichnet. Der Vektor $y := A_B^{-T} c_B$ sei dual zulässig, d.h es ist $\bar{c}_N \geq 0$ mit $\bar{c}_j := c_j - a_j^T y$ für $j \in N$. Die Kosten von x sind gegeben durch $c_0 := c^T x = b^T y$.

- Falls $x_B \geq 0$ bzw. x primal zulässig ist, dann STOP: x ist eine Lösung von (P) (und y eine Lösung des hierzu dualen linearen Programms (D)).

- Wähle $r \in \{1, \dots, m\}$ mit $(A_B^{-1} b)_r < 0$.

Oft wählt man $r \in \{1, \dots, m\}$ so, daß $(A_B^{-1} b)_r = \min_{i=1, \dots, m} (A_B^{-1} b)_i$, obwohl diese Wahl nicht unbedingt den größtmöglichen Anstieg der Zielfunktion garantiert.

- Berechne $u := A_B^{-T} e_r$ und $v_N := A_N^T u$, also $v_j := a_j^T u$ für $j \in N$. Falls $v_N \geq 0$, dann STOP: Das lineare Programm (P) ist nicht zulässig, da $\sup(D) = +\infty$.

- Bestimme bzw. wähle $s \in N$ mit $v_s < 0$ so, daß

$$\frac{\bar{c}_s}{v_s} = \max_{j \in N} \left\{ \frac{\bar{c}_j}{v_j} : v_j < 0 \right\} =: \gamma^*.$$

6. Setze $B^+ := \{j(1), \dots, j(r-1), s, j(r+1), \dots, j(m)\}$, $N^+ := \{1, \dots, n\} \setminus B^+$
 Anschließend berechne man (genau wie beim primalen Simplexverfahren)

$$w := A_B^{-1} a_s, \quad A_{B^+}^{-1} := \left(I - \frac{(w - e_r) e_r^T}{w_r} \right) A_B^{-1}.$$

Den Basisanteil $x_{B^+}^+ := A_{B^+}^{-1} b$ der neuen Basislösung x^+ berechnet man aus

$$x_j^+ := \begin{cases} (A_B^{-1} b)_i - \frac{(A_B^{-1} b)_r}{w_r} w_i & \text{für } j = j(i), i \neq r, \\ \frac{(A_B^{-1} b)_r}{w_r} & \text{für } j = s. \end{cases}$$

Die neue dual zulässige Lösung $y^+ := A_{B^+}^{-1} c_{B^+}$ sowie die Kosten $c_0^+ := c^T x^+$ erhält man aus

$$y^+ := y + \gamma^* u, \quad c_0^+ := c_0 + \gamma^* (A_B^{-1} b)_r.$$

Schließlich berechne man $\bar{c}_{N^+} := c_{N^+} - A_{N^+}^T y^+ =: (\bar{c}_j)_{j \in N^+}$ durch

$$\bar{c}_j := \begin{cases} -\gamma^* & \text{für } j = j(r), \\ \bar{c}_j - \gamma^* v_j & \text{für } j \in N^+ \setminus \{j(r)\}. \end{cases}$$

7. Vertausche $j(r)$ und s , setze $(x, y, B, N, c_0) := (x^+, y^+, B^+, N^+, c_0^+)$ und gehe zu 2.

Beispiel: Gegeben sei das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

mit den Daten

c^T			5	3	3	6	0	0	0		
A	b		-6	1	2	4	1	0	0	14	
			3	-2	-1	-5	0	1	0	-25	
			-2	1	0	2	0	0	1	14	

Mit $B := \{5, 6, 7\}$ und $N := \{1, 2, 3, 4\}$ ist $c_B = 0$, $c_N \geq 0$ und $A_B = I$. Hiermit kann das duale Simplexverfahren gestartet werden. Die berechneten Tableaus schreiben wir wie beim revidierten primalen Simplexverfahren in der Form

s	y^T	c_0
B	A_B^{-1}	x_B

Hierbei ist B die aktuelle Menge der Basisindizes, $s \notin B$ wird in die Basis aufgenommen, y ist dual zulässig und x_B der Basisanteil der (erst zum Schluß zulässigen)

Basislösung x mit den Kosten $c_0 = c^T x = b^T y$. Ferner wird der zu entfernende Basisindex wieder eingerahmt. Man erhält die folgenden Tableaus:

4	0	0	0	0
5	1	0	0	14
6	0	1	0	-25
7	0	0	1	14
2	0	$-\frac{6}{5}$	0	30
5	1	$\frac{4}{5}$	0	-6
4	0	$-\frac{1}{5}$	0	5
7	0	$\frac{2}{5}$	1	4
-1	-2	0	36	
2	$-\frac{5}{3}$	$-\frac{4}{3}$	0	10
4	$\frac{2}{3}$	$\frac{1}{3}$	0	1
7	$\frac{1}{3}$	$\frac{2}{3}$	1	2

Aus dem letzten, optimalen Tableau liest man die optimale, zulässige Basislösung $x^* := (0, 10, 0, 1, 0, 0, 2)^T$ zu den Basisindizes $B^* := \{2, 4, 7\}$ mit den minimalen Kosten $c_0^* := 36$ ab. Mit $y^* := (-1, -2, 0)^T$ ist gleichzeitig eine Lösung des dualen Programms berechnet worden. \square

Aufgaben

1. Bei gegebenen $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ betrachte man die diskrete, lineare Tschebyscheffsche Approximationsaufgabe, $f(x) := \|Ax - b\|_\infty$ auf dem \mathbb{R}^n zu minimieren. Diese Aufgabe ist (siehe Aufgabe 2 in Abschnitt 6.1) äquivalent dem linearen Programm

(P) Minimiere δ unter den Nebenbedingungen $\delta \geq 0$, $-\delta e \leq Ax - b \leq \delta e$,

wobei $e := (1, \dots, 1)^T \in \mathbb{R}^m$.

- (a) Sei $b \notin \text{Bild}(A)$ und $A^T = (a_1 \ \dots \ a_m)$. Mit Hilfe des starken Dualitätssatzes der linearen Optimierung zeige man: Ist (x^*, δ^*) eine Lösung von (P) und $I^* := \{i \in \{1, \dots, m\} : |(Ax^*)_i - b_i| = \delta^*\}$, so existieren für $i \in I^*$ nichtnegative λ_i , die nicht alle verschwinden, mit

$$\sum_{i \in I^*} \lambda_i \operatorname{sign}((Ax^*)_i - b_i) a_i = 0.$$

- (b) Man zeige, daß die vorige notwendige Optimalitätsbedingung auch hinreichend ist.

2. Man schreibe ein Programm, das zu vorgegebenen $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ eine Lösung der diskreten, linearen Tschebyscheffschen Approximationsaufgabe, $f(x) := \|Ax - b\|_\infty$ auf dem \mathbb{R}^n zu minimieren, berechnet. Hierbei gehe man so vor wie in dem Beispiel auf Seite 113, indem man auf das duale Programm das (revidierte) Simplexverfahren anwendet und hiermit in der ersten Zeile eines optimalen Tableaus eine Lösung des eigentlich interessierenden Problems erhält (genauer muß man die Zeile noch mit -1 multiplizieren). Anschließend teste man das Programm und bestimme zu $t_i := (i-1)/10$, $i = 1, \dots, 11$, ein Polynom $p^* \in \Pi_4$ mit

$$\max_{i=1, \dots, 11} |p^*(t_i) - \exp(t_i)| \leq \max_{i=1, \dots, 11} |p(t_i) - \exp(t_i)| \quad \text{für alle } p \in \Pi_4.$$

Hinweis: Als Lösung wird man

$$p^*(t) \approx 1.000026 + 0.998714 t + 0.510077 t^2 + 0.139717 t^3 + 0.069722 t^4$$

erhalten. Ferner ist

$$\delta^* := \max_{i=1,\dots,11} |p^*(t_i) - \exp(t_i)| \approx 0.000026.$$

Skizziert man den Defekt $p^*(t) - \exp(t)$ über dem Intervall $[0, 1]$, so erhält man das in Abbildung 6.4 angegebene Bild.

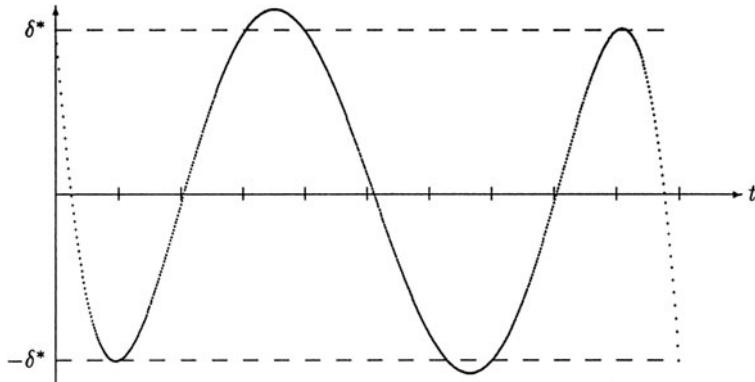


Abbildung 6.4: Der Defekt $p^*(t) - \exp(t)$

3. Sei $A \in \mathbb{R}^{m \times n}$. Man zeige, daß die beiden linearen Programme

(P) Minimiere α unter den Nebenbedingungen $x \geq 0, Ax \leq \alpha e, e^T x = 1$

und

(D) Maximiere β unter den Nebenbedingungen $y \geq 0, A^T y \geq \beta e, e^T y = 1$,

wobei e der Vektor im \mathbb{R}^n bzw. \mathbb{R}^m ist, dessen Komponenten alle gleich 1 sind, dual zu einander sind. Mit Hilfe des starken Dualitätssatzes beweise man die Aussage des *Hauptsatzes der Theorie der Matrixspiele*:

$$\max_{y \in Y} \min_{x \in X} y^T Ax = \min_{x \in X} \max_{y \in Y} y^T Ax,$$

wobei

$$X := \{x \in \mathbb{R}^n : x \geq 0, e^T x = 1\}, \quad Y := \{y \in \mathbb{R}^m : y \geq 0, e^T y = 1\}.$$

4. Sei $A \in \mathbb{R}^{m \times n}$. Mit Hilfe des Farkas-Lemmas (Lemma 3.4) zeige man: Das System

$$(I) \quad Ax = 0, \quad x \geq 0, \quad x \neq 0$$

besitzt genau dann keine Lösung, wenn das System

$$(II) \quad A^T y > 0$$

eine Lösung besitzt.

5. Das lineare Programm in Normalform

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

mit $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ besitze eine Lösung $x^* \in M$. Sei y^* eine Lösung des dualen Programms

$$(D) \quad \text{Maximiere } b^T y \text{ auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Mit $c_{n+1} \in \mathbb{R}$ und $a_{n+1} \in \mathbb{R}^m$ betrachte man das erweiterte lineare Programm

$$(\tilde{P}) \quad \left\{ \begin{array}{l} \text{Minimiere } \left(\begin{array}{c} c \\ c_{n+1} \end{array} \right)^T \left(\begin{array}{c} x \\ x_{n+1} \end{array} \right) \text{ unter den Nebenbedingungen} \\ \left(\begin{array}{c} x \\ x_{n+1} \end{array} \right) \geq 0, \quad \left(\begin{array}{cc} A & a_{n+1} \end{array} \right) \left(\begin{array}{c} x \\ x_{n+1} \end{array} \right) = b. \end{array} \right.$$

Sei c_{n+1} so groß, daß $c_{n+1} > a_{n+1}^T y^*$. Dann gilt:

$$(a) \quad \left(\begin{array}{c} x^* \\ 0 \end{array} \right) \text{ ist eine Lösung von } (\tilde{P}).$$

$$(b) \quad \text{Ist } \left(\begin{array}{c} \tilde{x} \\ \tilde{x}_{n+1} \end{array} \right) \text{ eine Lösung von } (\tilde{P}), \text{ so ist notwendig } \tilde{x}_{n+1} = 0 \text{ und } \tilde{x} \text{ eine Lösung von } (P).$$

Hinweis: Das zu (\tilde{P}) duale Programm ist

$$(\tilde{D}) \quad \text{Maximiere } b^T y \text{ unter den Nebenbedingungen } \left(\begin{array}{c} A^T \\ a_{n+1}^T \end{array} \right) y \leq \left(\begin{array}{c} c \\ c_{n+1} \end{array} \right).$$

Wegen des starken Dualitätssatzes (angewandt auf (P) und (D)) ist $c^T x^* = b^T y^*$. Die Gültigkeit des ersten Teiles der Aufgabe erkennt man, wenn man berücksichtigt, daß $\left(\begin{array}{c} x^* \\ 0 \end{array} \right)$ bzw. y^* zulässig für (\tilde{P}) bzw. (\tilde{D}) sind, und anschließend den schwachen Dualitätssatz anwendet. Im zweiten Teil der Aufgabe nehme man im Widerspruch zur Behauptung an, es sei $\tilde{x}_{n+1} > 0$ und zeige

$c^T x^* = c^T \tilde{x} + c_{n+1} \tilde{x}_{n+1} > c^T \tilde{x} + (a_{n+1} \tilde{x}_{n+1})^T y^* = c^T \tilde{x} + (b - A \tilde{x})^T y^* \geq b^T y^* = c^T x^*$, ein Widerspruch. Also ist notwendig $\tilde{x}_{n+1} = 0$ und daher \tilde{x} zulässig für (P) . Wegen $c^T x^* = c^T \tilde{x}$ ist auch \tilde{x} eine Lösung von (P) .

6. Bei gegebenen $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $d \in \mathbb{R}^m$ betrachte man für $t \in [0, 1]$ das lineare Programm in Normalform

$$(P_t) \quad \text{Minimiere } c^T x \text{ auf } M_t := \{x \in \mathbb{R}^n : x \geq 0, Ax = b + td\},$$

dessen duales Programm durch

$$(D_t) \quad \text{Maximiere } (b + td)^T y \text{ auf } N := \{y \in \mathbb{R}^m : A^T y \leq c\}$$

gegeben ist. Es werde vorausgesetzt, daß M_0 , M_1 und N nicht leer sind. Man zeige der Reihe nach:

- (a) $M_t \neq \emptyset$ für alle $t \in [0, 1]$.
- (b) Für alle $t \in [0, 1]$ sind (P_t) und (D_t) lösbar.
- (c) Die durch $w(t) := \min(P_t)$ definierte Funktion $w: [0, 1] \rightarrow \mathbb{R}$ ist konvex.
- (d) Die Aufgabe

(D_{opt}) Maximiere $d^T y_0$ auf $N_{\text{opt}} := \{y_0 \in \mathbb{R}^m : y_0 \text{ ist Lösung von } (D_0)\}$

besitzt eine Lösung y_0^* .

- (e) Ist y_0^* eine Lösung von (D_{opt}) , so ist y_0^* für alle hinreichend kleinen $t > 0$ auch eine Lösung von (D_t) .
- (f) Für alle hinreichend kleinen $t > 0$ ist

$$\min(P_t) = \min(P_0) + t \max(D_{\text{opt}}) = \min(P_0) + t d^T y_0^*,$$

wobei y_0^* eine Lösung von (D_{opt}) ist.

7. Man programmiere das duale Simplexverfahren zur Lösung linearer Programme der Form

Minimiere $c^T x$ unter den Nebenbedingungen $x \geq 0, Ax \leq b$

mit nichtnegativem Kostenvektor c . Anschließend teste man das Programm an dem folgenden Beispiel, bei dem die Daten durch

c^T		3	1	2	4	
A	b	-5	2	-3	-1	20
		3	-4	1	4	-15
		2	5	-5	2	12

gegeben sind.

8. Man formuliere und beweise eine Aufgabe 4 in Abschnitt 6.2 entsprechende Aussage zum dualen Simplexverfahren, welches mit vollständigen Tableaus arbeitet.
 9. Gegeben sei das lineare Programm in Normalform

(P) Minimiere $c^T x$ auf $M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$.

Hierbei sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $\text{Rang}(A) = m$. Das lineare Programm (P) sei lösbar, das revidierte primale oder duale Simplexverfahren liefere ein optimales Tableau

	y^T	c_0
B	A_B^{-1}	x_B

Bei gegebenen $a \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ betrachte man anschließend die Aufgabe

(\tilde{P}) Minimiere $c^T x$ auf $\tilde{M} := M \cap \{x \in \mathbb{R}^n : a^T x \leq \beta\}$.

Man überlege sich, wie man (\tilde{P}) mit Hilfe des dualen Simplexverfahrens unter Benutzung des für (P) optimalen Tableaus lösen kann.

Hinweis: In Normalform ist die Aufgabe (\tilde{P}) nach Einführung einer Schlupfvariablen für die zusätzliche Ungleichungsrestriktion durch die Daten

c^T	0	
A	0	b
a^T	1	β

gegeben. Das duale Simplexverfahren kann mit dem Tableau

	y^T	0	c_0
B	A_B^{-1}	0	x_B
$n + 1$	$-a_B^T A_B^{-1}$	1	$\beta - a_B^T x_B$

gestartet werden.

6.4 Das Karmarkar-Verfahren

Das Simplexverfahren hat sich in der Praxis außerordentlich bewährt. Allerdings gaben V. KLEE, G. MINTY (1972) ein Beispiel eines linearen Programms an, das grob gesagt darin besteht, die höchste Ecke eines verzerrten Würfels im \mathbb{R}^n zu bestimmen, und bei dem das Simplexverfahren sämtliche 2^n Ecken durchläuft. Daher ist das Simplexverfahren „im schlechtesten Fall“ ein schlechtes Verfahren (dafür aber „im Mittel“ ein gutes Verfahren, siehe K. H. BORGWARDT (1987)). Von L. G. KHACHIAN (1979) wurde gezeigt, daß es für lineare Programme einen Algorithmus gibt, der im schlechtesten Fall ein „gutes“ Verfahren ist. Über das Ergebnis von Khachian, das hier nicht präzisiert werden soll, wurde in vielen Zeitungen und Magazinen berichtet, es wurde allerdings oft gründlich mißverstanden².

Zwar wird inzwischen in einigen neueren Lehrbüchern der linearen Optimierung (z. B. C. H. PAPADIMITRIOU, K. STEIGLITZ (1982), K. G. MURTY (1983) und A. SCHRIJVER (1986)) über das Khachian-Verfahren berichtet, man ist sich aber wohl darüber einig, daß es sich für die Praxis nicht eignet, sondern daß durch das Khachian-Verfahren eine für die Komplexitätstheorie linearer Optimierungsaufgaben wichtige theoretische Problemstellung gelöst wurde.

Etwas anders verhält es sich mit dem 1984 vorgestellten Karmarkar-Verfahren (siehe N. KARMARKAR (1984)). Auch dieses Verfahren ist im schlechtesten Fall ein gutes Verfahren (wieder eine Aussage, die hier nicht präzisiert werden soll), von dem aber zumindestens der Autor behauptete, daß es bei Problemen aus der Praxis wesentlich schneller als das Simplexverfahren sei. Diese Behauptung scheint noch nicht bewiesen zu sein. Es gibt aber Anzeichen dafür, daß geeignete Implementationen des

²So konnte man z. B. im Dezember 1979 im SPIEGEL unter der Überschrift „Schnelles Öl“ lesen: „Wahrscheinlich wird sich durch Khachians Erkenntnis die dagegen doch recht umständliche Simplex-Methode weithin ersetzen lassen.“ Über das Simplexverfahren selber erfährt man: „Professor Dantzig von der Stanford University ersann jedoch einen Ausweg, der seiner Zunft zwar nicht ganz geheuer scheint, meist aber zu brauchbaren Ergebnissen führt.“ In dem interessanten Aufsatz von E. L. LAWLER (1980) kann man nachlesen, daß die Reaktion der Presse in den USA und England ähnlich war.

Karmarkar-Verfahrens gerade für große lineare Optimierungsaufgaben zumindestens konkurrenzfähig sind. Ganz sicher ist aber, daß die Arbeit von Karmarkar eine ganz neue Entwicklung auf dem Gebiet der linearen Optimierung initiierte, was durch außerordentlich viele Veröffentlichungen dokumentiert ist. Die Übersichtsaufsätze von U. ZIMMERMANN (1988) und D. GOLDFARB, M. J. TODD (1989) geben hiervon einen Eindruck.

Wir wollen in diesem Abschnitt die wesentlichen Ideen des Karmarkar-Verfahrens schildern, wobei allerdings komplexitätstheoretische Fragen nicht berücksichtigt werden. Hierzu wird zunächst von einer sogenannten Karmarkar-Normalform des linearen Programms ausgegangen. Für diese Normalform wird das Karmarkar-Verfahren motiviert und analysiert. Am Schluß dieses Abschnittes werden wir auf die Überführung eines allgemeinen linearen Programms auf Karmarkar-Normalform eingehen. Natürlich kann hier nur ein kleiner Einblick in diese neue, noch keineswegs abgeschlossene Entwicklung auf dem Gebiet der linearen Optimierung gegeben werden.

6.4.1 Das Karmarkar-Verfahren und seine Motivation

Gegeben sei das lineare Programm

$$(P) \quad \text{Minimiere } c^T x \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{pmatrix} A \\ e^T \end{pmatrix} x = \begin{pmatrix} 0 \\ n \end{pmatrix}, x \geq 0 \right\}.$$

Hierbei seien $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ mit $n \geq 2$ gegeben und $e := (1, \dots, 1)^T \in \mathbb{R}^n$. Zusätzlich wird vorausgesetzt, daß

$$(V) \quad \text{Rang}(A) = m, \quad Ae = 0, \quad \min(P) = 0.$$

Wir sagen, das lineare Programm (P) sei in *Karmarkar-Normalform*, wenn (V) erfüllt ist. Auf eine Möglichkeit, ein (allgemeines) lineares Programm auf Karmarkar-Normalform zurückzuführen, werden wir in Unterabschnitt 6.4.3 eingehen. Das von Karmarkar entwickelte Verfahren läßt sich sehr einfach angeben und lautet folgendermaßen:

- Gegeben sei das lineare Programm (P), die Voraussetzung (V) sei erfüllt. Sei $\alpha \in (0, 1)$, $r := \sqrt{n/(n-1)}$ und $x^0 := e$.
- Für $k = 0, 1, \dots$:

Falls $c^T x^k = 0$, dann: STOP, x^k ist eine Lösung von (P).

$$\text{Setze } D_k := \text{diag}(x_1^k, \dots, x_n^k), \quad B_k := \begin{pmatrix} AD_k \\ e^T \end{pmatrix}.$$

$$\text{Berechne } p^k := [I - B_k^T (B_k B_k^T)^{-1} B_k] D_k c,$$

$$y^{k+1} := e - \alpha r \frac{p^k}{\|p^k\|_2},$$

$$x^{k+1} := \frac{n}{(x^k)^T y^{k+1}} D_k y^{k+1}.$$

Bevor im nächsten Lemma die *Durchführbarkeit* des Karmarkar-Verfahrens bewiesen wird (grob bedeutet das den Nachweis, daß nicht durch Null dividiert wird), wollen wir jetzt versuchen, das Verfahren zu motivieren. Hierzu nehmen wir an, $x \in M$ mit $x > 0$ sei eine aktuelle Näherung, die keine Lösung ist, für die also $c^T x > 0$. Wie im Algorithmus angegeben, sei $D := \text{diag}(x_1, \dots, x_n)$ die $n \times n$ -Diagonalmatrix, die die Komponenten von x in der Diagonalen enthält. Mit dem (kompakten) Simplex

$$\Sigma := \{z \in \mathbb{R}^n : e^T z = n, z \geq 0\}$$

ist das Ausgangsproblem gegeben durch

$$(P) \quad \text{Minimiere } c^T z \quad \text{unter den Nebenbedingungen } z \in \text{Kern}(A) \cap \Sigma.$$

Nun definiere man die *projektive Transformation* $T: \Sigma \rightarrow \Sigma$ durch

$$y = T(z) := \frac{n}{e^T D^{-1} z} D^{-1} z.$$

Dann ist $T(x) = e$, d. h. die aktuelle Näherung x wird in den Schwerpunkt e von Σ abgebildet. Ferner bildet T das Simplex Σ eindeutig auf sich ab. Die inverse Transformation $T^{-1}: \Sigma \rightarrow \Sigma$ ist durch

$$z = T^{-1}(y) := \frac{n}{x^T y} Dy$$

gegeben. Schreibt man das Ausgangsproblem in der transformierten Variablen y , ersetzt man also z durch $T^{-1}(y)$, so erhält man das äquivalente lineare Quotienten-Programm

$$\text{Minimiere } n \frac{(Dc)^T y}{x^T y} \quad \text{unter der Nebenbedingung } y \in \text{Kern}(AD) \cap \Sigma.$$

Wegen der Voraussetzung $\min(P) = 0$ ist dieses Programm wiederum äquivalent zu

$$(P_T) \quad \text{Minimiere } (Dc)^T y \quad \text{unter der Nebenbedingung } y \in \text{Kern}(AD) \cap \Sigma.$$

Genauer ist x^* ist genau dann eine Lösung von (P) , wenn $y^* := T(x^*)$ eine Lösung von (P_T) ist. Nun ist das transformierte Problem (P_T) genauso schwer oder leicht zu lösen wie das Ausgangsproblem (P) . Auf den ersten Blick hat es lediglich den Vorteil, daß der Schwerpunkt e des Simplex Σ zulässig für (P_T) ist. Die entscheidende Idee von Karmarkar besteht nun darin, (P_T) als *Relaxation* einer geschlossen lösbareren, nichtlinear restringierten Optimierungsaufgabe aufzufassen. Grundlage hierfür ist

Lemma 4.1 Sei

$$\Sigma := \{y \in \mathbb{R}^n : e^T y = n, y \geq 0\}, \quad K[e; \delta] := \{y \in \mathbb{R}^n : e^T y = n, \|y - e\|_2 \leq \delta\}.$$

Dann gilt: Mit $r := \sqrt{n/(n-1)}$, $R := \sqrt{n(n-1)}$ und $\alpha \in (0, 1)$ ist

$$K[e; r] \subset \Sigma \subset K[e; R], \quad K[e; \alpha r] \subset \{y \in \mathbb{R}^n : e^T y = n, y > 0\}.$$

Beweis: Der einfache Beweis bleibt dem Leser überlassen, siehe Aufgabe 1. \square

Geometrisch gesprochen ist $K[e; r]$ Inkugel und $K[e; R]$ Umkugel zum Simplex Σ in der Hyperebene $H := \{y \in \mathbb{R}^n : e^T y = n\}$. Wegen $K[e; \alpha r] \subset \Sigma$ ist das lineare Programm (P_T^α) eine Relaxation der Optimierungsaufgabe

(P_T^α) Minimiere $(Dc)^T y$ unter der Nebenbedingung $y \in \text{Kern}(AD) \cap K[e; \alpha r]$.

Die Aufgabe (P_T^α) besitzt, wie im folgenden Lemma bewiesen wird, eine eindeutige, geschlossene angebbare Lösung y^+ . Das wollen wir hier schon geometrisch motivieren. Ersetzt man y durch $e - u$, so erkennt man, daß y^+ genau dann eine Lösung von (P_T^α) ist, wenn $u^+ := e - y^+$ eine Lösung von

$(*)$ Maximiere $(Dc)^T u$ unter der Nebenbedingung $u \in \text{Kern}(B) \cap B[0; \alpha r]$

ist, wobei

$$B := \begin{pmatrix} AD \\ e^T \end{pmatrix}, \quad B[0; \alpha r] := \{u \in \mathbb{R}^n : \|u\|_2 \leq \alpha r\}$$

gesetzt ist. Die Aufgabe $(*)$ besteht also darin, eine lineare, reellwertige Funktion auf dem Schnitt des linearen Teilraumes $\text{Kern}(B)$ mit der euklidischen Kugel $B[0; \alpha r]$ zu minimieren. In Abbildung 6.5 wird veranschaulicht, daß man die Lösung u^+ von $(*)$ in zwei Schritten erreicht. Zunächst gewinnt man den Vektor p , indem man

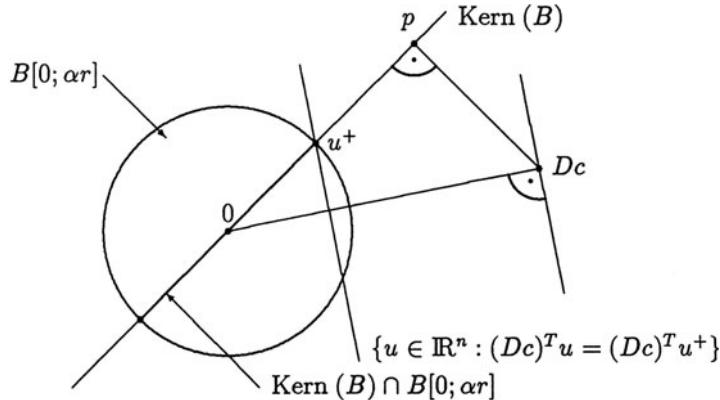


Abbildung 6.5: Die Konstruktion der Lösung u^+ von $(*)$

Dc auf $\text{Kern}(B)$ (orthogonal) projiziert. Anschließend multipliziert man p so mit einem positiven Faktor, daß der resultierende Vektor auf dem Rand von $B[0; \alpha r]$ liegt, d. h. man berechnet $u^+ := \alpha p / \|p\|_2$. Daher wird man erwarten, daß $y^+ := e - u^+$ eine Lösung von (P_T^α) ist. Wegen $y^+ \in \text{Kern}(AD) \cap K[e; \alpha r]$ und Lemma 4.1 ist $y^+ \in \text{Kern}(AD) \cap \Sigma$ und $y^+ > 0$. Durch Rücktransformation erhält man die neue aktuelle Näherung $x^+ := T^{-1}(y^+)$. Diese ist zulässig für das Ausgangsproblem (P) , d. h. es ist $x^+ \in \text{Kern}(A) \cap \Sigma$, ferner ist $x^+ > 0$. Anders als beim Simplexverfahren

sind die Näherungen beim Karmarkar-Verfahren *positive* Vektoren, also aus dem *Innern*³ des nichtnegativen Orthanten im \mathbb{R}^n . In Abbildung 6.6 fassen wir die Schritte des Karmarkar-Verfahrens, von einer aktuellen, zulässigen Näherung $x \in M$ mit $x > 0$ zur nächsten Näherung $x^+ \in M$ mit $x^+ > 0$ zu gelangen, noch einmal zusammen.

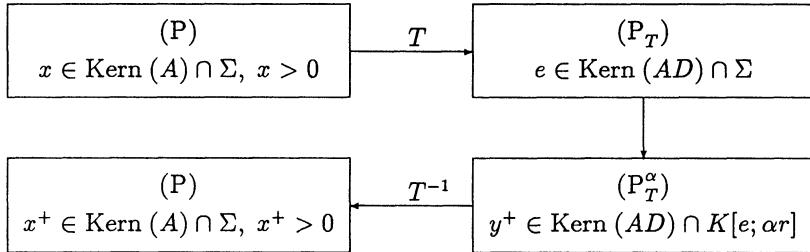


Abbildung 6.6: Ein Schritt des Karmarkar-Verfahrens

Aus dem folgenden Lemma wird die Durchführbarkeit des Karmarkar-Verfahrens folgen. Ferner wird in ihm zum Schluß ein Ergebnis formuliert und bewiesen, das später im Konvergenzbeweis benötigt wird.

Lemma 4.2 Gegeben sei das lineare Programm (P) in Karmarkar-Normalform, die Voraussetzung (V) sei erfüllt. Sei $x \in M$ mit $x > 0$ eine aktuelle Näherung und $c^T x > 0$, also x keine Lösung von (P). Man setze

$$D := \text{diag}(x_1, \dots, x_n), \quad B := \begin{pmatrix} AD \\ e^T \end{pmatrix}.$$

Dann gilt:

1. Es ist $\text{Rang}(B) = m + 1$ und daher BB^T nichtsingulär.
2. Es ist $p := [I - B^T(BB^T)^{-1}B]Dc \neq 0$.
3. $y^+ := e - \alpha r p / \|p\|_2$ ist die eindeutige Lösung von

Minimiere $(Dc)^T y$ unter der Nebenbedingung $y \in \text{Kern}(AD) \cap K[e; \alpha r]$,

wobei $\alpha \in (0, 1)$, $r := \sqrt{n/(n-1)}$ und

$$K[e; \alpha r] := \{y \in \mathbb{R}^n : e^T y = n, \|y - e\|_2 \leq \alpha r\}.$$

4. Es ist $(Dc)^T y^+ \leq [1 - \alpha/(n-1)] c^T x$.

³Karmarkars Trick: Bei seiner schrittweisen Näherung an die Ideallösung schneidet er gleichsam durch den Monsterrörper möglicher Lösungen und sucht von innen her nach jener Ecke, welche das optimale Ergebnis repräsentiert (SPIEGEL (1984)).

Beweis: Wir zeigen, daß der Kern von B^T nur aus dem Nullvektor (im \mathbb{R}^{m+1}) besteht. Sei also

$$0 = B^T \begin{pmatrix} z \\ z_{m+1} \end{pmatrix} = DA^T z + z_{m+1} e.$$

Wegen $Ae = 0$ ist $ADA^T z = 0$. Da der Rang von A nach Voraussetzung maximal und D eine positive Diagonalmatrix ist, folgt $z = 0$ und dann auch $z_{m+1} = 0$. Daher ist $\text{Rang}(B) = m+1$, der Vektor $p := [I - B^T(BB^T)^{-1}B]Dc$ ist definiert.

Angenommen, es sei $p = 0$. Dann ist $Dc \in \text{Bild}(B^T)$, es existieren also $z \in \mathbb{R}^m$ und $z_{m+1} \in \mathbb{R}$ mit

$$Dc = DA^T z + z_{m+1} e \quad \text{bzw.} \quad c = A^T z + z_{m+1} D^{-1} e.$$

Ist $x^* \in M$ eine Lösung von (P), so ist nach Voraussetzung

$$0 = c^T x^* = (\underbrace{Ax^*}_=)^T z + z_{m+1} \underbrace{e^T D^{-1} x^*}_{>0}$$

und damit $z_{m+1} = 0$. Aus $c = A^T z$ folgt $c^T x = 0$, ein Widerspruch.

Offenbar ist $Bp = 0$ und daher $y^+ := e - \alpha r p / \|p\|_2 \in \text{Kern}(AD) \cap K[e; \alpha r]$. Für beliebiges $y \in \text{Kern}(AD) \cap K[e; \alpha r]$ ist wegen $B(y^+ - y) = 0$ und der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} (Dc)^T (y^+ - y) &= [p + B^T(BB^T)^{-1}BDc]^T (y^+ - y) \\ &= p^T (y^+ - y) \\ &= p^T (e - y) - \alpha r \|p\|_2 \\ &\leq \|p\|_2 \underbrace{\|e - y\|_2}_{\leq \alpha r} - \alpha r \|p\|_2 \\ &\leq 0. \end{aligned}$$

Ist $(Dc)^T (y^+ - y) = 0$, so ist $\|e - y\|_2 = \alpha r$ und $e - y = \lambda p$ mit $\lambda > 0$, da Gleichheit bei der Anwendung der Cauchy-Schwarzschen Ungleichung eingetreten ist. Aus beiden Gleichungen zusammen folgt $y = y^+$. Insgesamt ist auch die dritte Aussage des Lemmas bewiesen.

Sei y^* Lösung von

$$(P_T) \quad \text{Minimiere } (Dc)^T y \quad \text{unter der Nebenbedingung } y \in \text{Kern}(AD) \cap \Sigma.$$

Auch der Wert dieses transformierten Programms ist Null, es ist also $(Dc)^T y^* = 0$. Ferner ist $y^* \neq e$, da andernfalls x schon eine Lösung von (P) wäre. Dagegen ist y^+ die Lösung von

$$(P_T^\alpha) \quad \text{Minimiere } (Dc)^T y \quad \text{unter der Nebenbedingung } y \in \text{Kern}(AD) \cap K[e; \alpha r].$$

Aus $\Sigma \subset K[e; R]$ mit $R := \sqrt{n(n-1)}$ (siehe Lemma 4.1) folgt $\|y^* - e\|_2 \leq R$, so daß

$$e + \frac{\alpha r}{R} (y^* - e) \in \text{Kern}(AD) \cap K[e; \alpha r]$$

zulässig für (P_T^α) ist. Da ferner y^+ die Lösung von (P_T^α) ist, folgt

$$\begin{aligned}(Dc)^T y^+ &\leq (Dc)^T \left(e + \frac{\alpha r}{R} (y^* - e) \right) \\&= \left(1 - \frac{\alpha r}{R} \right) \underbrace{(Dc)^T e}_{=c^T x} \\&= \left(1 - \frac{\alpha}{n-1} \right) c^T x.\end{aligned}$$

Damit ist das Lemma bewiesen. \square

Bemerkung: Die Hauptarbeit in jedem Schritt des Karmarkar-Verfahrens besteht in der Berechnung von $p := [I - B^T(BB^T)^{-1}B]Dc$. Da $p \in \text{Kern}(B)$ und $p - Dc$ senkrecht auf $\text{Kern}(B)$ steht, ist p die *orthogonale Projektion* von Dc auf $\text{Kern}(B)$. Unter Benutzung von $B^T = \begin{pmatrix} DA^T & e \end{pmatrix}$ erhält man

$$\begin{aligned}p &:= [I - B^T(BB^T)^{-1}B]Dc \\&= \left[I - \begin{pmatrix} DA^T & e \end{pmatrix} \begin{pmatrix} AD^2 A^T & AD e \\ e^T DA^T & n \end{pmatrix}^{-1} \begin{pmatrix} AD \\ e^T \end{pmatrix} \right] Dc \\&= \left[I - \begin{pmatrix} DA^T & e \end{pmatrix} \begin{pmatrix} AD^2 A^T & 0 \\ 0^T & n \end{pmatrix}^{-1} \begin{pmatrix} AD \\ e^T \end{pmatrix} \right] Dc \\&= \left[I - (AD)^T (AD^2 A^T)^{-1} (AD) - \frac{1}{n} ee^T \right] Dc.\end{aligned}$$

Setzt man zur Abkürzung

$$q := (AD^2 A^T)^{-1} AD Dc = [(AD)(AD)^T]^{-1} AD Dc,$$

so ist

$$p = D(c - A^T q) - \frac{c^T x}{n} e.$$

Der Vektor q genügt den Normalgleichungen $(AD)(AD)^T q = (AD)(Dc)$, d. h. q ergibt sich als Lösung des linearen Ausgleichsproblems

$$\text{Minimiere } \|(AD)^T q - Dc\|_2 = \|D(A^T q - c)\|_2, \quad q \in \mathbb{R}^m.$$

Wegen $m = \text{Rang}(A) = \text{Rang}(AD) = \text{Rang}(AD)^T$ sind eine ganze Reihe von Methoden zur Berechnung von q denkbar, erinnert sei nur an die Cholesky-Zerlegung sowie die QR -Zerlegung nach Householder und Givens. Wichtig ist hierbei, wie stets in diesem Zusammenhang, von der Zerlegung zu einer Iterationsstufe auf die der nächsten zu schließen. Hierzu sei lediglich auf D. F. SHANNO (1988) verwiesen. Zu beachten sind aber auch iterative Verfahren zur Lösung linearer Ausgleichsprobleme, vor allem das im nächsten Kapitel beschriebene *Verfahren der konjugierten Gradienten*. Dies gilt insbesondere dann, wenn die Matrix A dünn besetzt ist. \square

6.4.2 Die Konvergenz des Karmarkar-Verfahrens

Der Konvergenzaussage zum Karmarkar-Verfahren werden zwei Lemmata vorangestellt. Für deren Beweis verweisen wir z. B. auf W. WALTER (1990, S. 47) bzw. A. SCHRIJVER (1986, S. 192) und J. FRANKLIN (1987). Eine andere Beweisidee zu Lemma 4.4 wird in Aufgabe 3 angegeben.

Lemma 4.3 Ist $x = (x_j) \in \mathbb{R}^n$ mit $x_j \geq 0$ für $j = 1, \dots, n$, so gilt die Ungleichung vom geometrisch-arithmetischen Mittel:

$$\left(\prod_{j=1}^n x_j \right)^{1/n} \leq \frac{1}{n} \sum_{j=1}^n x_j$$

Lemma 4.4 Sei $\alpha \in (0, 1)$ und $r := \sqrt{n/(n-1)}$. Dann ist

$$\prod_{j=1}^n y_j \geq (1 - \alpha) \left(1 + \frac{\alpha}{n-1} \right)^{n-1} \quad \text{für alle } y \in K[e; \alpha r],$$

wobei $K[e; \alpha r] := \{y \in \mathbb{R}^n : e^T y = n, \|y - e\|_2 \leq \alpha r\}$.

Nun formulieren wir eine Konvergenzaussage zu dem im vorigen Unterabschnitt formulierten Karmarkar-Verfahren. Wir werden dabei die dort eingeführten Bezeichnungen benutzen.

Satz 4.5 Gegeben sei das lineare Programm (P) in Karmarkar-Normalform, die Voraussetzung (V) sei also erfüllt. Für $\alpha \in (0, 1)$ und $n \in \mathbb{N}$ definiere man

$$q(\alpha, n) := \frac{[1 - \alpha/(n-1)]^n}{(1 - \alpha)[1 + \alpha/(n-1)]^{n-1}}.$$

Dann gilt:

1. Es ist

$$q(\alpha, n) \leq \lim_{n \rightarrow \infty} q(\alpha, n) = \frac{e^{-2\alpha}}{1 - \alpha} =: q(\alpha).$$

Für $n \geq 3$ wird $q(\cdot, n)$ auf $(0, 1)$ minimal für $\alpha_n := (n-1)/(2n-3)$, $q(\cdot)$ wird auf $(0, 1)$ minimal für $\alpha := 1/2$.

2. Sei $x \in M$ mit $x > 0$ eine aktuelle Näherung, die noch keine Lösung von (P) ist, und x^+ die neue Näherung. Dann gilt die Ungleichung

$$\prod_{j=1}^n \frac{c^T x^+}{x_j^+} \leq q(\alpha, n) \prod_{j=1}^n \frac{c^T x}{x_j}.$$

3. Bricht das Karmarkar-Verfahren nicht vorzeitig mit einer Lösung von (P) ab, so erzeugt es eine Folge $\{x^k\} \subset M$ mit

$$\frac{c^T x^k}{c^T x^0} \leq q(\alpha, n)^{k/n} \quad \text{für } k = 0, 1, \dots$$

Wählt man im Karmarkar-Verfahren $\alpha := 1/2$, so ist

$$\frac{c^T x^k}{c^T x^0} \leq \left(\frac{2}{e}\right)^{k/n} \quad \text{für } k = 0, 1, \dots,$$

ferner ist jeder Häufungspunkt von $\{x^k\}$ eine Lösung von (P).

Beweis: Die Aussagen über $q(\alpha, n)$ sind mit elementaren Hilfsmitteln der Analysis beweisbar, der Beweis wird übergangen.

Es ist

$$\begin{aligned} \prod_{j=1}^n \frac{c^T x^+}{x_j^+} \Big/ \prod_{j=1}^n \frac{c^T x}{x_j} &= \prod_{j=1}^n \left(\frac{c^T x^+}{c^T x} \cdot \frac{x_j}{x_j^+} \right) \\ &= \prod_{j=1}^n \left(\frac{(Dc)^T y^+}{c^T x} \cdot \frac{x_j}{(Dy^+)_j} \right) \\ &\quad (\text{wegen } x^+ := \frac{n}{x^T y^+} Dy^+) \\ &= \prod_{j=1}^n \frac{(Dc)^T y^+}{c^T x} \prod_{j=1}^n \frac{1}{y_j^+} \\ &\leq \left(1 - \frac{\alpha}{n-1} \right)^n \prod_{j=1}^n \frac{1}{y_j^+} \\ &\quad (\text{wegen Lemma 4.2}) \\ &\leq \left(1 - \frac{\alpha}{n-1} \right)^n \frac{1}{(1-\alpha)[1+\alpha/(n-1)]^{n-1}} \\ &\quad (\text{wegen } y^+ \in K[e; \alpha r] \text{ und Lemma 4.4}) \\ &= q(\alpha, n). \end{aligned}$$

Damit ist der zweite Teil des Satzes bewiesen.

Setzt man im gerade bewiesenen Teil des Satzes $x := x^k$ und $x^+ := x^{k+1}$, so erhält man

$$\prod_{j=1}^n \frac{c^T x^{k+1}}{x_j^{k+1}} \leq q(\alpha, n) \prod_{j=1}^n \frac{c^T x^k}{x_j^k} \quad \text{für } k = 0, 1, \dots$$

und durch „Zurückspulen“

$$\prod_{j=1}^n \frac{c^T x^k}{x_j^k} \leq q(\alpha, n)^k \prod_{j=1}^n \frac{c^T x^0}{x_j^0} \quad \text{für } k = 0, 1, \dots$$

Unter Benutzung von $x^0 = e$ folgt hieraus mit Hilfe der Ungleichung vom geometrisch-arithmetischen Mittel

$$\frac{c^T x^k}{c^T x^0} \leq q(\alpha, n)^{k/n} \left(\prod_{j=1}^n x_j^k \right)^{1/n} \leq q(\alpha, n)^{k/n} \frac{1}{n} \sum_{j=1}^n x_j^k = q(\alpha, n)^{k/n}.$$

Ist $q(\alpha, n) < 1$, was z. B. für $\alpha = 1/2$ der Fall ist, so ist $\lim_{k \rightarrow \infty} c^T x^k = 0$. Ist daher x^* ein Häufungspunkt der Folge $\{x^k\} \subset M$, so ist $x^* \in M$ und $c^T x^* = 0$, also x^* eine Lösung von (P). \square

Bemerkung: Ist $x \in M$ mit $x > 0$ im Karmarkar-Verfahren eine aktuelle Näherung, die noch keine Lösung ist, so wird durch

$$p := [I - B^T(BB^T)^{-1}B]Dc \quad \text{mit} \quad B := \begin{pmatrix} AD \\ e^T \end{pmatrix}$$

die Projektion von Dc auf Kern (B) berechnet und anschließend

$$y^+ := e - \alpha r \frac{p}{\|p\|_2}, \quad x^+ := \frac{n}{x^T y^+} Dy^+$$

gesetzt. Entscheidend für den Konvergenzbeweis war die in Lemma 4.2 bewiesene Ungleichung

$$(Dc)^T y^+ = (Dc)^T \left(e - \alpha r \frac{p}{\|p\|_2} \right) \leq \left(1 - \frac{\alpha}{n-1} \right) (Dc)^T e$$

bzw. die äquivalente Ungleichung

$$(Dc)^T \left(e - R \frac{p}{\|p\|_2} \right) \leq 0$$

mit $R := \sqrt{n(n-1)}$. Der obige Konvergenzbeweis zeigt, daß es nicht nötig ist, p als *exakte* Projektion von Dc auf Kern (B) zu wählen. Vielmehr genügt es, p so zu bestimmen, daß

$$Bp = 0, \quad p \neq 0 \quad \text{und} \quad (Dc)^T \left(e - R \frac{p}{\|p\|_2} \right) \leq 0.$$

Genau diese Beobachtung machten D. GOLDFARB, S. MEHROTRA (1988a). □

Bemerkung: In Satz 4.5 wurde

$$\prod_{j=1}^n \frac{c^T x^+}{x_j^+} \leq q(\alpha, n) \prod_{j=1}^n \frac{c^T x}{x_j}$$

bewiesen. Führt man wie N. KARMAKAR (1984) die sogenannte *logarithmische Potentialfunktion* f durch

$$f(z) := n \ln c^T z - \sum_{j=1}^n \ln z_j = \ln \left(\prod_{j=1}^n \frac{c^T z}{z_j} \right)$$

ein, so besagt diese Ungleichung, daß $f(x) - f(x^+) \geq -\ln q(\alpha, n)$. Die Wahl $\alpha := 1/2$ führt auf $f(x) - f(x^+) \geq 1 - \ln 2 \approx 0.3$, in diesem Falle vermindert sich der Wert der logarithmischen Potentialfunktion in jedem Schritt mindestens um 0.3. In der Praxis wird man nicht mit einer *konstanten* Schrittweite $\alpha \in (0, 1)$ arbeiten, sondern diese in jedem Schritt so wählen, daß die Verminderung der logarithmischen Potentialfunktion möglichst groß ist. □

6.4.3 Zurückführung eines linearen Programms auf Karmarkar-Normalform

Es ist ziemlich einfach, ein allgemeines lineares Programm auf Simplex-Normalform zurückzuführen. Nicht offensichtlich ist dagegen, wie eine entsprechende Überführung auf Karmarkar-Normalform aussehen könnte. Das ist auch mit ein Grund dafür, daß einige Varianten zum Karmarkar-Verfahren vorgeschlagen wurden, bei denen die Voraussetzung, der Wert $\min(P)$ sei bekannt bzw. gleich Null, durch die Annahme ersetzt wird, es sei eine untere Schranke für den Wert $\min(P)$ des gegebenen linearen Programms gegeben (siehe u. a. K. M. ANSTREICHER (1986, 1989), M. J. TODD, B. P. BURRELL (1986), D. GOLDFARB, S. MEHROTRA (1988a)). Hierauf wollen wir nicht näher eingehen, sondern im wesentlichen den Vorschlag von N. KARMAKAR (1984) schildern, wie man ein allgemeines lineares Programm auf Karmarkar-Normalform zurückführen kann. Hierbei werden wir auch die Ausführungen bei D. GOLDFARB, S. MEHROTRA (1989) berücksichtigen.

Wir könnten von einem linearen Programm in Simplex-Normalform ausgehen und diesem (unter geeigneten Voraussetzungen) ein äquivalentes lineares Programm in Karmarkar-Normalform zuordnen. Statt dessen ziehen wir es vor, das lineare Programm

$$(P_0) \quad \text{Minimiere } c_0^T u \quad \text{auf } M_0 := \{u \in \mathbb{R}^l : u \geq 0, A_0 u \geq b_0\}$$

als Ausgangsproblem zu wählen und die Zurückführung von (P_0) auf Karmarkar-Normalform zu untersuchen. Dem Hauptergebnis dieses Unterabschnittes, nämlich Satz 4.7, schicken wir ein einfaches Lemma voraus.

Lemma 4.6 *Mit $A_0 \in \mathbb{R}^{k \times l}$, $b_0 \in \mathbb{R}^k$ und $c_0 \in \mathbb{R}^l$ sei das lineare Programm*

$$(P_0) \quad \text{Minimiere } c_0^T u \quad \text{auf } M_0 := \{u \in \mathbb{R}^l : u \geq 0, A_0 u \geq b_0\}$$

gegeben. Die Menge der optimalen Lösungen von (P_0) sei nichtleer und beschränkt. Dann ist

$$\{u \in \mathbb{R}^l : u \geq 0, A_0 u \geq b_0, c_0^T u \leq 0\} = \{0\}.$$

Beweis: Angenommen, $u \in \mathbb{R}^l$ sei ein vom Nullvektor verschiedener, nichtnegativer Vektor mit $A_0 u \geq 0$ und $c_0^T u \leq 0$. Sei $u_0 \in M_0$ eine Lösung von (P_0) . Dann ist $u_0 + tu \in M_0$ für alle $t \geq 0$. Ist $c_0^T u < 0$, so wäre $\inf(P_0) = -\infty$, ein Widerspruch. Ist dagegen $c_0^T u = 0$, so wäre $u_0 + tu$ für alle $t \geq 0$ eine Lösung von (P_0) , was einen Widerspruch zur vorausgesetzten Beschränktheit der Lösungsmenge von (P_0) bedeutet. Insgesamt ist das Lemma bewiesen. \square

Im folgenden sei wieder e der Vektor geeigneter Länge, dessen Komponenten alle gleich 1 sind.

Satz 4.7 *Gegeben seien das lineare Programm*

$$(P_0) \quad \text{Minimiere } c_0^T u \quad \text{auf } M_0 := \{u \in \mathbb{R}^l : u \geq 0, A_0 u \geq b_0\}$$

und das hierzu duale lineare Programm

$$(D_0) \quad \text{Maximiere } b_0^T v \text{ auf } N_0 := \{v \in \mathbb{R}^k : v \geq 0, A_0^T v \leq c_0\}.$$

Hierbei seien $A_0 \in \mathbb{R}^{k \times l}$, $b_0 \in \mathbb{R}^k$ und $c_0 \in \mathbb{R}^l$. Über (P_0) werde vorausgesetzt:

- (a) Es ist $b_0 \neq 0$ oder $c_0 \neq 0$.
- (b) Die Menge der (optimalen) Lösungen von (P_0) ist nicht leer und beschränkt.
- (c) Es existiert ein $\bar{u} \in \mathbb{R}^l$ mit $\bar{u} \geq 0$ und $A_0 \bar{u} > b_0$.

Man setze $m := k + l + 1$, $n := 2m$, definiere

$$B := \begin{pmatrix} A_0 & -I & 0 & 0 \\ 0 & 0 & A_0^T & I \\ c_0^T & 0^T & -b_0^T & 0^T \end{pmatrix} \in \mathbb{R}^{m \times 2(m-1)}, \quad d := \begin{pmatrix} b_0 \\ c_0 \\ 0 \end{pmatrix} \in \mathbb{R}^m,$$

und bilde das lineare Programm

$$(P) \quad \left\{ \begin{array}{l} \text{Minimiere } \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \alpha \\ \beta \end{pmatrix} \text{ auf} \\ M := \left\{ \begin{pmatrix} p \\ \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^n : \begin{pmatrix} B & -d & d - Be \\ e^T & 1 & 1 \end{pmatrix} \begin{pmatrix} p \\ \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ n \end{pmatrix}, \begin{pmatrix} p \\ \alpha \\ \beta \end{pmatrix} \geq 0 \right\}. \end{array} \right.$$

Dann gilt:

1. Das lineare Programm (P) ist in Karmarkar-Normalform.
2. Das lineare Programm (P) ist im folgenden Sinne äquivalent zu (P_0) und (D_0) : Sei $u^* \in M_0$ eine Lösung von (P_0) und $v^* \in N_0$ eine Lösung von (D_0) . Dann ist

$$x^* := \frac{n}{e^T w^* + 1} \begin{pmatrix} w^* \\ 1 \\ 0 \end{pmatrix} \in M \quad \text{mit} \quad w^* := \begin{pmatrix} u^* \\ A_0 u^* - b_0 \\ v^* \\ c_0 - A_0^T v^* \end{pmatrix}$$

eine Lösung von (P) .

Ist umgekehrt $x^* \in M$ eine Lösung von (P) und $x^{*T} = (p^{*T}, \alpha^*, \beta^*)$, so ist $\beta^* = 0$ und $\alpha^* > 0$. Durch $w^* := (1/\alpha^*) p^*$ ist eine Lösung von $Bw = d$, $w \geq 0$ gegeben. Lösungen u^* von (P_0) und v^* von (D_0) entnimmt man aus $w^{*T} = (u^{*T}, y^{*T}, v^{*T}, z^{*T})$.

Beweis: Zur Abkürzung setze man $A := (B \ -d \ d - Be) \in \mathbb{R}^{m \times n}$. Um nachzuweisen, daß (P) ein lineares Programm in Karmarkar-Normalform ist, haben wir $\text{Rang}(A) = m$, $Ae = 0$ und $\min(P) = 0$ zu zeigen. Wie man leicht nachweist, ist $\text{Kern}(B^T) = \{0\}$ bzw. $\text{Rang}(B) = m$, falls nicht $b_0 = 0$ und $c_0 = 0$, was durch die

Voraussetzung (a) ausgeschlossen ist. Daher ist $\text{Rang}(A) = \text{Rang}(B) = m$. Ferner ist

$$Ae = Be + (-d) \cdot 1 + (d - Be) \cdot 1 = 0$$

(man beachte, daß e ein Vektor geeigneter Länge ist, dessen Komponenten alle gleich 1 sind).

Nach Voraussetzung (b) besitzt (P_0) eine Lösung $u^* \in M_0$. Wegen des starken Dualitätssatzes besitzt das duale Problem (D_0) eine Lösung $v^* \in N_0$ und es ist $c_0^T u^* = b_0^T v^*$. Definiert man

$$x^* := \frac{n}{e^T w^* + 1} \begin{pmatrix} w^* \\ 1 \\ 0 \end{pmatrix} \quad \text{mit} \quad w^* := \begin{pmatrix} u^* \\ A_0 u^* - b_0 \\ v^* \\ c_0 - A_0^T v^* \end{pmatrix},$$

so ist $Bw^* = d$, $w^* \geq 0$ und daher x^* für (P) zulässig mit Zielfunktionswert 0, folglich eine Lösung von (P) . Insbesondere ist $\min(P) = 0$.

Nun sei $x^{*T} = (p^{*T}, \alpha^*, \beta^*)$ eine Lösung von (P) . Wegen $\min(P) = 0$ folgt $\beta^* = 0$. Angenommen, es wäre $\alpha^* = 0$. Dann wäre

$$Bp^* = 0, \quad e^T p^* = n, \quad p^* \geq 0,$$

das System $Bp = 0$, $p \geq 0$ hätte also die nichttriviale Lösung p^* . Man zerlege $p^* \in \mathbb{R}^{2(m-1)} = \mathbb{R}^{2(k+l)} = \mathbb{R}^{l+k+k+l}$ durch $p^{*T} = (u^T, y^T, v^T, z^T)$. Unter Benutzung der Definition von B erhält man

$$A_0 u - y = 0, \quad A_0^T v + z = 0, \quad c_0^T u - b_0^T v = 0.$$

Wäre $u \neq 0$, so wäre $c_0^T u > 0$ wegen Lemma 4.6. Dann wäre $0 < c_0^T u = b_0^T v$, zusammen mit $A_0^T v \leq 0$ würde $\sup(D_0) = +\infty$ folgen, ein Widerspruch zur Lösbarkeit von (D_0) . Damit ist $u = 0$, folglich auch $y = 0$ sowie $b_0^T v = c_0^T u = 0$. Zu zeigen bleibt $v = 0$, da sich dann auch $z = 0$ und insgesamt $p^* = 0$ ergibt. Wäre $v \neq 0$, so folgte mit Voraussetzung (c)

$$0 < v^T (\underbrace{A_0 \bar{u} - b_0}_{>0}) = \underbrace{\bar{u}^T A_0^T v}_{\leq 0} - \underbrace{b_0^T v}_{=0} \leq 0,$$

ein Widerspruch. Insgesamt ist die Annahme, daß $\alpha^* = 0$ ist bzw. eine nichttriviale Lösung p^* zu $Bp = 0$, $p \geq 0$ existiert, zum Widerspruch geführt worden. Der Rest der Behauptung ist trivial, er ergibt sich aus dem schwachen Dualitätssatz. Der Satz ist damit vollständig bewiesen. \square

Bemerkung: Man erkennt, daß die Überführung des linearen Programms (P_0) in Karmarkar-Normalform (P) in mehreren Schritten erfolgt. Durch Kombination von (P_0) mit seinem dualen Programm (D_0) erhält man zunächst die Aufgabe

(P_1)

Bestimme eine Lösung von $Bw = d$, $w \geq 0$.

Ähnlich wie bei der Phase I des Simplexverfahrens kann man diesem Zulässigkeitsproblem durch Einführung einer künstlichen Variablen λ das lineare Programm

$$(P_2) \quad \text{Minimiere } \lambda \text{ auf } \left\{ \begin{pmatrix} w \\ \lambda \end{pmatrix} : \begin{pmatrix} B & d - Be \end{pmatrix} \begin{pmatrix} w \\ \lambda \end{pmatrix} = d, \begin{pmatrix} w \\ \lambda \end{pmatrix} \geq 0 \right\}$$

zuordnen. (P_2) besitzt als zulässiges Programm, dessen Zielfunktion nach unten (durch 0) beschränkt ist, eine Lösung. (P_1) (und damit auch (P_0) sowie (D_0)) ist genau dann lösbar, wenn $\min(P_2) = 0$. Der letzte Schritt dient dazu, das lineare Programm (P_2) zu „homogenisieren“ und endgültig auf Karmarkar-Normalform (P) zu bringen. \square

Aufgaben

1. Man beweise Lemma 4.1.

2. Man löse die Optimierungsaufgaben

$$\text{Minimiere } c^T x \text{ auf } \Sigma := \{x \in \mathbb{R}^n : e^T x = n, x \geq 0\}$$

sowie

$$\text{Minimiere } c^T x \text{ auf } K[e; \delta] := \{x \in \mathbb{R}^n : e^T x = n, \|x - e\|_2 \leq \delta\}$$

mit gegebenen $c \in \mathbb{R}^n$ und $\delta > 0$.

3. Bei gegebenem $\alpha \in (0, 1)$ und $r := \sqrt{n/(n-1)}$ betrachte man die Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(y) := \prod_{j=1}^n y_j \text{ auf } K[e; \alpha r] := \{y \in \mathbb{R}^n : e^T y = n, \|y - e\|_2 \leq \alpha r\}.$$

Man zeige:

(a) (P) besitzt eine Lösung y^* und es ist notwendig $y^* > 0$ und $\|y^* - e\|_2 = \alpha r$.

(b) Sei y^* eine Lösung von (P) mit $y_1^* \leq y_2^* \leq \dots \leq y_n^*$ (man beachte, daß das Permutieren der Komponenten einer Lösung wieder zu einer Lösung führt). Dann ist $y_2^* = \dots = y_n^*$.

(c) Ist $e^T y^* = n$, $\|y^* - e\|_2 = \alpha r$ und $y_1^* \leq y_2^* = \dots = y_n^*$, so ist notwendig

$$y_1^* = 1 - \alpha, \quad y_j^* = 1 + \frac{\alpha}{n-1} \quad \text{für } j = 2, \dots, n.$$

(d) Es ist

$$\prod_{j=1}^n y_j \geq (1 - \alpha) \left(1 + \frac{\alpha}{n-1}\right)^{n-1} \quad \text{für alle } y \in K[e; \alpha r].$$

4. Zur Lösung von

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$$

mit $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $\text{Rang}(A) = m$ betrachte man den folgenden Algorithmus (siehe E. R. Barnes (1986)):

- Seien $\alpha \in (0, 1)$ und $x^0 \in M$ mit $x^0 > 0$ gegeben.
- Für $k = 0, 1, \dots$:
 - Setze $D_k := \text{diag}(x_1^k, \dots, x_n^k)$, berechne $y^k := (AD_k^2 A^T)^{-1} AD_k^2 c$.
 - Falls $c - A^T y^k = 0$, dann: STOP, x^k ist Lösung von (P).
 - Berechne $x^{k+1} := x^k - \alpha \frac{D_k^2(c - A^T y^k)}{\|D_k(c - A^T y^k)\|_2}$.

Man zeige: Bricht der Algorithmus nicht vorzeitig mit einer Lösung von (P) ab, so liefert er eine Folge $\{x^k\} \subset M$ mit

$$x^k > 0, \quad c^T x^{k+1} = c^T x^k - \alpha \|D_k(c - A^T y^k)\|_2 \quad \text{für } k = 0, 1, \dots$$

Ferner zeige man: Ist $x^k \in M$ mit $x^k > 0$ und $c - A^T y^k \neq 0$, so ist x^{k+1} die Lösung des Hilfsproblems

$$\text{Minimiere } c^T x \text{ unter den Nebenbedingungen } Ax = b, \|D_k(x - x^k)\|_2 \leq \alpha.$$

5. Man stelle einen Satz 4.7 entsprechenden Satz für den Fall auf, daß das Ausgangsproblem in Simplex-Normalform gegeben ist:

$$(P_0) \quad \text{Minimiere } c_0^T u \text{ auf } M_0 := \{u \in \mathbb{R}^l : u \geq 0, A_0 u = b_0\},$$

wobei $A_0 \in \mathbb{R}^{k \times l}$, $b_0 \in \mathbb{R}^k$ und $c_0 \in \mathbb{R}^l$.

Hinweis: Hierzu kann man D. GOLDFARB, S. MEHROTRA (1988b) konsultieren.

6. Gegeben sei das lineare Programm in Karmarkar-Normalform

$$(P) \quad \text{Minimiere } c^T x \text{ auf } M := \left\{ x \in \mathbb{R}^n : \begin{pmatrix} A \\ e^T \end{pmatrix} x = \begin{pmatrix} 0 \\ n \end{pmatrix}, x \geq 0 \right\},$$

es sei also $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ mit $n \geq 2$ und

$$(V) \quad \text{Rang}(A) = m, \quad Ae = 0, \quad \min(P) = 0.$$

Man programmiere das Karmarkar-Verfahren und löse hiermit, unter Benutzung von Satz 4.7, das lineare Programm (siehe das Beispiel auf Seite 85)

$$(P_0) \quad \left\{ \begin{array}{ll} \text{Minimiere} & -x_1 - x_2 \text{ unter den Nebenbedingungen} \\ & x_1 + 3x_2 \leq 13, \\ x_1 \geq 0, & x_2 \geq 0, \quad 3x_1 + x_2 \leq 15, \\ & -x_1 + x_2 \leq 3. \end{array} \right.$$

Kapitel 7

Unrestringierte Optimierungsaufgaben

In diesem Kapitel beschäftigen wir uns mit einer scheinbar sehr einfachen Klasse nichtlinearer Optimierungsaufgaben, nämlich solchen ohne Nebenbedingungen bzw. Restriktionen. Kürzer spricht man auch von *unrestringierten Optimierungsaufgaben*. Hier ist eine Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ gegeben. Das Problem besteht darin, einen Punkt zu finden, in dem diese Zielfunktion minimal ist, wofür wir

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

schreiben werden. Man unterscheidet zwischen

- einer *globalen Lösung* von (P), d. h. einem Punkt $x^* \in \mathbb{R}^n$ mit $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$,
- und einer *lokalen Lösung* von (P), d. h. einem Punkt $x^* \in \mathbb{R}^n$, zu dem es eine Umgebung U^* gibt, so daß $f(x^*) \leq f(x)$ wenigstens für alle $x \in U^*$ ist.

Wir werden uns damit begnügen, mit Hilfe von Iterationsverfahren lokale Lösungen von (P) (oder wenigstens solche Punkte, die als lokale Lösungen in Frage kommen), zu berechnen bzw. zu approximieren.

Unrestringierte Optimierungsaufgaben kommen in vielen Bereichen vor, wobei *nichtlineare Ausgleichsprobleme* bzw. *diskrete, nichtlineare Approximationsaufgaben* besonders wichtig sind. Wie in anderen Gebieten der numerischen Mathematik kann man auch in der unrestringierten Optimierung nicht von *dem* besten Lösungsverfahren sprechen. Die Auswahl eines geeigneten Verfahrens wird u. a. von der „Glattheit“ der Zielfunktion f (liegt diese samt ihren Ableitungen analytisch vor, wie teuer ist die Auswertung?) und n , der Anzahl der Variablen, abhängen. Wir werden versuchen, einen Einblick in das reizvolle Gebiet der numerischen Behandlung unrestringierter Optimierungsaufgaben zu geben. Als Literatur hierzu empfehlen wir u. a. R. FLETCHER (1987), P. E. GILL, W. MURRAY, M. H. WRIGHT (1981), J. E. DENNIS, R. B. SCHNABEL (1983) sowie einen Übersichtsartikel von J. E. DENNIS, R. B. SCHNABEL (1989).

7.1 Grundlagen

7.1.1 Einführung

Machen wir uns die Aufgabenstellung bei der unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

anschaulich klar! Angenommen, wir sind im Gebirge und ein dichter Nebel überrascht uns. Wir wollen ins Tal¹, sagen wir zum niedrigsten Punkt. Wegen des Nebels können wir nur die unmittelbare Umgebung unseres jeweiligen Standpunktes erfassen. Eine naheliegende Strategie besteht darin, zunächst einmal zu prüfen, ob es eine von dem aktuellen Punkt ausgehende Richtung gibt, in der es sofort wenigstens „ein kleines Stück“ abwärts geht. Mathematisch gesprochen: Wir sind im Punkt $x \in \mathbb{R}^n$ und fragen, ob es eine Richtung $p \in \mathbb{R}^n$ und ein $t_0 > 0$ mit $f(x + tp) < f(x)$ für alle $t \in (0, t_0]$ gibt. Wenn der Nebel sehr dicht ist, so werden wir sogar nur den Wert²

$$f'(x; p) := \lim_{t \rightarrow 0+} \frac{f(x + tp) - f(x)}{t}$$

für jede Richtung p , in die wir gehen dürfen (die unrestringierte Optimierung zeichnet sich dadurch aus, daß man in *jede* Richtung gehen darf, ohne den „Zulässigkeitsbereich“ zu verlassen), „ertasten“ können. Können wir eine Richtung p mit $f'(x; p) < 0$ finden, so sind wir sicher, nicht in einer „Mulde“ gelandet, geschweige denn am Ziel zu sein, und werden daher einen oder mehrere Schritte in eine solche Richtung gehen, die uns weiter abwärts führt. D. h. wir wählen eine sogenannte *Abstiegsrichtung*, also ein $p \in \mathbb{R}^n$ mit $f'(x; p) < 0$, und anschließend eine *Schrittweite* $t > 0$ mit $f(x + tp) < f(x)$. Dann ist $x_+ := x + tp$ unser neuer aktueller Standpunkt und wir können mit der Suche nach der Talstation fortfahren.

Nehmen wir nun an, wir seien an einem Punkt x^* angelangt, für den

$$f'(x^*; p) \geq 0 \quad \text{für jede Richtung } p$$

gilt. Einen solchen Punkt werden wir *stationär* bzw. eine *stationäre* Lösung von (P) nennen. In diesem Falle sind wir sicher, daß unser aktueller Standpunkt x^* wenigstens eine *notwendige* Bedingung dafür erfüllt, eine „Mulde“ bzw. ein lokales Minimum zu sein. Denn ist x^* eine lokale Lösung von (P) und $p \in \mathbb{R}^n$ eine beliebige Richtung, so ist

$$f(x^* + tp) - f(x^*) \geq 0 \quad \text{für alle hinreichend kleinen } t > 0,$$

woraus nach Division durch t und Grenzübergang $t \rightarrow 0+$ (nach wie vor wird angenommen, daß der Limes existiert) $f'(x^*; p) \geq 0$ folgt. Erst durch eine subtilere Untersuchung der Umgebung wird man feststellen können, ob man wirklich in einem lokal niedrigsten Punkt ist, d. h. ob eine *hinreichende* Bedingung für ein lokales

¹Vernünftigerweise würde man in einem solchen Falle dort bleiben, wo man ist, bis sich der Nebel gelichtet hat. Das würde uns bei der Erläuterung eines Iterationsverfahrens aber nichts nützen!

²Natürlich nehmen wir hier an, daß dieser Limes existiert. Bei einem „glatten“ Gebirge wird das der Fall sein.

Minimum erfüllt ist. Wissen wir dagegen, daß das Tal bzw. die zu minimierende Funktion f *konvex* ist, d. h. daß man zu je zwei Punkten des Gebirges die gesamte Zwischenstrecke überblicken kann bzw. daß

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \quad \text{für alle } x, y \in \mathbb{R}^n, t \in [0, 1]$$

gilt, so können wir uns überlegen, daß wir in unserem stationären Punkt x^* sogar schon am Ziel sind. Hierzu müssen wir die Höhe $f(x)$ eines beliebigen anderen Punktes x mit der Höhe $f(x^*)$ des stationären Punktes x^* vergleichen. Wegen der Konvexität ist

$$f(x^* + t(x - x^*)) \leq (1-t)f(x^*) + tf(x) \quad \text{für alle } t \in [0, 1],$$

woraus

$$\frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \leq f(x) - f(x^*) \quad \text{für alle } t \in (0, 1]$$

und nach Grenzübergang $f'(x^*; x - x^*) \leq f(x) - f(x^*)$ folgt. Ist also x^* stationär und $f: \mathbb{R}^n \rightarrow \mathbb{R}$ konvex, so ist $0 \leq f'(x^*; x - x^*) \leq f(x) - f(x^*)$ für alle $x \in \mathbb{R}^n$ und damit x^* sogar eine globale Lösung von (P).

Damit haben wir einige der grundlegenden Begriffe der unrestringierten Optimierung angesprochen. Die meisten Verfahren bestehen aus einer *Richtungs-* und einer *Schrittweitenstrategie*. Die erste gibt an, in welche Abstiegsrichtung von einem nichtstationären Punkt ausgehend man sich fortbewegt, die zweite sagt aus, wie weit dies zu geschehen hat. Einer zweiten Klasse von Verfahren, den sogenannten *Trust-Region-Verfahren*, liegt eine etwas andere Idee zugrunde. Auf diese Verfahren werden wir später ebenfalls eingehen, da sie in neuerer Zeit insbesondere bei nichtlinearen Ausgleichsproblemen Bedeutung gewonnen haben. Ferner haben wir gesehen, daß *notwendige* und *hinreichende Optimalitätsbedingungen* eine Rolle spielen werden, daß man ferner eventuell mit *stationären* Lösungen zufrieden sein muß und der *Konvexitätsbegriff* besonders wichtig ist. Schließlich wird die *Glattheit* bzw. *Differenzierbarkeit* der Zielfunktion f von Interesse sein, da wir z. B. in obiger Motivation die Existenz von $f'(x; p)$ angenommen haben.

7.1.2 Notwendige Optimalitätsbedingungen erster Ordnung

Wir betrachten bei vorgegebener Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n.$$

Unser Ziel in diesem Unterabschnitt ist es, notwendige Bedingungen erster Ordnung dafür anzugeben, daß ein $x^* \in \mathbb{R}^n$ lokale Lösung von (P) ist. „Erster Ordnung“ bedeutet in diesem Zusammenhang, daß lediglich Ableitungen erster Ordnung von f auftreten.

Definition 1.1 Ist $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ in einem Punkt $x \in \mathbb{R}^n$ stetig partiell differenzierbar, existieren also alle partiellen Ableitungen $\partial F_i / \partial x_j$, $i = 1, \dots, m$, $j = 1, \dots, n$,

aller Komponenten F_i von F in einer Umgebung von x und sind diese in x stetig, so heißt F *in x stetig differenzierbar*. Man nennt

$$F'(x) = \left(\frac{\partial F_i}{\partial x_j}(x) \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \in \mathbb{R}^{m \times n}$$

die *Funktionalmatrix* von F in x . Ist $f: \mathbb{R}^n \rightarrow \mathbb{R}$ in $x \in \mathbb{R}^n$ stetig differenzierbar, so heißt

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^T$$

der *Gradient* von f in x . Die Funktion $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ heißt *stetig differenzierbar auf einer offenen Menge $D \subset \mathbb{R}^n$* , wofür wir kürzer $F \in C^1(D; \mathbb{R}^m)$ schreiben werden, wenn F in jedem Punkt $x \in D$ stetig differenzierbar ist. Statt $C^1(D; \mathbb{R}^1)$ schreiben wir $C^1(D)$.

Beispiel: Bei nichtlinearen Ausgleichsaufgaben nach der Methode der kleinsten Quadrate handelt es sich um die Aufgabe, die Zielfunktion

$$f(x) := \frac{1}{2} \sum_{i=1}^m F_i(x)^2$$

zu minimieren, wobei $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$. Sind alle F_i in x stetig differenzierbar, so auch die Zielfunktion f und es ist

$$\frac{\partial f}{\partial x_j}(x) = \sum_{i=1}^m \frac{\partial F_i}{\partial x_j}(x) F_i(x), \quad j = 1, \dots, n.$$

Faßt man die F_i als Komponenten einer Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ auf (die Zielfunktion f läßt sich dann als $f(x) = \frac{1}{2} \|F(x)\|_2^2$ schreiben), so ist der Gradient von f in x durch

$$\nabla f(x) = F'(x)^T F(x)$$

gegeben. □

Von Zielfunktionen der Form

$$f(x) = \max_{i=1, \dots, m} F_i(x), \quad f(x) = \max_{i=1, \dots, m} |F_i(x)|, \quad f(x) = \sum_{i=1}^m |F_i(x)|$$

kann man nicht erwarten, daß sie stetig differenzierbar sind, selbst dann, wenn es die F_i sind. Daher ist es sinnvoll, auch einen schwächeren Ableitungsbegriff einzuführen.

Definition 1.2 Die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ heißt *in x ∈ Rn in die Richtung p ∈ Rn richtungsdifferenzierbar*, wenn

$$F'(x; p) := \lim_{t \rightarrow 0+} \frac{F(x + tp) - F(x)}{t}$$

existiert. In diesem Falle heißt $F'(x; p)$ die *Richtungsableitung von F in x in Richtung p*. Die Funktion F heißt im Punkte x *richtungsdifferenzierbar*, wenn F in jede Richtung $p \in \mathbb{R}^n$ richtungsdifferenzierbar ist. In diesem Falle nennt man die Abbildung $F'(x; \cdot): \mathbb{R}^n \rightarrow \mathbb{R}$ die *Gateaux-Variation* von F in x .

Natürlich ist eine in einem Punkt x stetig differenzierbare Funktion dort auch richtungsdifferenzierbar. Dies notieren wir in dem folgenden Lemma.

Lemma 1.3 Ist $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ in $x \in \mathbb{R}^n$ stetig differenzierbar, so ist F in x richtungsdifferenzierbar, und die Gateaux-Variation $F'(x; \cdot): \mathbb{R}^n \rightarrow \mathbb{R}^m$ ist durch $F'(x; p) = F'(x)p$ gegeben.

Der folgende Satz liefert eine erste notwendige Optimalitätsbedingung.

Satz 1.4 Sei $x^* \in \mathbb{R}^n$ eine lokale Lösung von (P). Ist f in x^* richtungsdifferenzierbar, so ist $f'(x^*; p) \geq 0$ für jedes $p \in \mathbb{R}^n$. Ist f in x^* sogar stetig differenzierbar, so ist $\nabla f(x^*) = 0$.

Beweis: Da x^* eine lokale Lösung von (P) ist, existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^*$. Bei vorgegebenem $p \in \mathbb{R}^n$ existiert ein $t_0 > 0$ derart, daß $x^* + tp \in U^*$ und daher $f(x^*) \leq f(x^* + tp)$ für alle $t \in [0, t_0]$. Folglich ist

$$f'(x^*; p) = \lim_{t \rightarrow 0+} \frac{f(x^* + tp) - f(x^*)}{t} \geq 0 \quad \text{für alle } p \in \mathbb{R}^n.$$

Ist f in x^* sogar stetig differenzierbar, so ist $f'(x^*; p) = \nabla f(x^*)^T p \geq 0$ für alle $p \in \mathbb{R}^n$ und (setze z. B. $p := -\nabla f(x^*)$) daher $\nabla f(x^*) = 0$. \square

Die nächste Definition ist in der Einführung schon angekündigt worden.

Definition 1.5 Ist $f: \mathbb{R}^n \rightarrow \mathbb{R}$ in $x^* \in \mathbb{R}^n$ richtungsdifferenzierbar mit Gateaux-Variation $f'(x^*; \cdot)$, so heißt x^* ein *stationärer Punkt* von f oder eine *stationäre Lösung* von (P), wenn $f'(x^*; p) \geq 0$ für alle $p \in \mathbb{R}^n$.

Wegen Satz 1.4 ist eine lokale Lösung von (P), in der f richtungsdifferenzierbar ist, notwendig eine stationäre Lösung.

Beispiel: Eine beliebte Testfunktion bei unrestringierten Optimierungsaufgaben ist die sogenannte *Rosenbrock-Funktion*

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Zeichnet man sich bei gegebenem $c > 0$ die Niveaulinien $\{x \in \mathbb{R}^2 : f(x) = c\}$ auf, so erkennt man, daß die Suche nach einem Minimum von f der Suche nach dem tiefsten Punkt in einem langgestreckten, „bananenförmigen“ Tal entspricht.

Als Gradienten von f berechnet man

$$\nabla f(x) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}.$$

Der einzige stationäre Punkt von f ist offenbar $x^* = (1, 1)$. Dieser Punkt ist sogar die eindeutige globale Lösung der zugehörigen unrestringierten Optimierungsaufgabe. Dies ist ein besonders glücklicher Umstand und keineswegs die Regel. \square

Wie wir uns schon in der Einführung klar machen, ist ein stationärer Punkt einer *konvexen Funktion* sogar eine globale Lösung der zugehörigen unrestringierten Optimierungsaufgabe. Hierbei heißt bekanntlich eine Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ *konvex* auf einer konvexen Menge $D \subset \mathbb{R}^n$, wenn

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \quad \text{für alle } x, y \in D \text{ und } t \in [0, 1].$$

Wir wollen zwar im folgenden den Schwerpunkt auf „glatte“, unrestringierte Optimierungsaufgaben legen, aber gleichzeitig einige Grundlagen für sogenannte „nichtglatte“ (oder besser: „halbglatte“) Aufgaben bereitstellen. Interessant in diesem Zusammenhang ist das folgende Lemma.

Lemma 1.6 Sei $f: \mathbb{R}^n \rightarrow \mathbb{R}$ (auf dem \mathbb{R}^n) *konvex*. Dann ist f in jedem $x \in \mathbb{R}^n$ *richtungsdifferenzierbar* und es ist

$$f'(x; p) \leq f(x + p) - f(x) \quad \text{für alle } x, p \in \mathbb{R}^n.$$

Die Gateaux-Variation $f'(x; \cdot): \mathbb{R}^n \rightarrow \mathbb{R}$ von f in x ist *nichtnegativ homogen*, d. h.

$$f'(x; \alpha p) = \alpha f'(x; p) \quad \text{für alle } \alpha \geq 0, p \in \mathbb{R}^n,$$

subadditiv, d. h.

$$f'(x; p + q) \leq f'(x; p) + f'(x; q) \quad \text{für alle } p, q \in \mathbb{R}^n$$

und *konvex*.

Beweis: Zu $x, p \in \mathbb{R}^n$ definiere man $\phi: (0, 1] \rightarrow \mathbb{R}$ durch $\phi(t) := [f(x + tp) - f(x)]/t$. Wir zeigen die Existenz von $\lim_{t \rightarrow 0+} \phi(t)$, indem wir nachweisen, daß ϕ auf $(0, 1]$ nach unten beschränkt und monoton nicht fallend ist.

Wegen der Konvexität von f ist

$$f(x) = f\left(\frac{1}{1+t}(x + tp) + \frac{t}{1+t}(x - p)\right) \leq \frac{1}{1+t}f(x + tp) + \frac{t}{1+t}f(x - p)$$

für alle $t \in (0, 1]$, woraus $f(x) - f(x - p) \leq \phi(t)$ für alle $t \in (0, 1]$ folgt. Ist ferner $0 < s \leq t \leq 1$, so ist

$$f(x + sp) - f(x) = f\left(\frac{s}{t}(x + tp) + \frac{t-s}{t}x\right) - f(x) \leq \frac{s}{t}[f(x + tp) - f(x)]$$

wieder wegen der Konvexität von f , woraus $\phi(s) \leq \phi(t)$ folgt. Die Existenz von $f'(x; p) = \lim_{t \rightarrow 0+} \phi(t)$ ist damit gesichert. Wegen $\phi(t) \leq \phi(1)$ erhalten wir insbesondere $f'(x; p) \leq \phi(1) = f(x + p) - f(x)$.

Der Nachweis der restlichen Behauptungen ist einfach, er bleibt dem Leser überlassen. \square

Wegen Lemma 1.6 wissen wir, daß konvexe Funktionen eine Gateaux-Variation besitzen. Diese aber im konkreten Fall auszurechnen, kann schwieriger sein. Vektornormen auf dem \mathbb{R}^n sind prominente Vertreter konvexer Funktionen. Im folgenden Satz wird die Gateaux-Variation der Maximumsnorm berechnet.

Satz 1.7 Sei $f: \mathbb{R}^n \rightarrow \mathbb{R}$ durch $f(x) := \|x\|_\infty = \max_{j=1,\dots,n} |x_j|$ definiert. Dann ist f in jedem $x \in \mathbb{R}^n$ rückwärts differenzierbar, die Gateaux-Variation ist durch

$$f'(x; p) = \begin{cases} \|p\|_\infty & \text{falls } x = 0, \\ \max_{j \in J(x)} \operatorname{sign}(x_j)p_j & \text{falls } x \neq 0 \end{cases}$$

mit $J(x) := \{j \in \{1, \dots, n\} : |x_j| = \|x\|_\infty\}$ gegeben.

Beweis: Wir nehmen o. B. d. A. $x \neq 0$ an. Sei eine Richtung $p \in \mathbb{R}^n$ vorgegeben. Aus Stetigkeitsgründen ist $|x_j + tp_j| < \|x + tp\|_\infty$ für alle $j \notin J(x)$ und alle hinreichend kleinen $t > 0$. Daher ist

$$\frac{f(x + tp) - f(x)}{t} = \max_{j \in J(x)} \frac{|x_j + tp_j| - |x_j|}{t} = \max_{j \in J(x)} \operatorname{sign}(x_j)p_j$$

für alle hinreichend kleinen $t > 0$, woraus nach dem Grenzübergang $t \rightarrow 0+$ die Behauptung folgt. \square

Bemerkung: Ähnlich kann die Gateaux-Variation der durch

$$f(x) := \max_{j=1,\dots,n} x_j \quad \text{bzw.} \quad f(x) := \|x\|_1 = \sum_{j=1}^n |x_j|$$

definierten konvexen Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ berechnet werden (siehe Aufgabe 1). \square

Die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n,$$

bei der $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine i. allg. nichtlineare, glatte Abbildung und $\|\cdot\|$ eine Vektornorm auf dem \mathbb{R}^m ist, nennt man eine *diskrete, nichtlineare Approximationsaufgabe*. Wichtigste Spezialfälle sind $\|\cdot\| = \|\cdot\|_2$, $\|\cdot\| = \|\cdot\|_\infty$ (dann spricht man von einer *diskreten Tschebyscheffschen Approximationsaufgabe*) und $\|\cdot\| = \|\cdot\|_1$ (*diskrete L₁-Approximationsaufgabe*). I. allg. wird hier $m > n$ sein, so daß man (P) als die Aufgabe auffassen kann, den Defekt des überbestimmten nichtlinearen Gleichungssystems $F(x) = 0$ bezüglich der Norm $\|\cdot\|$ zu minimieren. Hierdurch und durch die sogenannte (unrestringierte) *Min-Max-Optimierungsaufgabe*, bei der die Zielfunktion f die Form $f(x) := \max_{i=1,\dots,m} F_i(x)$ hat, sind die wohl wichtigsten „halbglatten“ Optimierungsaufgaben gegeben. Ein Überblick (Theorie und Algorithmen) über genau diese Klasse von Optimierungsaufgaben wird bei R. GONIN, A. H. MONEY (1989) gegeben. Wenn die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ bzw. die Komponentenfunktionen F_i , $i = 1, \dots, m$, hinreichend glatt sind, kann man immer noch hoffen, daß die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) rückwärts differenzierbar ist. Diese Vermutung soll im folgenden Satz für die bei der diskreten Tschebyscheff-Approximation auftretende Zielfunktion exemplarisch nachgewiesen werden.

Satz 1.8 Mit der Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ sei $f: \mathbb{R}^n \rightarrow \mathbb{R}$ durch

$$f(x) := \|F(x)\|_\infty = \max_{i=1,\dots,m} |F_i(x)|$$

definiert. Dann gilt: Ist F in $x^* \in \mathbb{R}^n$ stetig differenzierbar, so ist f in x^* richtungs-differenzierbar mit der Gateaux-Variation

$$f'(x^*; p) = \begin{cases} \max_{i=1,\dots,m} |\nabla F_i(x^*)^T p| & \text{falls } F(x^*) = 0, \\ \max_{i \in I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p & \text{falls } F(x^*) \neq 0, \end{cases}$$

wobei $I(x^*) := \{i \in \{1, \dots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}$.

Beweis: Mit $g: \mathbb{R}^m \rightarrow \mathbb{R}$, definiert durch $g(y) = \|y\|_\infty = \max_{i=1,\dots,m} |y_i|$, ist $f = g \circ F$. Die Abbildung F ist nach Voraussetzung in x^* stetig differenzierbar, während g eine konvexe Funktion ist, deren Gateaux-Variation $g'(y; \cdot)$ wir wegen Satz 1.7 kennen. Behauptet wird, daß f in x^* die Gateaux-Variation

$$f'(x^*; p) = g'(F(x^*); F'(x^*)p)$$

besitzt. Hieran erkennt man sehr deutlich, daß der Aussage des Satzes eine *Kettenregel* zugrunde liegt. Wir wollen den Beweis so führen, daß dieser Zusammenhang deutlich wird.

Seien $p \in \mathbb{R}^n$ und eine Nullfolge $\{t_k\} \subset \mathbb{R}_+$ vorgegeben. Definiert man

$$r_k := F(x^* + t_k p) - F(x^*) - t_k F'(x^*)p,$$

so gilt $\lim_{k \rightarrow \infty} r_k / t_k = 0$, da F als in x^* stetig differenzierbare Funktion insbesondere in x^* auch (total) differenzierbar ist. Damit erhält man

$$\begin{aligned} \frac{f(x^* + t_k p) - f(x^*)}{t_k} &= \frac{g \circ F(x^* + t_k p) - g \circ F(x^*)}{t_k} \\ &= \frac{g(F(x^*) + t_k F'(x^*)p + r_k) - g(F(x^*))}{t_k} \\ &= \frac{g(F(x^*) + t_k F'(x^*)p) - g(F(x^*))}{t_k} \\ &\quad + \frac{g(F(x^*) + t_k F'(x^*)p + r_k) - g(F(x^*) + t_k F'(x^*)p)}{t_k}. \end{aligned}$$

Da g in $F(x^*)$ eine Gateaux-Variation $g'(F(x^*); \cdot)$ besitzt, konvergiert der erste Summand gegen $g'(F(x^*); F'(x^*)p)$. Greifen wir auf die Definition von g als Maximum-norm auf dem \mathbb{R}^m zurück, so erhalten wir für den Betrag des zweiten Summanden

$$\left| \frac{\|F(x^*) + t_k F'(x^*)p + r_k\|_\infty - \|F(x^*) + t_k F'(x^*)p\|_\infty}{t_k} \right| \leq \frac{\|r_k\|_\infty}{t_k}.$$

Wegen $\lim_{k \rightarrow \infty} r_k / t_k = 0$ konvergiert dieser zweite Summand gegen Null. Daher ist $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ bewiesen, woraus die Behauptung des Satzes folgt. \square

Bemerkung: Eine etwas genauere Inspektion des Beweises von Satz 1.8 zeigt, daß neben der stetigen Differenzierbarkeit von F die Existenz der Gateaux-Variation und die Lipschitzstetigkeit der durch $g(y) := \max_{i=1,\dots,m} |y_i|$ definierten Abbildung

$g: \mathbb{R}^m \rightarrow \mathbb{R}$ entscheidend für die Existenz der Gateaux-Variation von $f = g \circ F$ in x^* sowie die Gültigkeit der Kettenregel $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ sind. Analog kann die Existenz der Gateaux-Variation weiterer halbglatte Zielfunktionen $f = g \circ F$ nachgewiesen werden, siehe z. B. Aufgabe 3. \square

Wir fassen die notwendigen Optimalitätsbedingungen (erster Ordnung) für eine lokale Lösung einer diskreten Tschebyscheffschen Approximationsaufgabe in dem folgenden Satz zusammen.

Satz 1.9 Gegeben sei die diskrete Tschebyscheffsche Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty = \max_{i=1,\dots,m} |F_i(x)|, \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, seien in $x^* \in \mathbb{R}^n$ stetig differenzierbar und es sei $F(x^*) \neq 0$. Dann gilt: Ist x^* eine lokale Lösung von (P), so ist x^* auch eine stationäre Lösung von (P), d. h. es gilt

$$(*) \quad f'(x^*; p) = \max_{i \in I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p \geq 0 \quad \text{für alle } p \in \mathbb{R}^n,$$

wobei $I(x^*) := \{i \in \{1, \dots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}$. Ferner ist (*) äquivalent zu der Existenz von reellen Zahlen λ_i^* , $i \in I(x^*)$, mit

$$(**) \quad \lambda_i^* \geq 0 \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* = 1, \quad \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla F_i(x^*) = 0.$$

Beweis: Eine lokale Lösung von (P) ist notwendig eine stationäre Lösung. Wegen Satz 1.8 ist für eine lokale Lösung x^* also notwendig (*) erfüllt. Die Äquivalenz von (*) und (**) ist daher die eigentliche Aussage dieses Satzes.

Zunächst wollen wir die einfache Richtung beweisen und nehmen an, (**) würde gelten. Für alle $p \in \mathbb{R}^n$ ist dann

$$0 = \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p \leq \max_{i \in I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p,$$

es gilt also (*).

Der Beweis der umgekehrten Richtung ist schwieriger, hier kommt man nicht ohne etwas tiefere Hilfsmittel aus. Jetzt nehmen wir also an, (*) würde gelten bzw. x^* sei eine stationäre Lösung von (P). Zum Beweis benutzen wir das *Farkas-Lemma* (siehe Lemma 3.4 in Abschnitt 6.3). Um dem Leser das Blättern zu ersparen, zitieren wir es hier:

- Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Dann besitzt das System $Ax = b$, $x \geq 0$ genau dann keine Lösung, wenn das System $A^T y \leq 0$, $b^T y > 0$ eine Lösung besitzt.

Sei $q \geq 1$ die Anzahl der Elemente in der nichtleeren Indexmenge $I(x^*)$ und B die $n \times q$ -Matrix, deren Spalten $\text{sign}[F_i(x^*)] \nabla F_i(x^*)$, $i \in I(x^*)$, sind. Ferner sei $e := (1, \dots, 1)^T \in \mathbb{R}^q$. Die Annahme, daß (**) nicht gilt, bedeutet dann gerade, daß

$$\begin{pmatrix} B \\ e^T \end{pmatrix} \lambda = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \lambda \geq 0$$

nicht lösbar ist. Das Farkas-Lemma liefert, daß das System

$$B^T p + \gamma e = \begin{pmatrix} B^T & e \end{pmatrix} \begin{pmatrix} p \\ \gamma \end{pmatrix} \leq 0, \quad \gamma = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \gamma \end{pmatrix} > 0$$

lösbar ist. Die Zeilen von B^T sind $\text{sign}[F_i(x^*)]\nabla F_i(x^*)^T$. Daher erhalten wir die Existenz eines $p \in \mathbb{R}^n$ und einer positiven Zahl γ mit $B^T p \leq -\gamma e < 0$ bzw.

$$\text{sign}[F_i(x^*)]\nabla F_i(x^*)^T p \leq -\gamma < 0 \quad \text{für alle } i \in I(x^*),$$

ganz offensichtlich ein Widerspruch zu (*). Der Satz ist damit bewiesen. \square

Bemerkung: Entsprechend zu Satz 1.9 können auch stationäre Lösungen für die Min-Max-Aufgabe sowie die diskrete L_1 -Approximationsaufgabe charakterisiert werden (siehe die Aufgaben 4 und 5). \square

7.1.3 Notwendige und hinreichende Optimalitätsbedingungen zweiter Ordnung

Nachdem wir uns bisher mit *notwendigen Optimalitätsbedingungen erster Ordnung* (es gehen nur Ableitungen erster Ordnung ein) beschäftigt haben, kommen wir nun zu notwendigen und hinreichenden Optimalitätsbedingungen zweiter Ordnung für lokale Lösungen der unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n.$$

Zunächst betrachten wir den „glatten“ Fall und definieren analog zu 1.1:

Definition 1.10 Ist $f: \mathbb{R}^n \rightarrow \mathbb{R}$ in einem Punkt $x \in \mathbb{R}^n$ zweimal stetig partiell differenzierbar, existieren also alle partiellen Ableitungen $\partial^2 f / \partial x_i \partial x_j$, $i, j = 1, \dots, n$, in einer Umgebung von x und sind diese in x stetig, so heißt f in x *zweimal stetig differenzierbar*. Man nennt die symmetrische Matrix

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$$

die *Hessesche* von f in x . Die Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ heißt *zweimal stetig differenzierbar auf der offenen Menge $D \subset \mathbb{R}^n$* , wofür wir kürzer $f \in C^2(D)$ schreiben werden, wenn f in jedem Punkt $x \in D$ zweimal stetig differenzierbar ist.

Beispiel: Für

$$f(x) := \frac{1}{2} \sum_{k=1}^m F_k(x)^2 = \frac{1}{2} \|F(x)\|_2^2$$

mit in $x \in \mathbb{R}^n$ zweimal stetig differenzierbaren $F_k: \mathbb{R}^n \rightarrow \mathbb{R}$, $k = 1, \dots, m$, erhält man

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \sum_{k=1}^m \frac{\partial F_k}{\partial x_i}(x) \frac{\partial F_k}{\partial x_j}(x) + \sum_{k=1}^m \frac{\partial^2 F_k}{\partial x_i \partial x_j}(x) F_k(x),$$

so daß

$$\nabla^2 f(x) = F'(x)^T F'(x) + \sum_{k=1}^m F_k(x) \nabla^2 F_k(x)$$

die Hessesche von f ist. \square

Aus der Analysis (siehe z. B. O. FORSTER (1984, S. 61 ff.)) sind die folgenden notwendigen und hinreichenden Optimalitätsbedingungen zweiter Ordnung bekannt.

Satz 1.11 Die Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ sei auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Dann gilt:

1. Ist x^* eine lokale Lösung von (P), so ist $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ ist (symmetrisch und) positiv semidefinit.
2. Ist $\nabla f(x^*) = 0$ und ist $\nabla^2 f(x^*)$ positiv definit, so ist x^* eine isolierte, lokale Lösung von (P), d. h. es gibt eine Umgebung U^* von x^* mit $f(x^*) < f(x)$ für alle $x \in U^* \setminus \{x^*\}$.

Vom mathematischen Standpunkt aus interessanter sind notwendige und hinreichende Optimalitätsbedingungen zweiter Ordnung für „nichtglatte“, unrestringierte Optimierungsaufgaben. Exemplarisch wollen wir die diskrete Tschebyscheffsche Approximationsaufgabe betrachten und eine hinreichende Optimalitätsbedingung zweiter Ordnung im folgenden Satz formulieren und anschließend beweisen.

Satz 1.12 Gegeben sei die diskrete Tschebyscheffsche Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty = \max_{i=1,\dots,m} |F_i(x)|, \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar und es sei $F(x^*) \neq 0$. Sei

$$I(x^*) := \{i \in \{1, \dots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}.$$

Es wird vorausgesetzt, daß reelle Zahlen λ_i^* , $i \in I(x^*)$, existieren mit:

1. Es ist

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* = 1, \quad \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) = 0,$$

d. h. in x^* ist die notwendige Optimalitätsbedingung (**) aus Satz 1.9 erfüllt.

2. Mit $T^* := \{p \in \mathbb{R}^n : \nabla F_i(x^*)^T p = 0 \text{ für alle } i \in I(x^*) \text{ mit } \lambda_i^* > 0\}$ ist

$$p^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}[F_i(x^*)] \nabla^2 F_i(x^*) \right\} p > 0 \quad \text{für alle } p \in T^* \setminus \{0\}.$$

Dann ist x^* eine isolierte, lokale Lösung von (P), d. h. es existiert eine Umgebung U^* von x^* mit $\|F(x^*)\|_\infty < \|F(x)\|_\infty$ für alle $x \in U^* \setminus \{x^*\}$.

Beweis: Angenommen, die Behauptung sei falsch. Dann gibt es eine gegen x^* konvergierende Folge $\{x_k\}$, $x_k \neq x^*$ für alle k , mit $f(x_k) \leq f(x^*)$. Man stelle x_k dar in der Form $x_k = x^* + t_k p_k$ mit $t_k > 0$ und $\|p_k\| = 1$ (hierbei ist $\|\cdot\|$ irgendeine Vektornorm). Wegen $\lim_{k \rightarrow \infty} x_k = x^*$ ist $\lim_{k \rightarrow \infty} t_k = 0$. Aus $\{p_k\}$ kann eine konvergente Teilfolge ausgewählt werden. Daher nehmen wir o. B. d. A. an, es sei $x_k = x^* + t_k p_k$ mit einer Nullfolge $\{t_k\} \subset \mathbb{R}_+$ und $\lim_{k \rightarrow \infty} t_k = p \neq 0$.

Wegen $x_k = x^* + t_k p_k + r_k$ mit $r_k := t_k(p_k - p)$ und $\lim_{k \rightarrow \infty} r_k/t_k = 0$ kann leicht analog zum Beweis von Satz 1.8 gezeigt werden, daß

$$\lim_{k \rightarrow \infty} \frac{f(x^* + t_k p_k) - f(x^*)}{t_k} = f'(x^*; p).$$

Wegen $f(x^* + t_k p_k) \leq f(x^*)$ und $t_k > 0$ ist $f'(x^*; p) \leq 0$. Satz 1.8 liefert uns

$$f'(x^*; p) = \max_{i \in I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p \leq 0.$$

Aus der ersten Voraussetzung erhalten wir

$$\sum_{i \in I(x^*)} \lambda_i^* \underbrace{\text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p}_{\leq 0} = 0$$

und hieraus $p \in T^*$. Für $i \in I(x^*)$ ist

$$|F_i(x_k)| \leq \|F(x_k)\|_\infty \leq \|F(x^*)\|_\infty = |F_i(x^*)|,$$

ferner ist $\text{sign}[F_i(x_k)] = \text{sign}[F_i(x^*)]$ für alle hinreichend großen k . Für alle $i \in I(x^*)$ und alle hinreichend großen k ist daher

$$\begin{aligned} 0 &\geq |F_i(x_k)| - |F_i(x^*)| \\ &= \text{sign}[F_i(x^*)] [F_i(x^* + t_k p_k) - F_i(x^*)] \\ &= t_k \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p_k + \frac{1}{2} t_k^2 p_k^T \{ \text{sign}[F_i(x^*)] \nabla^2 F_i(z_{ik}) \} p_k \end{aligned}$$

mit $z_{ik} = x^* + \vartheta_{ik} t_k p_k$ und $\vartheta_{ik} \in (0, 1)$. Eine Multiplikation dieser Ungleichung mit $\lambda_i^* \geq 0$, $i \in I(x^*)$, und anschließendes Aufsummieren liefert unter erneuter Benutzung der ersten Voraussetzung, daß

$$0 \geq t_k \underbrace{\left\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla F_i(x^*) \right\}^T p_k}_{=0} + \frac{1}{2} t_k^2 p_k^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla^2 F_i(z_{ik}) \right\} p_k$$

bzw.

$$p_k^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla^2 F_i(z_{ik}) \right\} p_k \leq 0$$

für alle hinreichend großen k . Mit $k \rightarrow \infty$ folgt wegen $p_k \rightarrow p$ und $z_{ik} \rightarrow x^*$, daß

$$p^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}[F_i(x^*)] \nabla^2 F_i(x^*) \right\} p \leq 0,$$

was wegen $p \in T^* \setminus \{0\}$ ein Widerspruch zur zweiten Voraussetzung ist. \square

Bemerkung: Von A. BEN-TAL, J. ZOWE (1982) werden für eine allgemeinere Klasse von nichtglatten, unrestringierten Optimierungsaufgaben notwendige und hinreichende Optimalitätsbedingungen bewiesen. Auf die Min-Max-Aufgabe und die diskrete L_1 -Approximation wird in den Aufgaben 6 und 7 eingegangen. \square

7.1.4 Glatte konvexe Funktionen

In diesem kurzen Unterabschnitt sollen glatte, d. h. einmal oder zweimal stetig differenzierbare Funktionen betrachtet und ihre Konvexität durch geeignete Bedingungen charakterisiert werden.

Definition 1.13 Eine Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ heißt *gleichmäßig konvex* auf der konvexen Menge $D \subset \mathbb{R}^n$, falls eine Konstante $c > 0$ existiert mit

$$(*) \quad (1-t)f(x) + tf(y) - f((1-t)x + ty) \geq \frac{c}{2} t(1-t) \|x - y\|_2^2$$

für alle $x, y \in D, t \in [0, 1]$.

(Dagegen heißt f bekanntlich *konvex* auf D , wenn $(*)$ mit $c = 0$ gilt.)

Beispiel: Quadratische Funktionen sind die einfachsten Beispiele für glatte, nichtlineare und konvexe Funktionen. Genauer seien $c \in \mathbb{R}^n$ und eine symmetrische, positiv semidefinite Matrix $Q \in \mathbb{R}^{n \times n}$ gegeben und hiermit $f: \mathbb{R}^n \rightarrow \mathbb{R}$ durch

$$f(x) := c^T x + \frac{1}{2} x^T Q x$$

definiert. Für $x, y \in \mathbb{R}^n$ und $t \in [0, 1]$ ist

$$(1-t)f(x) + tf(y) - f((1-t)x + ty) = \frac{t(1-t)}{2} (x - y)^T Q (x - y) \geq 0,$$

womit die Konvexität von f bewiesen ist. Ist Q sogar positiv definit, so ist f gleichmäßig konvex (als Konstante c kann der kleinste Eigenwert von Q gewählt werden). Als Gradienten von f in x erhält man $\nabla f(x) = c + Qx$. Damit ist ein $x^* \in \mathbb{R}^n$ genau dann ein stationärer Punkt von f bzw. (für positiv semidefinites Q) eine globale Lösung der zugehörigen unrestringierten Optimierungsaufgabe, wenn x^* dem linearen Gleichungssystem $c + Qx = 0$ genügt. Für positiv definites Q bzw. gleichmäßig konvexes f ist dieses lineare Gleichungssystem eindeutig lösbar. Offenbar ist $\nabla^2 f(x) = Q$. \square

In den beiden folgenden Sätzen wird die Konvexität und die gleichmäßige Konvexität einer glatten Funktion f durch ihre ersten bzw. zweiten Ableitungen charakterisiert.

Satz 1.14 Sei $D \subset \mathbb{R}^n$ konvex und $f: \mathbb{R}^n \rightarrow \mathbb{R}$ auf einer offenen Obermenge von D stetig differenzierbar. Dann gilt:

1. f ist genau dann auf D konvex, wenn

$$\nabla f(x)^T (y - x) \leq f(y) - f(x) \quad \text{für alle } x, y \in D.$$

2. f ist genau dann auf D gleichmäßig konvex (mit einer Konstanten $c > 0$), wenn

$$\frac{c}{2} \|y - x\|_2^2 + \nabla f(x)^T (y - x) \leq f(y) - f(x) \quad \text{für alle } x, y \in D.$$

Beweis: Für alle $x, y \in D$ und $t \in [0, 1]$ sei

$$(1-t)f(x) + tf(y) \geq f((1-t)x + ty) + \frac{c}{2}t(1-t)\|y - x\|_2^2$$

mit einer Konstanten $c \geq 0$. Es sei f also konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). Dann ist

$$f(y) - f(x) \geq \frac{f(x + t(y - x)) - f(x)}{t} + \frac{c}{2}(1-t)\|y - x\|_2^2 \quad \text{für alle } t \in (0, 1].$$

Mit $t \rightarrow 0+$ folgt

$$f(y) - f(x) \geq \nabla f(x)^T(y - x) + \frac{c}{2}\|y - x\|_2^2.$$

Damit ist eine Richtung (nämlich " \Rightarrow ") bewiesen. Für die andere Richtung " \Leftarrow " nehmen wir an, mit einer Konstanten $c \geq 0$ sei

$$\frac{c}{2}\|y - x\|_2^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \quad \text{für alle } x, y \in D.$$

Seien $x, y \in D$ und $t \in [0, 1]$ vorgegeben. Dann ist $z := (1-t)x + ty \in D$ wegen der Konvexität von D und daher nach Voraussetzung

$$\begin{aligned} f(x) - f(z) &\geq \nabla f(z)^T(x - z) + \frac{c}{2}\|x - z\|_2^2, \\ f(y) - f(z) &\geq \nabla f(z)^T(y - z) + \frac{c}{2}\|y - z\|_2^2. \end{aligned}$$

Eine Multiplikation dieser Ungleichungen mit $(1-t)$ bzw. t und anschließende Addition ergibt

$$\begin{aligned} (1-t)f(x) + tf(y) - f((1-t)x + ty) &\geq \frac{c}{2}[(1-t)\|x - z\|_2^2 + t\|y - z\|_2^2] \\ &= \frac{c}{2}t(1-t)\|x - y\|_2^2. \end{aligned}$$

Also ist f konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). □

Satz 1.15 Sei $D \subset \mathbb{R}^n$ konvex und $f: \mathbb{R}^n \rightarrow \mathbb{R}$ auf einer offenen Obermenge von D zweimal stetig differenzierbar. Dann gilt:

1. Ist $\nabla^2 f(x)$ positiv semidefinit für alle $x \in D$, so ist f auf D konvex.
2. Existiert eine Konstante $c > 0$ mit

$$c\|p\|_2^2 \leq p^T \nabla^2 f(x)p \quad \text{für alle } x \in D \text{ und alle } p \in \mathbb{R}^n,$$

so ist f auf D gleichmäßig konvex (mit der Konstanten c).

3. Ist D auch offen, so gelten in 1. und 2. auch die Umkehrungen.

Beweis: Die ersten beiden Aussagen werden zusammen bewiesen, indem wir annehmen, es existiere eine Konstante $c \geq 0$ mit $c\|p\|_2^2 \leq p^T \nabla^2 f(x)p$ für alle $x \in D$ und alle $p \in \mathbb{R}^n$.

Seien $x, y \in D$. Definiert man $\phi: [0, 1] \rightarrow \mathbb{R}$ durch $\phi(t) := f(x + t(y - x))$, so ist

$$\phi(1) - \phi(0) - \phi'(0) = \frac{1}{2}\phi''(t_0) \quad \text{mit } t_0 \in (0, 1)$$

bzw.

$$f(y) - f(x) - \nabla f(x)^T(y - x) = \frac{1}{2}(y - x)^T \nabla^2 f(\underbrace{x + t_0(y - x)}_{\in D})(y - x) \geq \frac{c}{2} \|y - x\|_2^2.$$

Aus Satz 1.14 folgt, daß f konvex ($c = 0$) bzw. gleichmäßig konvex (mit der Konstanten $c > 0$) ist.

Nun sei D auch offen und f auf D konvex ($c = 0$) bzw. gleichmäßig konvex (mit der Konstanten $c > 0$). Seien $x \in D$ und $p \in \mathbb{R}^n$ beliebig. Wegen der Offenheit von D ist $x + tp \in D$ für alle hinreichend kleinen $|t|$. Nach Satz 1.14 gilt für diese t :

$$\begin{aligned} f(x + tp) - f(x) &\geq \nabla f(x)^T(tp) + \frac{c}{2} t^2 \|p\|_2^2, \\ f(x) - f(x + tp) &\geq -\nabla f(x + tp)^T(tp) + \frac{c}{2} t^2 \|p\|_2^2, \end{aligned}$$

so daß nach Addition $[\nabla f(x + tp) - \nabla f(x)]^T(tp) \geq ct^2\|p\|_2^2$. Wegen

$$\begin{aligned} p^T \nabla^2 f(x)p &= \lim_{t \rightarrow 0} \frac{[\nabla f(x + tp) - \nabla f(x)]^T p}{t} \\ &= \lim_{t \rightarrow 0} \frac{[\nabla f(x + tp) - \nabla f(x)]^T(tp)}{t^2} \\ &\geq c\|p\|_2^2 \end{aligned}$$

ist die Behauptung bewiesen. \square

Bemerkung: Konvexe Funktionen spielen aus mehreren Gründen eine wichtige Rolle in der Optimierung, insbesondere auch bei unrestringierten Optimierungsaufgaben. Nur für konvexe Zielfunktionen wird man *globale* Konvergenzaussagen (bei diesen wird *nicht* vorausgesetzt, daß der Startwert x_0 hinreichend nahe bei einer Lösung liegt) für später zu untersuchende Verfahren erwarten können. Ist andererseits in einer lokalen Lösung x^* die hinreichende Bedingung zweiter Ordnung erfüllt, ist also f in einer Umgebung von x^* zweimal stetig differenzierbar und ist $\nabla^2 f(x^*)$ positiv definit, so existiert eine offene, konvexe Umgebung D von x^* und eine Konstante $c > 0$ derart, daß

$$(*) \quad c\|p\|_2^2 \leq p^T \nabla^2 f(x)p \quad \text{für alle } x \in D \text{ und alle } p \in \mathbb{R}^n.$$

Denn ist $\lambda_{\min}^* > 0$ der kleinste Eigenwert von $\nabla^2 f(x^*)$, so ist $\lambda_{\min}^* \|p\|_2^2 \leq p^T \nabla^2 f(x^*)p$ für alle $p \in \mathbb{R}^n$. Bestimmt man daher eine offene, konvexe Umgebung D von x^* (etwa eine offene Kugel um x^* mit hinreichend kleinem Radius) so, daß

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\|_2 \leq \frac{\lambda_{\min}^*}{2} \quad \text{für alle } x \in D,$$

so ist für beliebiges $x \in D$ und alle $p \in \mathbb{R}^n$:

$$\begin{aligned} p^T \nabla^2 f(x) p &= p^T \nabla^2 f(x^*) p + p^T [\nabla^2 f(x) - \nabla^2 f(x^*)] p \\ &\geq \lambda_{\min}^* \|p\|_2^2 - \|\nabla^2 f(x) - \nabla^2 f(x^*)\|_2 \|p\|_2^2 \\ &\geq \frac{\lambda_{\min}^*}{2} \|p\|_2^2 \end{aligned}$$

und damit (*) erfüllt. Man beachte, daß (*) gleichwertig mit der Aussage ist, daß alle Eigenwerte von $\nabla^2 f(x)$ für jedes $x \in D$ größer oder gleich c sind. Wir haben uns damit überlegt: Ist f auf einer Umgebung von x^* zweimal stetig differenzierbar und ist $\nabla^2 f(x^*)$ positiv definit, so ist f auf einer hinreichend kleinen, offenen, konvexen Umgebung von x^* gleichmäßig konvex. Die hinreichende Optimalitätsbedingung zweiter Ordnung bzw. die Voraussetzung (*) spielt besonders bei Aussagen über die Konvergenzgeschwindigkeit eines konvergenten Verfahrens eine wichtige Rolle, da (*) ja gerade besagt, daß wenigstens lokal sich die Zielfunktion außerordentlich angenehm benimmt. \square

Beispiel: Die Rosenbrock-Funktion

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

besitzt $x^* = (1, 1)$ als einziges lokales, und damit globales Minimum. Als Hessesche von f in x berechnet man

$$\nabla^2 f(x) = \begin{pmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix},$$

so daß

$$\nabla^2 f(x^*) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \quad \text{mit den Eigenwerten } \begin{array}{rcl} \lambda_1^* & \approx & 1001.6, \\ \lambda_2^* & \approx & 0.4. \end{array}$$

Also ist $\nabla^2 f(x^*)$ positiv definit, ein Eigenwert aber klein, der andere groß. Dies ist ein Indiz dafür, daß x^* niedrigster Punkt in einem „langgestreckten Tal“ ist. Die Rosenbrock-Funktion ist nicht auf dem gesamten \mathbb{R}^2 konvex, da $\nabla^2 f(x)$ offenbar auch negative Eigenwerte besitzen kann. \square

Mit Hilfe von Satz 1.15 kann die Konvexität gegebener Funktionen nachgewiesen werden, ähnlich wie im eindimensionalen Fall, bei dem man zum Nachweis der Konvexität die zweite Ableitung berechnet und zeigt, daß diese nichtnegativ ist. Beispiele hierfür finden sich in den Aufgaben 9 und 10.

Aufgaben

1. Die Gateaux-Variation der durch $f(x) := \max_{j=1,\dots,n} x_j$ bzw. $f(x) := \|x\|_1 = \sum_{j=1}^n |x_j|$ definierten konvexen Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ist durch

$$f'(x; p) = \max_{j \in J(x)} p_j \quad \text{mit } J(x) := \{j \in \{1, \dots, n\} : x_j = f(x)\}$$

bzw.

$$f'(x; p) = \sum_{j \in J(x)} |p_j| + \sum_{j \notin J(x)} \operatorname{sign}(x_j) p_j \quad \text{mit } J(x) := \{j \in \{1, \dots, n\} : x_j = 0\}$$

gegeben.

2. Sei $f = g \circ F$ mit einer konvexen, lipschitzstetigen Funktion $g: \mathbb{R}^m \rightarrow \mathbb{R}$ und einer in $x^* \in \mathbb{R}^n$ stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$. Dann gilt:

- (a) f besitzt in x^* eine durch $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ gegebene Gateaux-Variation.
- (b) Ist $\{t_k\} \subset \mathbb{R}_+$ eine Nullfolge und $\{r_k\} \subset \mathbb{R}^n$ eine Folge mit $\lim_{k \rightarrow \infty} r_k/t_k = 0$, so ist

$$\lim_{k \rightarrow \infty} \frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k} = f'(x^*; p) \quad \text{für jedes } p \in \mathbb{R}^n.$$

3. Ist $f(x) := \max_{i=1, \dots, m} F_i(x)$ mit in x^* stetig differenzierbaren $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, so besitzt f in x^* eine Gateaux-Variation, die durch

$$f'(x^*; p) = \max_{i \in I(x^*)} \nabla F_i(x^*)^T p \quad \text{mit } I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = f(x^*)\}$$

gegeben ist. Entsprechend besitzt $f(x) := \sum_{i=1}^m |F_i(x)|$ in x^* die Gateaux-Variation

$$f'(x^*; p) = \sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*)^T p$$

mit $I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = 0\}$ (siehe auch R. A. EL-ATTAR ET AL. (1979, Lemma 2.2)).

4. Ist $f(x) := \max_{i=1, \dots, m} F_i(x)$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, so ist

$$(*) \quad f'(x^*; p) = \max_{i \in I(x^*)} \nabla F_i(x^*)^T p \geq 0 \quad \text{für alle } p \in \mathbb{R}^n,$$

wobei $I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = f(x^*)\}$, äquivalent zu der Existenz von reellen Zahlen $\lambda_i^*, i \in I(x^*)$, mit

$$(**) \quad \lambda_i^* \geq 0 \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* = 1, \quad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

5. Ist $f(x) := \sum_{i=1}^m |F_i(x)|$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, so ist

$$(*) \quad \left\{ \begin{array}{l} f'(x^*; p) = \sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*)^T p \geq 0 \\ \text{für alle } p \in \mathbb{R}^n, \end{array} \right.$$

wobei $I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = 0\}$, äquivalent zu der Existenz von reellen Zahlen $\lambda_i^*, i \in I(x^*)$, mit

$$(**) \quad \lambda_i^* \in [-1, 1] \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) = 0.$$

Hinweis: Daß $(**)$ wieder $(*)$ impliziert, sieht man durch genaueres Hinsehen. Die Umkehrung ist ein wenig komplizierter als in Satz 1.9, daher geben wir den Beweis an (siehe auch C. CHARALAMBOUS (1979, Theorem 2) und R. A. EL-ATTAR ET AL. (1979, Lemma 2.4)). Sei q die Anzahl der Elemente von $I(x^*)$, o. B. d. A. sei $q \geq 1$ (andernfalls ist die Implikation $(*) \Rightarrow (**)$ trivial). Mit $B \in \mathbb{R}^{n \times q}$ bezeichne man die Matrix, deren Spalten durch $\nabla F_i(x^*)$, $i \in I(x^*)$, gegeben sind. Schließlich sei wieder $e := (1, \dots, 1)^T \in \mathbb{R}^q$ und zur Abkürzung $c := \sum_{i \notin I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)$. Die Annahme, $(**)$ würde nicht gelten, bedeutet dann genau, daß das System $B\lambda = -c$, $-e \leq \lambda \leq e$ keine Lösung $\lambda \in \mathbb{R}^q$ besitzt. Um wie beim Beweis von Satz 1.9 das Farkas-Lemma anwenden zu können, muß dieses System sozusagen auf Simplex-Normalform gebracht werden. Tut man dies (Einführung von nichtnegativen Schlupfvariablen y und z sowie Darstellung von λ als Differenz nichtnegativer μ und ν), so erhält man, daß das System

$$\begin{pmatrix} B & -B & 0 & 0 \\ I & -I & I & 0 \\ -I & I & 0 & I \end{pmatrix} \begin{pmatrix} \mu \\ \nu \\ y \\ z \end{pmatrix} = \begin{pmatrix} -c \\ e \\ e \end{pmatrix}, \quad \begin{pmatrix} \mu \\ \nu \\ y \\ z \end{pmatrix} \geq 0$$

nicht lösbar ist. Nun ist das Farkas-Lemma anwendbar, es zeigt die Lösbarkeit von

$$\begin{pmatrix} B^T & I & -I \\ -B^T & -I & I \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} p \\ u \\ v \end{pmatrix} \leq 0, \quad \begin{pmatrix} -c \\ e \\ e \end{pmatrix}^T \begin{pmatrix} p \\ u \\ v \end{pmatrix} > 0.$$

Daher existiert ein $p \in \mathbb{R}^n$ sowie $u_i, v_i \leq 0$, $i \in I(x^*)$, mit

$$\nabla F_i(x^*)^T p + u_i - v_i = 0 \quad (i \in I(x^*)), \quad \sum_{i \notin I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p < \sum_{i \in I(x^*)} (u_i + v_i).$$

Dann ist aber

$$\begin{aligned} f'(x^*; p) &= \sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T p \\ &< \sum_{i \in I(x^*)} [\underbrace{|u_i - v_i|}_{\leq 0} + (u_i + v_i)] \\ &\leq 0, \end{aligned}$$

ein Widerspruch zu $(*)$.

6. Gegeben sei die unrestringierte Min-Max-Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \max_{i=1, \dots, m} F_i(x), \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

$$I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = f(x^*)\}.$$

Es wird vorausgesetzt, daß reelle Zahlen λ_i^* , $i \in I(x^*)$, existieren mit:

(a) Es ist

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* = 1, \quad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

(b) Mit $T^* := \{p \in \mathbb{R}^n : \nabla F_i(x^*)^T p = 0 \text{ für alle } i \in I(x^*) \text{ mit } \lambda_i^* > 0\}$ ist

$$p^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) \right\} p > 0 \quad \text{für alle } p \in T^* \setminus \{0\}.$$

Dann ist x^* eine isolierte, lokale Lösung von (P).

7. Gegeben sei die diskrete L_1 -Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \sum_{i=1}^m |F_i(x)|, \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

$$I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = 0\}.$$

Es wird vorausgesetzt, daß reelle Zahlen λ_i^* , $i \in I(x^*)$, existieren mit:

(a) Es ist $\lambda_i^* \in [-1, 1]$, $i \in I(x^*)$, und

$$\sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \text{sign}[F_i(x^*)] \nabla F_i(x^*) = 0.$$

(b) Mit

$$T^* := \left\{ p \in \mathbb{R}^n : \nabla F_i(x^*)^T p \begin{cases} = 0 & \text{für } i \in I(x^*) \text{ mit } |\lambda_i^*| < 1, \\ \geq 0 & \text{für } i \in I(x^*) \text{ mit } \lambda_i^* = 1, \\ \leq 0 & \text{für } i \in I(x^*) \text{ mit } \lambda_i^* = -1 \end{cases} \right\}$$

ist

$$p^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) + \sum_{i \notin I(x^*)} \text{sign}[F_i(x^*)] \nabla^2 F_i(x^*) \right\} p > 0 \quad \text{für alle } p \in T^* \setminus \{0\}.$$

Dann ist x^* eine isolierte, lokale Lösung von (P).

Hinweis: Angenommen, die Aussage sei falsch. Wie im Beweis von Satz 1.12 zeige man die Existenz einer Richtung $p \neq 0$ mit $f'(x^*; p) \leq 0$. Hieraus schließe man auf $\lambda_i^* \nabla F_i(x^*)^T p = |\nabla F_i(x^*)^T p|$ für alle $i \in I(x^*)$, daraus auf $p \in T^*$ und folgere analog zum Beweis von Satz 1.12 auf einen Widerspruch zur zweiten Voraussetzung.

8. Sind beim diskreten Tschebyscheffschen Approximationsaufgabe oder der diskreten L_1 -Approximation die Funktionen $F_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, affin linear, so ist eine stationäre Lösung sogar eine globale Lösung der entsprechenden Aufgabe.

9. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit. Die Funktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ sei durch $f(x) := \frac{1}{2}(x^T A x)^2$ definiert. Man berechne die Hessesche $\nabla^2 f(x)$ von f und beweise, daß f auf dem \mathbb{R}^n konvex ist.
10. Für $k = 1, \dots, m$ seien $c_k > 0$ und $a_k \in \mathbb{R}^n$. Hiermit definiere man auf dem \mathbb{R}^n die reellwertigen Funktionen f und g durch

$$f(x) := \sum_{k=1}^m c_k \exp(a_k^T x), \quad g(x) := \ln\left(\sum_{k=1}^m c_k \exp(a_k^T x)\right).$$

Funktionen dieser Form treten bei sogenannten *geometrischen Optimierungsaufgaben* auf. Man zeige, daß f und g auf dem \mathbb{R}^n konvex sind.

Hinweis: Daß f konvex ist, ist leicht zu sehen. Zum Nachweis der Konvexität von g zeige man, daß $\nabla^2 g(x)$ durch

$$\begin{aligned} \nabla^2 g(x) &= \frac{1}{f(x)^2} \left\{ \left(\sum_{k=1}^m c_k \exp(a_k^T x) \right) \left(\sum_{k=1}^m c_k \exp(a_k^T x) a_k a_k^T \right) \right. \\ &\quad \left. - \left(\sum_{k=1}^m c_k \exp(a_k^T x) a_k \right) \left(\sum_{k=1}^m c_k \exp(a_k^T x) a_k \right)^T \right\} \end{aligned}$$

gegeben ist und beweise mit Hilfe der Cauchy-Schwarzschen Ungleichung, daß $\nabla^2 g(x)$ für jedes $x \in \mathbb{R}^n$ positiv semidefinit ist.

7.2 Ein Modellalgorithmus

Wir betrachten die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

und nehmen an, die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ besitze in jedem $x \in \mathbb{R}^n$ eine Gateaux-Variation $f'(x; \cdot)$, sei also z. B. aus $C^1(\mathbb{R}^n)$. Die meisten der später zu untersuchenden konkreten Verfahren (lediglich die sogenannten Trust-Region-Verfahren bilden eine Ausnahme) lassen sich dem folgenden *Modellalgorithmus* unterordnen:

- Sei $x_0 \in \mathbb{R}^n$ gegeben.
- Für $k = 0, 1, \dots$:

Test auf Abbruch: Falls $f'(x_k; p) \geq 0$ für alle $p \in \mathbb{R}^n$, dann: STOP.

Wahl einer (Abstiegs-) Richtung: Bestimme $p_k \in \mathbb{R}^n$ mit $f'(x_k; p_k) < 0$.

Wahl einer Schrittweite: Bestimme $t_k > 0$ mit $f(x_k + t_k p_k) < f(x_k)$.

Bestimme neue Näherung: Setze $x_{k+1} := x_k + t_k p_k$.

Uns kommt es darauf an, deutlich zu machen, daß dieser Modellalgorithmus sich aus einer *Richtungs-* und einer *Schrittweitenstrategie* zusammensetzt. Durch eine Spezifikation dieser Strategien wird aus dem Modellalgorithmus ein auf die gegebene Aufgabe anwendbares, konkretes Verfahren.

7.2.1 Schrittweitenstrategien bei glatter Zielfunktion

Wir betrachten die unrestringierte Optimierungsaufgabe (P) und nehmen an, daß die folgenden Voraussetzungen für die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ erfüllt sind:

- (V) (a) Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (gewöhnlich Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt.
- (b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.
- (c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2 \quad \text{für alle } x, y \in L_0.$$

Auf die Wahl von Richtungsstrategien im Modellalgorithmus werden wir in den folgenden Abschnitten ausführlich eingehen. In diesem Unterabschnitt stellen wir uns auf den Standpunkt, es sei eine aktuelle Näherung $x = x_k \in L_0$ gegeben, es sei x keine stationäre Lösung von (P), also $\nabla f(x) \neq 0$, und $p = p_k$ eine *Abstiegsrichtung* für f in x bzw. $\nabla f(x)^T p < 0$. Z. B. ist $p := -\nabla f(x)$ (diese Richtungswahl führt auf das schon 1827 von Cauchy angegebene *Gradientenverfahren*, manchmal auch *Verfahren des steilsten Abstiegs* genannt) oder allgemeiner $p := -H\nabla f(x)$ mit einer symmetrischen, positiv definiten Matrix $H \in \mathbb{R}^{n \times n}$ eine Abstiegsrichtung.

Ziel dieses Unterabschnittes ist es, Strategien zur Berechnung einer Schrittweite $t > 0$ anzugeben, für welche die Verminderung $f(x) - f(x + tp)$ der Zielfunktion einerseits positiv und andererseits so groß ist, daß (einfache) Konvergenzaussagen für das entstehende, von einer speziellen Richtungsstrategie weitgehend unabhängige Verfahren gemacht werden können.

Das folgende Lemma wird sich als nützlich erweisen, wenn $f(x) - f(x + tp)$ nach unten abgeschätzt werden soll.

Lemma 2.1 Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x , d. h. es sei $\nabla f(x)^T p < 0$. Ist dann $\hat{t} = \hat{t}(x, p)$ die erste positive Nullstelle von $\psi(t) := f(x) - f(x + tp)$, so gilt:

$$f(x + tp) \leq f(x) + t \nabla f(x)^T p + t^2 \frac{\gamma}{2} \|p\|_2^2 \quad \text{für alle } t \in [0, \hat{t}], \quad -\frac{2 \nabla f(x)^T p}{\gamma \|p\|_2^2} \leq \hat{t}.$$

Beweis: Wegen Voraussetzung (V) existiert \hat{t} und es ist $x + tp \in L_0$ für alle $t \in [0, \hat{t}]$. Für $t \in [0, \hat{t}]$ erhält man

$$\begin{aligned} f(x + tp) &= f(x) + t \nabla f(x)^T p + \int_0^t [\nabla f(\underbrace{x + sp}_{\in L_0}) - \nabla f(x)]^T p \, ds \\ &\leq f(x) + t \nabla f(x)^T p + \int_0^t \gamma \|p\|_2^2 s \, ds \\ &\quad (\text{Voraussetzung (V) (c) und Cauchy-Schwarzsche Ungleichung}) \\ &= f(x) + t \nabla f(x)^T p + t^2 \frac{\gamma}{2} \|p\|_2^2, \end{aligned}$$

und damit die erste Behauptung. Die zweite folgt, indem man in der gerade eben bewiesenen Ungleichung $t = \hat{t}$ setzt. \square

Eine naheliegende Schrittweitenstrategie besteht darin, als Schrittweite $t^* > 0$ eine globale oder die erste stationäre Lösung der eindimensionalen Minimierungsaufgabe

$$\text{Minimiere } f(x + tp), \quad t \in [0, \infty)$$

zu wählen. In diesem Falle wird $t^* > 0$ also so bestimmt, daß

$$f(x + t^* p) = \min_{t \geq 0} f(x + tp)$$

bzw.

$$\nabla f(x + t^* p)^T p = 0 \quad \text{und} \quad \nabla f(x + tp)^T p < 0 \quad \text{für alle } t \in [0, t^*).$$

Unter der Voraussetzung (V) ist die Existenz dieser Schrittweiten gesichert. Man spricht von einer *exakten Schrittweite*, da eine eindimensionale Minimierungsaufgabe bzw. eine eindimensionale Nullstellenaufgabe zur Bestimmung der Schrittweite exakt zu lösen ist. Es ist klar, daß nur in Ausnahmefällen (z. B. bei quadratischen Zielfunktionen) die exakte Schrittweite (in endlich vielen Schritten) berechnet werden kann, i. allg. muß man sich mit einer Näherung begnügen. Trotzdem wollen wir die durch die exakte Schrittweite erreichbare Verminderung der Zielfunktion im folgenden Satz abschätzen.

Satz 2.2 Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x . Zur Abkürzung sei $\phi(t) := f(x + tp)$. Ist t^* die erste positive Nullstelle von $\phi'(\cdot)$ auf $[0, \infty)$, so ist

$$(*) \quad -\frac{\nabla f(x)^T p}{\gamma \|p\|_2^2} \leq t^*, \quad f(x) - f(x + t^* p) \geq \frac{1}{2\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

Beweis: Wegen der Voraussetzung (V) ist die Existenz der exakten Schrittweite t^* gesichert. ϕ ist monoton fallend auf $[0, t^*]$, ferner ist $t^* \leq \hat{t}$, wobei \hat{t} wie in Lemma 2.1 die erste positive Nullstelle von $\phi(0) - \phi(\cdot)$ ist. Der Beweis des Satzes erfolgt in zwei Schritten. Zunächst wird t^* nach unten abgeschätzt und dann mit Hilfe der Monotonie von ϕ und Lemma 2.1 die Behauptung bewiesen.

Wegen der Lipschitzstetigkeit von $\nabla f(\cdot)$ auf der Niveaumenge L_0 ist

$$0 = \nabla f(x + t^* p)^T p = \nabla f(x)^T p + [\nabla f(\underbrace{x + t^* p}_{\in L_0}) - \nabla f(x)]^T p \leq \nabla f(x)^T p + \gamma t^* \|p\|_2^2,$$

so daß

$$\tilde{t} := -\frac{\nabla f(x)^T p}{\gamma \|p\|_2^2} \leq t^*.$$

Hieraus erhält man

$$f(x + t^* p) \leq f(x + \tilde{t} p) \leq f(x) + \tilde{t} \nabla f(x)^T p + \tilde{t}^2 \frac{\gamma}{2} \|p\|_2^2 = f(x) - \frac{1}{2\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2,$$

die Behauptung ist bewiesen. \square

Da eine exakte Schrittweite i. allg. nicht in endlich vielen Schritten realisiert werden kann, sind zunehmend *inexakte* Schrittweiten in Theorie und Praxis untersucht und angewandt worden. Bei der sogenannten *Powell-Schrittweite*³ (siehe M. J. D. POWELL (1976), der diese inexakte Schrittweite bei der Untersuchung der globalen Konvergenz von Quasi-Newton-Verfahren in den Vordergrund stellte) wird bei vorgegebenen $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ ein $t > 0$ so bestimmt, daß

$$(a) \quad f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p$$

und

$$(b) \quad \nabla f(x + tp)^T p \geq \beta \nabla f(x)^T p$$

erfüllt sind. (a) ist für alle hinreichend kleinen $t > 0$ erfüllt und besagt, daß die erwünschte Verminderung $f(x) - f(x + tp)$ der Zielfunktion nicht kleiner ist als ein kleines Vielfaches von $-\alpha t \nabla f(x)^T p$. Die Forderung (b) soll sichern, daß nicht zu kleine Schrittweiten gewählt werden. Die Situation machen wir uns in Abbildung 7.1 anschaulich klar. Man kann sich überlegen, daß man in endlich vielen Schritten eine

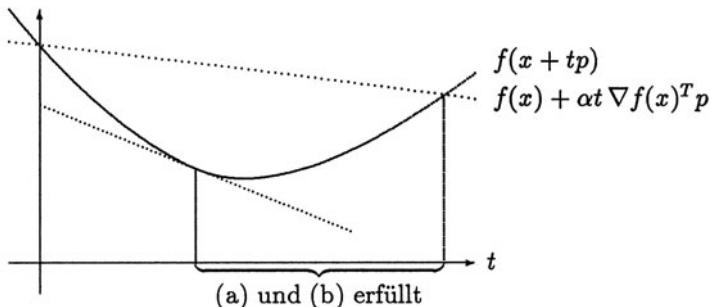


Abbildung 7.1: Veranschaulichung der Powell-Schrittweite

Powell-Schrittweite berechnen kann. Einen Algorithmus hierfür findet man bei J. E. DENNIS, R. B. SCHNABEL (1983, S. 328), siehe auch H. SCHWETLICK (1979, S. 187). Im folgenden Satz wollen wir zeigen, daß stets eine Schrittweite $t > 0$ existiert, die den Forderungen (a) und (b) genügt, ferner soll die durch eine Powell-Schrittweite erreichbare Verminderung der Zielfunktion nach unten abgeschätzt werden.

Satz 2.3 Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x . Seien $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ gegeben und

$$T_P(x, p) := \{t > 0 : f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p, \quad \nabla f(x + tp)^T p \geq \beta \nabla f(x)^T p\}$$

die Menge der Powell-Schrittweiten in x in Richtung p . Dann gilt:

³Gelegentlich wird diese Schrittweite auch *Wolfe-Schrittweite* genannt, siehe P. WOLFE (1969).

1. Es ist $T_P(x, p) \neq \emptyset$.
2. Es existiert eine Konstante $\theta > 0$, die nur von α, β und γ (der Lipschitzkonstanten von $\nabla f(\cdot)$ auf L_0) abhängt, nicht aber von x oder p , mit

$$(*) \quad f(x) - f(x + tp) \geq \theta \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2 \quad \text{für alle } t \in T_P(x, p).$$

Beweis: Zur Abkürzung setze man $\Phi(t) := f(x) - f(x + tp) + \alpha t \nabla f(x)^T p$. Ist t^* die erste positive Nullstelle von $\nabla f(x + tp)^T p$, so ist

$$\Phi'(0) = -(1 - \alpha) \nabla f(x)^T p > 0, \quad \Phi'(t^*) = \alpha \nabla f(x)^T p < 0.$$

Wegen $\Phi(0) = 0$ existiert daher ein $t \in (0, t^*)$ mit $\Phi(t) > 0$ und $\Phi'(t) = 0$. Offenbar ist $t \in T_P(x, p)$.

Mit \hat{t} werde wieder die erste positive Nullstelle von $\psi(t) := f(x) - f(x + tp)$ bezeichnet. Bei gegebenem $t \in T_P(x, p)$ machen wir eine Fallunterscheidung. Ist $t \leq \hat{t}$, so folgt aus

$$-(1 - \beta) \nabla f(x)^T p \leq [\underbrace{\nabla f(x + tp) - \nabla f(x)}_{\in L_0}]^T p \leq \gamma t \|p\|_2^2,$$

daß

$$f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p \leq f(x) - \frac{\alpha(1 - \beta)}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

Ist $t > \hat{t}$, so erhält man mit der unteren Schranke für \hat{t} aus Lemma 2.1, daß

$$f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p \leq f(x) - \frac{2\alpha}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

Insgesamt ist die Behauptung mit $\theta := \alpha(1 - \beta)/\gamma$ bewiesen. \square

Sehr oft spielt die Schrittweite $t = 1$ eine besondere Rolle. Dies ist insbesondere dann der Fall, wenn $p \approx -\nabla^2 f(x)^{-1} \nabla f(x)$, die Abstiegsrichtung also nahe bei der sogenannten *Newton-Richtung* liegt. Denn dann wird man hoffen, in der Nähe einer lokalen Lösung x^* von einem gedämpften Verfahren (die neue Näherung ist $x_+ = x + tp$ mit einer geeigneten Schrittweite $t > 0$) zu einem ungedämpften Verfahren (hier ist $x_+ = x + p$, die Schrittweite also $t = 1$) übergehen zu können. Dies ist die Motivation für die *Armijo-Schrittweite*. Grob gesagt testet man hier zunächst, ob bei einem vorgegebenem $\alpha \in (0, \frac{1}{2})$ die Ungleichung $f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p$ für $t := 1$ erfüllt ist. Ist dies der Fall, so wird $t = 1$ als Schrittweite akzeptiert. Andernfalls wird t „kontrolliert verkleinert“ und dieselbe Ungleichung erneut getestet. Sobald diese erfüllt ist (wir wissen, daß sie für alle hinreichend kleinen $t > 0$ gilt), wird die entsprechende Schrittweite akzeptiert.

Genauer sieht die Berechnung der Armijo-Schrittweite folgendermaßen aus:

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ gegeben. Setze $\rho_0 := 1$.

- Für $j = 0, 1, \dots$:

Falls $f(x + \rho_j p) \leq f(x) + \alpha \rho_j \nabla f(x)^T p$, dann: $t := \rho_j$, STOP.

Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Ist z. B. $l = u =: \rho$, so ist die Armijo-Schrittweite durch $t = \rho^j$ gegeben, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x + \rho^j p) \leq f(x) + \alpha \rho^j \nabla f(x)^T p$$

ist. Von S. P. HAN (1981) wird $\rho_0 := 1$ und

$$\rho_{j+1} := \max(0.1 \rho_j, \rho_j^*) \quad \text{mit} \quad \rho_j^* := -\frac{0.5 \rho_j^2 \nabla f(x)^T p}{f(x + \rho_j p) - f(x) - \rho_j \nabla f(x)^T p}$$

gesetzt. Ist $f(x + \rho_j p) > f(x) + \alpha \rho_j \nabla f(x)^T p$, die zu testende Ungleichung im j -ten Schritt also nicht erfüllt, so zeigt eine einfache Rechnung, daß

$$0.1 \rho_j \leq \rho_{j+1} \leq \underbrace{\frac{0.5}{1-\alpha}}_{<1} \rho_j.$$

Die von Han benutzte Schrittweite fällt also mit $l := 0.1$ und $u := 0.5/(1-\alpha)$ unter obiges Konzept. Man rechnet leicht nach, daß bei ρ_j^* gerade das Minimum des quadratischen Polynoms q liegt, das den Interpolationsbedingungen $q(0) = f(x)$, $q'(0) = \nabla f(x)^T p$ und $q(\rho_j) = f(x + \rho_j p)$ genügt. Eine etwas raffiniertere Version geben J. E. DENNIS, R. B. SCHNABEL (1983, S. 126 ff. und S. 325 ff.) an. Hierbei wird außer im ersten Schritt, bei dem $t = 1$ getestet wird, ein kubisches Polynom q (und dessen Minimum) bestimmt, das mit $f(x + tp)$ in den beiden letzten getesteten Schrittweiten übereinstimmt und für das $q(0) = f(x)$, $q'(0) = \nabla f(x)^T p$.

Der Vorteil der Armijo-Schrittweite liegt vor allem darin, daß sie sehr leicht zu implementieren ist. Im folgenden Satz wird die durch eine Armijo-Schrittweite erreichte Verminderung der Zielfunktion nach unten abgeschätzt.

Satz 2.4 Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x . Seien $\alpha \in (0, \frac{1}{2})$, $0 < l \leq u < 1$ und hiermit eine Armijo-Schrittweite $t = \rho_j$ gegeben. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u , nicht aber von x oder p abhängt, mit

$$(**) \quad f(x) - f(x + tp) \geq \theta \min \left[-\nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2 \right].$$

Beweis: Offenbar muß die zu testende Ungleichung nach endlich vielen Schritten erfüllt sein. Ist $j = 0$ bzw. $t = \rho_0 = 1$, so ist $f(x + tp) \leq f(x) + \alpha \nabla f(x)^T p$. Ist dagegen $j > 0$, so gelten mit $s := \rho_{j-1}$ zwei Ungleichungen:

$$f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p, \quad f(x + sp) > f(x) + \alpha s \nabla f(x)^T p.$$

Ferner ist $ls \leq t$. Mit \hat{t} wie in Lemma 2.1 machen wir eine Fallunterscheidung. Für $s \leq \hat{t}$ ist

$$f(x) + \alpha s \nabla f(x)^T p < f(x + sp) \leq f(x) + s \nabla f(x)^T p + s^2 \frac{\gamma}{2} \|p\|_2^2,$$

daher

$$\frac{2l(\alpha - 1)}{\gamma} \frac{\nabla f(x)^T p}{\|p\|_2^2} \leq ls \leq t$$

und folglich

$$f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p \leq f(x) - \frac{2\alpha(1-\alpha)l}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

Ist dagegen $s > \hat{t}$, so ist wiederum wegen Lemma 2.1

$$-\frac{2l \nabla f(x)^T p}{\gamma \|p\|_2^2} \leq l\hat{t} < ls \leq t$$

und daher

$$f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p \leq f(x) - \frac{2\alpha l}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

Der Satz ist damit bewiesen. \square

Bemerkungen: Die Voraussetzung $\alpha \in (0, \frac{1}{2})$ bei der Definition der Powell- und der Armijo-Schrittweite könnte durch $\alpha \in (0, 1)$ ersetzt werden und die Sätze 2.3 und 2.4 würden immer noch gelten. Bei dem Nachweis der superlinearen Konvergenz von Newton- und Quasi-Newton-Verfahren wird klar werden, weshalb $\alpha \in (0, \frac{1}{2})$ vorausgesetzt wird.

Schrittweiten t , für die eine Aussage wie bei der exakten Schrittweite oder der Powell-Schrittweite gemacht werden kann, für die also unter den Voraussetzungen (V) (a)–(c) eine von x und p unabhängige Konstante $\theta > 0$ mit

$$(*) \quad f(x) - f(x + tp) \geq \theta \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2$$

existiert, wurden von W. WARTH, J. WERNER (1977) *effizient* genannt. Entsprechend werden Schrittweiten t , wie z. B. die Armijo-Schrittweite, zu denen es unter den Voraussetzungen (V) (a)–(c) eine von x und p unabhängige Konstante $\theta > 0$ mit

$$(**) \quad f(x) - f(x + tp) \geq \theta \min \left[-\nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2 \right]$$

gibt, von P. KOSMOL (1989, S. 92) *semi-effizient* genannt. Wir vermeiden es, diese Begriffe zu benutzen, weil das Wort „effizient“ falsche Erwartungen wecken könnte. Trotzdem stellen sich die Beziehungen (*) und (**) als fundamental bei der Konvergenzanalyse heraus. Etwas vereinfacht gesagt: Hat man für eine Schrittweitenstrategie (*) bzw. (**) bewiesen, so kann man für die Konvergenzanalyse vergessen, wodurch die Schrittweitenstrategie spezifiziert ist, alleine die Richtungsstrategie spielt danach noch eine Rolle. \square

7.2.2 Konvergenz des Modellalgorithmus bei glatter Zielfunktion

Der folgende Satz gibt unter verhältnismäßig schwachen Voraussetzungen an die Zielfunktion f sowie an die benutzten Abstiegsrichtungen ein, wie man nicht anders erwarten kann, schwaches Konvergenzergebnis.

Satz 2.5 Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c). Als Schrittweite im Modellalgorithmus verwende man $t_k := t^*(x_k, p_k)$ (exakte Schrittweite), $t_k := t_P(x_k, p_k)$ (Powell-Schrittweite) oder $t_k := t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung sei $g_k := \nabla f(x_k)$. Ferner wird vorausgesetzt:

1. Es existiert eine Konstante $\sigma > 0$ mit $-\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \geq \sigma$, $k = 0, 1, \dots$
2. Es existiert eine Konstante $\tau > 0$ mit $\|p_k\|_2 \geq \tau \|g_k\|_2$, $k = 0, 1, \dots$

Dann gilt: Jeder Häufungspunkt der durch den Modellalgorithmus mit Abstiegsrichtungen p_k erzeugten Folge $\{x_k\}$ ist eine stationäre Lösung von (P). Besitzt (P) genau eine stationäre Lösung x^* in der Niveaumenge L_0 , so konvergiert die gesamte Folge $\{x_k\}$ gegen x^* .

Beweis: Wegen der Sätze 2.2, 2.3 und 2.4 existiert eine Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \min \left[-g_k^T p_k, \left(\frac{g_k^T p_k}{\|p_k\|_2} \right)^2 \right].$$

Wegen der Voraussetzungen 1. und 2. ist daher

$$f(x_k) - f(x_{k+1}) \geq \theta \sigma \min(\tau, \sigma) \|g_k\|_2^2 \quad \text{für } k = 0, 1, \dots$$

Da $\{f(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge ist, konvergiert damit $\{g_k\}$ gegen den Nullvektor.

Ist $x^* \in L_0$ ein Häufungspunkt von $\{x_k\}$, so ist x^* Limes einer konvergenten Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$. Da $\{g_{k(j)}\}$ einerseits gegen $\nabla f(x^*)$ und andererseits gegen 0 konvergiert, ist $\nabla f(x^*) = 0$, also x^* eine stationäre Lösung von (P).

Nun besitze (P) genau eine stationäre Lösung x^* in L_0 . Angenommen, $\{x_k\}$ konvergiert nicht gegen x^* . Dann existiert ein $\epsilon > 0$ und eine Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$ mit $\|x_{k(j)} - x^*\|_2 \geq \epsilon$, $j = 1, 2, \dots$. Da L_0 kompakt ist, kann aus $\{x_{k(j)}\}$ eine gegen ein $\hat{x} \in L_0$ konvergente Teilfolge ausgewählt werden. Als Häufungspunkt der Folge $\{x_k\}$ ist \hat{x} wegen des gerade bewiesenen ersten Teils eine stationäre Lösung von (P). Wegen $\|\hat{x} - x^*\|_2 \geq \epsilon$ ist $\hat{x} \neq x^*$. Dies ist ein Widerspruch dazu, daß x^* die einzige stationäre Lösung von (P) ist. \square

Bemerkungen: Man erkennt, daß in den Beweis des Konvergenzsatzes nicht die Definition der jeweiligen Schrittweiten eingeht, sondern lediglich die Folgerungen (*) bzw. (**) aus den Sätzen 2.2, 2.3 bzw. 2.4. Klar ist auch, daß man auf die zweite Voraussetzung in Satz 2.5, also die Existenz einer Konstanten $\tau > 0$ mit

$$\|p_k\|_2 \geq \tau \|g_k\|_2, \quad k = 0, 1, \dots$$

(hier und im folgenden wird die Abkürzung $g_k := \nabla f(x_k)$ benutzt), verzichten kann, wenn nur die exakte Schrittweite oder die Powell-Schrittweite (bzw. eine, für die $(*)$ gilt) verwendet wird.

Eine Folge von Abstiegsrichtungen $\{p_k\}$ wird *gradientenähnlich* genannt, wenn die erste Voraussetzung in Satz 2.5 erfüllt ist, wenn es also eine Konstante $\sigma > 0$ mit

$$-\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \geq \sigma, \quad k = 0, 1, \dots$$

gibt. Diese Voraussetzung besagt, daß der Winkel zwischen $-g_k$ und p_k gleichmäßig kleiner als der rechte Winkel sein muß. \square

Beispiel: Man betrachte den Modellalgorithmus mit einer Richtungsfolge $\{p_k\}$, wo bei $p_k = -H_k g_k$, $k = 0, 1, \dots$, mit einer symmetrischen und positiv definiten Matrix $H_k \in \mathbb{R}^{n \times n}$. Dann ist

$$-\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} = \frac{g_k^T H_k g_k}{\|g_k\|_2 \|H_k g_k\|_2} \geq \frac{1}{\sqrt{\|H_k\|_2 \|H_k^{-1}\|_2}} = \frac{1}{\sqrt{\text{cond}_2(H_k)}}.$$

Die erste Voraussetzung in Satz 2.5 ist daher erfüllt, wenn $\{\text{cond}_2(H_k)\}$ beschränkt ist. Hierbei haben wir benutzt: Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, bezeichnet ferner $A^{1/2}$ die symmetrische und positiv definite Quadratwurzel aus A , so ist für beliebiges $x \in \mathbb{R}^n \setminus \{0\}$:

$$\frac{x^T A x}{\|x\|_2 \|Ax\|_2} = \frac{\|A^{1/2} x\|_2^2}{\|A^{-1/2} A^{1/2} x\|_2 \|A^{1/2} A^{1/2} x\|_2} \geq \frac{1}{\text{cond}_2(A^{1/2})} = \frac{1}{\sqrt{\text{cond}_2(A)}}.$$

In Aufgabe 8 kann gezeigt werden, daß diese Abschätzung zu

$$\frac{x^T A x}{\|x\|_2 \|Ax\|_2} \geq \frac{2\sqrt{\text{cond}_2(A)}}{1 + \text{cond}_2(A)}$$

verbessert werden kann. Wegen

$$\|p_k\|_2 = \|H_k g_k\|_2 \geq \frac{1}{\|H_k^{-1}\|_2} \|g_k\|_2$$

ist die zweite Voraussetzung in Satz 2.5 erfüllt, wenn $\{\|H_k^{-1}\|_2\}$ beschränkt ist. Beide Voraussetzungen gelten, wenn $\{H_k\}$ eine Folge symmetrischer, gleichmäßig positiv definiter und beschränkter Matrizen ist, wenn es also Konstanten $0 < c \leq d$ mit

$$c \|z\|_2^2 \leq z^T H_k z \leq d \|z\|_2^2 \quad \text{für alle } z \in \mathbb{R}^n, k = 0, 1, \dots$$

gibt. Dies wiederum ist gleichbedeutend mit der Beschränktheit der Folgen $\{\|H_k\|_2\}$ und $\{\|H_k^{-1}\|_2\}$. Insbesondere ist dies natürlich für $H_k = I$, also das Gradientenverfahren, der Fall. Daher ist z. B. das Gradientenverfahren (d. h. $p_k := -g_k$), kombiniert mit einer der angegebenen Schrittweitenstrategien, global konvergent bei der Rosenbrock-Funktion

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Denn diese besitzt genau einen stationären Punkt, nämlich $x^* = (1, 1)$, ferner sind offenbar die Voraussetzungen (V) (a)–(c) erfüllt. \square

Nun betrachten wir noch die Anwendung des Modellalgorithmus bei einer glatten, gleichmäßig konvexen Zielfunktion. Hierzu fassen wir zunächst einige Hilfsmittel in dem folgenden Lemma zusammen.

Lemma 2.6 Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die folgenden Konvexitäts- und Glattheitsvoraussetzungen an die Zielfunktion f seien erfüllt:

- (K) (a) Mit einem $x_0 \in \mathbb{R}^n$ ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.
- (b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und auf L_0 gleichmäßig konvex, d. h. es existiert eine Konstante $c > 0$ mit

$$\frac{c}{2} \|y - x\|_2^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \quad \text{für alle } x, y \in L_0.$$

- (c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2 \quad \text{für alle } x, y \in L_0.$$

Dann ist die Niveaumenge L_0 kompakt, (P) besitzt daher eine globale Lösung x^* , diese liegt in L_0 und ist die einzige stationäre Lösung von (P) in L_0 . Ferner gilt die Fehlerabschätzung

$$\frac{c}{2} \|x - x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2c} \|\nabla f(x)\|_2^2 \quad \text{für alle } x \in L_0.$$

Beweis: Die Niveaumenge L_0 ist abgeschlossen. Für alle $x \in L_0$ ist wegen der gleichmäßigen Konvexität von f ferner

$$\frac{c}{2} \|x - x_0\|_2^2 + \nabla f(x_0)^T(x - x_0) \leq f(x) - f(x_0) \leq 0$$

und daher mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$L_0 \subset \left\{ x \in \mathbb{R}^n : \|x - x_0\|_2 \leq \frac{2}{c} \|\nabla f(x_0)\|_2 \right\}.$$

Insgesamt ist L_0 kompakt, die auf L_0 stetige Funktion f nimmt auf L_0 ihr (globales) Minimum an. Da eine globale Lösung von (P) nicht außerhalb von L_0 liegen kann, ist die Existenz einer globalen Lösung $x^* \in L_0$ bewiesen. Natürlich ist $\nabla f(x^*) = 0$, also x^* auch eine stationäre Lösung von (P). Wir zeigen nun noch die behaupteten Abschätzungen, aus denen insbesondere die Eindeutigkeit einer stationären Lösung von (P) in L_0 folgt.

Die erste Ungleichung folgt direkt aus der vorausgesetzten gleichmäßigen Konvexität von f , indem man $y = x$ und $x = x^*$ setzt. Bei festem $x \in L_0$ ist

$$-\frac{1}{2c} \|\nabla f(x)\|_2^2 \leq \frac{c}{2} \|x^* - x\|_2^2 + \nabla f(x)^T(x^* - x) \leq f(x^*) - f(x).$$

Dies erkennt man daran, daß die Aufgabe

$$\text{Minimiere } f_x(p) := \frac{c}{2} \|p\|_2^2 + \nabla f(x)^T p, \quad p \in \mathbb{R}^n$$

die eindeutige Lösung $p^* := -(1/c) \nabla f(x)$ besitzt. Insgesamt ist der Satz damit bewiesen. \square

Im folgenden Satz wird bei gleichmäßig konvexer Zielfunktion eine hinreichende Konvergenzbedingung für den Modellalgorithmus angegeben.

Satz 2.7 Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K) (a)–(c) aus Lemma 2.6 seien erfüllt. Zur Lösung von (P) betrachte man den Modellalgorithmus mit Abstiegsrichtungen p_k und Schrittweiten $t_k = t^*(x_k, p_k)$ (exakte Schrittweite), $t_k = t_P(x_k, p_k)$ (Powell-Schrittweite) oder $t_k = t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung sei $g_k := \nabla f(x_k)$ gesetzt. Schließlich sei

$$\delta_k := \begin{cases} \min \left[-\frac{g_k^T p_k}{\|g_k\|_2^2}, \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2 \right] & \text{falls } t_k = t_A(x_k, p_k), \\ \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2 & \text{falls } t_k = t^*(x_k, p_k), t_P(x_k, p_k). \end{cases}$$

Dann gilt:

1. Ist $\sum_{j=0}^{\infty} \delta_j = \infty$, so konvergiert die durch den Modellalgorithmus erzeugte Folge $\{x_k\}$ gegen die eindeutige (globale) Lösung x^* von (P).
2. Existiert ein $\delta > 0$ mit $\delta(k+1) \leq \sum_{j=0}^k \delta_j$ für $k = 0, 1, \dots$, so existieren Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\|_2 \leq C q^k$ für $k = 0, 1, \dots$.

Beweis: Wegen der Sätze 2.2, 2.3 und 2.4 sowie der Definition von δ_k existiert eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \delta_k \|g_k\|_2^2 \geq 2c\theta \delta_k [f(x_k) - f(x^*)],$$

wobei auch die Fehlerabschätzung aus Lemma 2.6 benutzt wurde. Daher ist

$$\begin{aligned} 0 \leq f(x_{k+1}) - f(x^*) &\leq (1 - 2c\theta \delta_k) [f(x_k) - f(x^*)] \\ &\leq \prod_{j=0}^k (1 - 2c\theta \delta_j) [f(x_0) - f(x^*)] \\ &\leq \exp \left(-2c\theta \sum_{j=0}^k \delta_j \right) [f(x_0) - f(x^*)]. \end{aligned}$$

Wegen $\sum_{j=0}^{\infty} \delta_j = \infty$ konvergiert $\{f(x_k)\}$ gegen $f(x^*)$. Wiederum wegen der Fehlerabschätzung in Lemma 2.6 folgt die Konvergenz von $\{x_k\}$ gegen x^* .

Existiert ein $\delta > 0$ mit $\delta(k+1) \leq \sum_{j=0}^k \delta_j$ für $k = 0, 1, \dots$, so ist

$$f(x_k) - f(x^*) \leq \exp \left(-2c\theta \sum_{j=0}^{k-1} \delta_j \right) [f(x_0) - f(x^*)] \leq \exp(-2c\theta\delta k) [f(x_0) - f(x^*)].$$

Mit Hilfe von $\|x_k - x^*\|_2 \leq \{2[f(x_k) - f(x^*)]/c\}^{1/2}$ (siehe Lemma 2.6) folgt daher

$$\|x_k - x^*\|_2 \leq \left\{ \frac{2[f(x_0) - f(x^*)]}{c} \right\}^{1/2} \exp(-c\theta\delta)^k, \quad k = 0, 1, \dots$$

Der Satz ist damit bewiesen. \square

Bemerkung: Die Schrittweitenstrategien gingen wiederum nur dadurch ein, daß die Aussagen der Sätze 2.2–2.4 benutzt wurden.

Wird im Modellalgorithmus unter den Voraussetzungen von Satz 2.7 stets die exakte Schrittweite oder die Powell-Schrittweite gewählt, so ist

$$\delta_k = \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2.$$

Die Bedingung $\sum_{j=0}^{\infty} \delta_j = \infty$ besagt, daß der Winkel zwischen $-g_k$ und p_k sich zwar einem rechten Winkel annähern, dies aber nicht zu schnell geschehen darf. \square

7.2.3 Das gedämpfte Gauß-Newton-Verfahren bei diskreten, nichtlinearen Approximationsaufgaben

In diesem Unterabschnitt wollen wir uns überlegen, daß man ganz analog zum glatten Fall auch (einfache) Konvergenzaussagen für ein Verfahren bei „halbglatten“, unrestrictierten Optimierungsaufgaben machen kann. Hier ist die Zielfunktion f durch $f = g \circ F$ mit einer konvexen Funktion $g: \mathbb{R}^m \rightarrow \mathbb{R}$ und einer glatten Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ gegeben. Wir werden uns auf diskrete, nichtlineare Approximationsaufgaben beschränken und die Aufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\| = g \circ F(x), \quad x \in \mathbb{R}^n$$

betrachten, bei der $g(y) := \|y\|$ mit einer gegebenen Norm $\|\cdot\|$ auf dem \mathbb{R}^m und $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine (glatte) Abbildung ist. Die wichtigsten Spezialfälle sind natürlich durch $\|\cdot\| = \|\cdot\|_1$, $\|\cdot\| = \|\cdot\|_2$ oder $\|\cdot\| = \|\cdot\|_\infty$ gegeben. Analog zu den Voraussetzungen in den letzten beiden Unterabschnitten wird diesmal gefordert:

- (V) (a) Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (Startpunkt des Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt.
- (b) Die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ ist auf einer offenen Obermenge von L_0 stetig differenzierbar.
- (c) Die Funktionalmatrix $F'(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|F'(x) - F'(y)\| \leq \gamma \|x - y\| \quad \text{für alle } x, y \in L_0.$$

(Hierbei ist rechts $\|\cdot\|$ eine feste Norm auf dem \mathbb{R}^n , während $\|\cdot\|$ links die der vorgegebenen Norm auf dem \mathbb{R}^m und der Norm auf dem \mathbb{R}^n zugeordnete Matrixnorm auf $\mathbb{R}^{m \times n}$ bedeutet⁴.)

⁴Die Lipschitzstetigkeit ist wegen der Äquivalenz der Normen eine normunabhängige Eigenschaft. Die eben getroffene Vereinbarung dient der Festlegung der Lipschitzkonstanten γ .

Unter diesen Voraussetzungen besitzt f in jedem $x \in L_0$ eine Gateaux-Variation $f'(x; \cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$, die durch $f'(x; p) = g'(F(x); F'(x)p)$ gegeben ist. Dieses Ergebnis haben wir in Satz 1.8 exemplarisch für $g(y) := \|y\|_\infty$ bewiesen.

Im folgenden Lemma wird gezeigt, wie man in einer aktuellen Näherung $x \in L_0$ eine Abstiegsrichtung berechnen oder feststellen kann, daß x eine stationäre Lösung von (P) ist.

Lemma 2.8 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Voraussetzungen (V) (a)–(c) seien erfüllt. Bei gegebenem $x \in L_0$ betrachte man die in x linearisierte Aufgabe

$$(LP_x) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|, \quad p \in \mathbb{R}^n.$$

Dann existiert eine (nicht notwendig eindeutige) Lösung p^* zu (LP_x) . Ferner gilt:

1. Ist $f_x(p^*) = f(x)$, so ist x eine stationäre Lösung von (P), d. h. es ist $f'(x; q) \geq 0$ für alle $q \in \mathbb{R}^n$.
2. Ist $f_x(p) < f(x)$ für ein $p \in \mathbb{R}^n$, so ist $f'(x; p) \leq f_x(p) - f(x) < 0$, also p eine Abstiegsrichtung für f in x . Insbesondere ist p^* eine Abstiegsrichtung für f in x , wenn $f_x(p^*) \neq f(x)$.

Beweis: Um die Lösbarkeit von (LP_x) nachzuweisen, betrachten wir die äquivalente Aufgabe

$$\text{Minimiere } g_x(q) := \|F(x) + q\|, \quad q \in \text{Bild}(F'(x)).$$

Da die Niveaumenge $\{q \in \mathbb{R}^m : g_x(q) \leq g_x(0)\} \cap \text{Bild}(F'(x))$ offensichtlich kompakt ist, folgt die Existenz einer Lösung p^* von (LP_x) .

Ist $f_x(p^*) = f(x)$, so ist $p = 0$ eine Lösung und damit auch eine stationäre Lösung der linearisierten Aufgabe (LP_x) . Folglich ist

$$f'_x(0; q) = g'(F(x); F'(x)q) = f'(x; q) \geq 0 \quad \text{für alle } q \in \mathbb{R}^n,$$

und daher x eine stationäre Lösung von (P).

Ist $f_x(p) < f(x)$, so ist wegen Lemma 1.6

$$f'(x; p) = g'(F(x); F'(x)p) \leq g(F(x) + F'(x)p) - g(F(x)) = f_x(p) - f(x) < 0,$$

d. h. p ist eine Abstiegsrichtung für die Zielfunktion f der diskreten, nichtlinearen Approximationsaufgabe (P) in der aktuellen Näherung x . Ist $f_x(p^*) \neq f(x) = f_x(0)$, so ist notwendig $f_x(p^*) < f(x)$ und damit p^* eine Abstiegsrichtung. Das Lemma ist damit bewiesen. \square

Bemerkung: Für $\|\cdot\| = \|\cdot\|_2$ ist (LP_x) ein lineares Ausgleichsproblem, das unter der Voraussetzung $m \geq n$ und Rang $F'(x) = n$ z. B. mit Hilfe einer QR-Zerlegung von $F'(x)$ gelöst werden kann (siehe Abschnitt 1.5). Ist $\|\cdot\| = \|\cdot\|_\infty$, so ist (LP_x) äquivalent dem linearen Programm

$$\text{Minimiere } \delta \text{ unter den Nebenbedingungen } \delta \geq 0, \quad -\delta e \leq F(x) + F'(x)p \leq \delta e,$$

wobei $e \in \mathbb{R}^m$ einmal wieder der Vektor ist, dessen Komponenten alle gleich Eins sind. Damit kann in diesem Falle eine Lösung von (LP_x) mit Hilfe des Simplexverfahrens (das man aber auf das zu (LP_x) duale lineare Programm anwenden sollte!) berechnet werden, siehe Abschnitt 6.3. Ähnliches gilt für andere, sogenannte *polyhedrale* Normen (siehe z. B. D. H. ANDERSON, M. R. OSBORNE (1977) und M. R. OSBORNE (1985, S. 128 ff.)). \square

Durch Lemma 2.8 ist eine Richtungsstrategie für ein Verfahren zur Lösung von (P) gegeben. Nun kommen wir zur Definition von Schrittweitenstrategien. Hierbei beschränken wir uns auf die Übertragung der Armijo-Schrittweite und nehmen an, $x \in L_0$ und $p \in \mathbb{R}^n$ mit $f_x(p) := \|F(x) + F'(x)p\| < f(x)$ seien vorgegeben. Wegen des eben bewiesenen Lemmas 2.8 ist p eine Abstiegsrichtung für die Zielfunktion f von (P) in x . Analog zum glatten Fall definieren wir die zugehörige *Armijo-Schrittweite* durch den folgenden Algorithmus:

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ gegeben, setze $\rho_0 := 1$.
- Für $j = 0, 1, \dots$:
 - Falls $f(x + \rho_j p) \leq f(x) + \alpha \rho_j [f_x(p) - f(x)]$, dann: $t := \rho_j$, STOP.
 - Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Es ist klar, daß die Armijo-Schrittweite existiert bzw. der obige Algorithmus nach endlich vielen Schritten abbricht. Denn wäre die zu testende Ungleichung für kein j erfüllt, so wäre $\{\rho_j\} \subset \mathbb{R}_+$ eine Nullfolge und

$$\alpha [f_x(p) - f(x)] \leq \lim_{j \rightarrow \infty} \frac{f(x + \rho_j p) - f(x)}{\rho_j} = f'(x; p) \leq f_x(p) - f(x)$$

würde den Widerspruch $0 \leq (1 - \alpha) [f_x(p) - f(x)]$ ergeben. Wie im glatten Fall kann $l = u =: \rho$ gewählt werden, so daß dann die Armijo-Schrittweite durch $t = \rho^j$ gegeben ist, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x + \rho^j p) \leq f(x) + \alpha \rho^j [f_x(p) - f(x)]$$

ist. Entsprechend S. P. HAN (1981) kann aber auch $\rho_0 := 1$ und

$$\rho_{j+1} := \max(0.1 \rho_j, \rho_j^*) \quad \text{mit} \quad \rho_j^* := \frac{0.5 \rho_j^2 [f(x) - f_x(p)]}{f(x + \rho_j p) - f(x) + \rho_j [f(x) - f_x(p)]}$$

gesetzt werden.

Das folgende Lemma wird dazu dienen, ganz entsprechend zu Lemma 2.1, die durch die Armijo-Schrittweite erzielte Verminderung der Zielfunktion nach unten abzuschätzen.

Lemma 2.9 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_x(p) := \|F(x) + F'(x)p\| < f(x)$, also p eine Abstiegsrichtung für f in x . Mit

$$t^* := \frac{2 [f(x) - f_x(p)]}{\gamma \|p\|^2}$$

ist dann

$$f(x + tp) \leq f(x) + t[f_x(p) - f(x)] + t^2 \frac{\gamma}{2} \|p\|^2 \quad \text{für alle } t \in [0, \min(1, t^*)].$$

Beweis: Wie in Lemma 2.1 sei \hat{t} die erste positive Nullstelle von $f(x) - f(x + tp)$. Dann ist $x + sp \in L_0$ für alle $s \in [0, \hat{t}]$. Für $t \in [0, \hat{t}]$ ist

$$\begin{aligned} F(x + tp) &= F(x) + tF'(x)p + \int_0^t [F'(x + sp) - F'(x)]p \, ds \\ &= (1-t)F(x) + t[F(x) + F'(x)p] + \int_0^t [F'(x + sp) - F'(x)]p \, ds. \end{aligned}$$

Für $t \in [0, \min(1, \hat{t})]$ folgt daher

$$f(x + tp) \leq (1-t)f(x) + tf_x(p) + t^2 \frac{\gamma}{2} \|p\|^2 = f(x) + t[f_x(p) - f(x)] + t^2 \frac{\gamma}{2} \|p\|^2.$$

Ist $\hat{t} \leq 1$, so folgt hieraus (setze $t = \hat{t}$), daß $t^* \leq \hat{t}$. Daher ist $\min(1, t^*) \leq \min(1, \hat{t})$ und das Lemma ist bewiesen. \square

Entsprechend Satz 2.4 kann auch hier die Verminderung der Zielfunktion abgeschätzt werden.

Satz 2.10 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_x(p) := \|F(x) + F'(x)p\| < f(x)$. Seien $\alpha \in (0, \frac{1}{2})$, $0 < l \leq u < 1$ gegeben und $t := \rho_j$ eine zugehörige Armijo-Schrittweite. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u , nicht aber von x oder p abhängt, mit

$$f(x) - f(x + tp) \geq \theta \min \left[f(x) - f_x(p), \left(\frac{f(x) - f_x(p)}{\|p\|} \right)^2 \right].$$

Beweis: Der Beweis verläuft völlig analog dem zu Satz 2.4. Statt Lemma 2.1 wird lediglich Lemma 2.9 angewandt. Er bleibt daher dem Leser überlassen. \square

Im folgenden Satz wird das (durch die Armijo-Schrittweite) *gedämpfte Gauß-Newton-Verfahren* zur Lösung von (P) angegeben und hierfür eine Konvergenzaussage gemacht (siehe auch D. H. ANDERSON, M. R. OSBORNE (1977), M. R. OSBORNE, G. A. WATSON (1978)).

Satz 2.11 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Die Voraussetzungen (V) (a)–(c) seien erfüllt. Zusätzlich sei $\text{Rang } F'(x) = n$ für alle $x \in L_0$. Zur Lösung von (P) betrachte man das folgende Verfahren:

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ (für die Armijo-Schrittweite) vorgegeben. Sei x_0 Startelement wie in Voraussetzung (V).
- Für $k = 0, 1, \dots$:

Sei p_k eine Lösung der in x_k linearisierten Approximationsaufgabe

$$(LP_k) \quad \text{Minimiere } f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad p \in \mathbb{R}^n.$$

Falls $f_k(p_k) = f(x_k)$, dann: x_k ist stationäre Lösung von (P), STOP.

Berechne Armijo-Schrittweite $t_k = \rho_j$.

Setze $x_{k+1} := x_k + t_k p_k$.

Dann gilt: Bricht dieser Algorithmus nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert er eine Folge $\{x_k\} \subset L_0$ mit $f(x_{k+1}) < f(x_k)$ für $k = 0, 1, \dots$ und der Eigenschaft, daß jeder Häufungspunkt x^* von $\{x_k\}$ eine stationäre Lösung von (P) ist. Besitzt insbesondere (P) genau eine stationäre Lösung x^* in L_0 , so konvergiert die gesamte Folge $\{x_k\}$ gegen x^* .

Beweis: Die Durchführbarkeit des Algorithmus ist wegen Lemma 2.8 sowie der Existenz der Armijo-Schrittweite klar. O. B. d. A. nehmen wir an, das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Da es sich um ein Abstiegsverfahren handelt, wird eine Folge $\{x_k\} \subset L_0$ mit $f(x_{k+1}) < f(x_k)$ erzeugt. Sei x^* ein Häufungspunkt von $\{x_k\}$. Der Nachweis dafür, daß x^* eine stationäre Lösung von (P) ist, erfolgt in drei Schritten.

- (a) Es existiert eine Konstante $C > 0$ mit der Eigenschaft: Ist $x \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $\|F(x) + F'(x)p\| \leq f(x)$, so ist $\|p\| \leq C f(x)$. Insbesondere ist daher $\|p_k\| \leq C f(x_k) \leq C f(x_0)$, also $\{p_k\}$ eine beschränkte Folge.

Denn: Sei $x \in L_0$ und $\|F(x) + F'(x)p\| \leq f(x)$. Dann ist $\|F'(x)p\| \leq 2f(x)$. Da $F'(x) \in \mathbb{R}^{m \times n}$ den Rang n besitzt, ist $F'(x)^T F'(x) \in \mathbb{R}^{n \times n}$ nichtsingulär und daher

$$\begin{aligned} \|p\| &= \|[F'(x)^T F'(x)]^{-1} F'(x)^T F'(x)p\| \\ &\leq \|[F'(x)^T F'(x)]^{-1} F'(x)^T\| \|F'(x)p\| \\ &\leq C f(x) \quad \text{mit} \quad C := 2 \max_{y \in L_0} \|[F'(y)^T F'(y)]^{-1} F'(y)^T\|. \end{aligned}$$

(Beim Beweis dieses Teiles geht die Rang-Voraussetzung an $F'(\cdot)$ ein.)

- (b) Es ist $\lim_{k \rightarrow \infty} [f(x_k) - f_k(p_k)] = 0$.

Denn: Wegen Satz 2.10 existiert eine (von k unabhängige) Konstante $\theta > 0$ mit

$$(*) \quad f(x_k) - f(x_{k+1}) \geq \theta \min \left[f(x_k) - f_k(p_k), \left(\frac{f(x_k) - f_k(p_k)}{\|p_k\|} \right)^2 \right], \quad k = 0, 1, \dots$$

Da $\{f(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge ist, gilt

$$\lim_{k \rightarrow \infty} [f(x_k) - f(x_{k+1})] = 0.$$

Hieraus, aus der Beschränktheit von $\{p_k\}$ sowie (*) folgt $\lim_{k \rightarrow \infty} [f(x_k) - f_k(p_k)] = 0$. (Dies ist der Teil des Beweises, in dem im wesentlichen die Schrittweitenstrategie eingeht.)

- (c) Da x^* ein Häufungspunkt von $\{x_k\}$ und $\{p_k\}$ beschränkt ist, existiert eine unendliche Teilmenge $K \subset \mathbb{N}$ derart, daß $\{x_k\}_{k \in K}$ gegen x^* und $\{p_k\}_{k \in K}$ gegen ein p^* konvergiert. Dann ist p^* eine Lösung von

$$(LP_*) \quad \text{Minimiere } f_*(p) := \|F(x^*) + F'(x^*)p\|, \quad p \in \mathbb{R}^n$$

mit $f_*(p^*) = f(x^*)$, so daß x^* nach Lemma 2.8 eine stationäre Lösung von (P) ist.

Denn: Sei $p \in \mathbb{R}^n$ beliebig. Nach Definition von p_k ist

$$\|F(x_k) + F'(x_k)p_k\| \leq \|F(x_k) + F'(x_k)p\|, \quad k = 0, 1, \dots$$

Läßt man hier $k \in K$ nach ∞ laufen, so erhält man, daß p^* eine Lösung von (LP_*) ist. Mit Teil (b) folgt

$$0 = \lim_{k \in K, k \rightarrow \infty} [f(x_k) - f_k(p_k)] = f(x^*) - f_*(p^*).$$

(In diesem Teil des Beweises geht die Richtungsstrategie ein.) Die letzte Behauptung des Satzes, daß sich aus der Eindeutigkeit einer stationären Lösung in L_0 die Konvergenz der gesamten Folge $\{x_k\}$ ergibt, ist einfach einzusehen und folgt wie in Satz 2.5. Insgesamt ist der Satz bewiesen. \square

Satz 2.11 macht eine (schwache) *globale* Konvergenzaussage unter der (starken) Voraussetzung, daß $F'(x) \in \mathbb{R}^{m \times n}$ maximalen Rang auf der Niveaumenge L_0 besitzt.

Zum Schluß dieses Abschnittes wollen wir annehmen, das Verfahren aus Satz 2.11 liefere eine gegen ein x^* konvergente Folge $\{x_k\}$. Uns interessiert die Frage, ob unter geeigneten Voraussetzungen $t_k = 1$ für alle hinreichend großen k gilt, ob also das gedämpfte Gauß-Newton-Verfahren nach endlich vielen Schritten in das ungedämpfte übergeht. Ist dies der Fall, so erwartet man ferner eine Aussage über die Konvergenzgeschwindigkeit der Folge $\{x_k\}$. Von entscheidender Bedeutung für die Beantwortung dieser Fragen ist die folgende Definition.

Definition 2.12 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Ein $x^* \in \mathbb{R}^n$ heißt *lokal stark eindeutige Lösung* von (P), wenn es positive Konstanten σ und δ mit

$$f(x) \geq f(x^*) + \sigma \|x - x^*\| \quad \text{für alle } x \in B[x^*; \delta] := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$$

gibt.

Die Bedeutung der (lokalen) starken Eindeutigkeit für die Konvergenzgeschwindigkeit von Iterationsverfahren wurde wohl zuerst von L. CROMME (1976, 1978) erkannt. Unser Ziel ist es, den folgenden Satz zu beweisen.

Satz 2.13 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei sei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\|\cdot\|$ eine Norm auf dem \mathbb{R}^m und $m \geq n$. Das Verfahren aus Satz 2.11 liefere eine gegen ein $x^* \in \mathbb{R}^n$ konvergente Folge $\{x_k\}$. Sei x^* eine lokal stark eindeutige Lösung von (P), F stetig differenzierbar und $F'(\cdot)$ lipschitzstetig auf einer Umgebung von x^* . Dann gilt:

1. Für alle hinreichend großen k ist $t_k = 1$. Nach endlich vielen Schritten geht also das gedämpfte Gauß-Newton-Verfahren in das ungedämpfte über.
2. Die Folge $\{x_k\}$ konvergiert von mindestens zweiter Ordnung gegen x^* , d. h. es existiert eine Konstante $C > 0$ derart, daß $\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$ für alle hinreichend großen k .

Beweis: Wegen der lokalen starken Eindeutigkeit von x^* und der Lipschitzstetigkeit von $F'(\cdot)$ auf einer Umgebung von x^* existieren positive Konstanten σ, δ und L mit

$$f(x) \geq f(x^*) + \sigma \|x - x^*\|, \quad \|F'(x) - F'(y)\| \leq L \|x - y\|$$

für alle $x, y \in B[x^*; \delta]$. O. B. d. A. bricht das Verfahren aus Satz 2.11 nicht vorzeitig ab, so daß $x_k \neq x^*$ für alle k angenommen werden kann. Der Beweis der beiden Behauptungen im Satz erfolgt in mehreren Schritten.

- (1) Sind $x_k, x_k + p \in B[x^*; \delta]$, so ist

$$|f(x_k + p) - f_k(p)| \leq \|F(x_k + p) - F(x_k) - F'(x_k)p\| \leq \frac{L}{2} \|p\|^2.$$

- (2) Es ist $\|p_k\| \leq 2 \|x_k - x^*\|$ für alle hinreichend großen k . Insbesondere konvergiert die Folge $\{p_k\}$ gegen den Nullvektor.

Denn: Wegen der vorausgesetzten Konvergenz der Folge $\{x_k\}$ gegen x^* existiert ein $k_0 \in \mathbb{N}$ mit

$$\|x_k - x^*\| \leq \frac{\delta}{3}, \quad \sigma - \frac{5L}{2} \|x_k - x^*\| > 0 \quad \text{für alle } k \geq k_0.$$

Wir wollen zeigen:

$$(*) \quad k \geq k_0, \quad \|p\| \geq 2 \|x_k - x^*\| \implies f_k(p) > f_k(x^* - x_k).$$

Ist (*) bewiesen, so wissen wir: Ist $k \geq k_0$, so ist $\|p_k\| \leq 2 \|x_k - x^*\|$. Daher bleibt (*) zu zeigen.

Hierzu nehmen wir zunächst an, es sei $\|p\| = 2 \|x_k - x^*\|$. Nach (zweimaliger) Anwendung von (1) und wegen der lokal starken Eindeutigkeit von x^* ist

$$\begin{aligned} f_k(p) - f_k(x^* - x_k) &\stackrel{(1)}{\geq} -2L \|x_k - x^*\|^2 + f(x_k + p) - f_k(x^* - x_k) \\ &\geq -2L \|x_k - x^*\|^2 + \sigma \|x_k + p - x^*\| + f(x^*) - f_k(x^* - x_k) \\ &\stackrel{(1)}{\geq} -\frac{5L}{2} \|x_k - x^*\| + \sigma \|x_k + p - x^*\| \\ &\geq \left(\sigma - \frac{5L}{2} \|x_k - x^*\| \right) \|x_k - x^*\| \\ &> 0 \end{aligned}$$

für alle $k \geq k_0$. Nun sei $\|p\| > 2\|x_k - x^*\|$. Man bestimme ein $\lambda \in (0, 1)$ mit

$$\|(1-\lambda)p + \lambda(x^* - x_k)\| = 2\|x_k - x^*\|$$

und setze anschließend $q := (1-\lambda)p + \lambda(x^* - x_k)$. Für alle $k \geq k_0$ ist dann

$$f_k(x^* - x_k) < f_k(q) \leq (1-\lambda)f_k(p) + \lambda f_k(x^* - x_k)$$

(wegen der Konvexität von f_k) und daher $f_k(x^* - x_k) < f_k(p)$. Damit ist (2) bewiesen.

- (3) Es existiert eine Konstante $c_0 > 0$ derart, daß $f(x_k) - f_k(p_k) \geq c_0 \|p_k\|$ für alle hinreichend großen k .

Denn: Sei $k_0 \in \mathbb{N}$ wie im Beweis von (2) bestimmt. Für $k \geq k_0$ ist wegen (1) und (2)

$$f(x_k) - f_k(p_k) \geq [f(x_k) - f(x^*)] + [f(x^*) - f_k(x^* - x_k)] \geq \frac{4\sigma}{5} \|x_k - x^*\| \geq \frac{2\sigma}{5} \|p_k\|.$$

Damit ist auch (3) bewiesen.

- (4) Bei vorgegebenem $\alpha \in (0, 1)$ ist

$$\frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)} \geq \alpha \quad \text{für alle hinreichend großen } k.$$

Insbesondere ist $t_k = 1$ und $x_{k+1} = x_k + p_k$ für alle hinreichend großen k .

Denn: Wegen (1), (2) und (3) ist

$$\frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)} = 1 + \frac{f_k(p_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)} \geq 1 - \frac{L}{2c_0} \|p_k\| \geq \alpha$$

für alle hinreichend großen k .

- (5) Die Folge $\{x_k\}$ konvergiert von mindestens zweiter Ordnung gegen x^* .

Denn: Wegen der lokal starken Eindeutigkeit von x^* und (1)–(4) ist

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq \frac{1}{\sigma} [f(x_k + p_k) - f(x^*)] \\ &= \frac{1}{\sigma} [f(x_k + p_k) - f_k(p_k) + f_k(p_k) - f(x^*)] \\ &\leq \frac{1}{\sigma} [2L\|x_k - x^*\|^2 + f_k(x^* - x_k) - f(x^*)] \\ &\leq \frac{5L}{2\sigma} \|x_k - x^*\|^2 \end{aligned}$$

für alle hinreichend großen k . Insgesamt ist der Satz bewiesen. \square

Bemerkungen: Die entscheidende Voraussetzung in Satz 2.13 ist die der lokal starken Eindeutigkeit von x^* , des Grenzwertes einer durch das gedämpfte Gauß-Newton-Verfahren aus Satz 2.11 erzeugten Folge $\{x_k\}$. Es liegt also nahe, diese Voraussetzung etwas „abzuklopfen“.

Ist $f(x^*) = 0$ bzw. $F(x^*) = 0$, so ist x^* eine lokal stark eindeutige Lösung der diskreten, nichtlinearen Approximationsaufgabe (P), wenn Rang $F'(x^*) = n$. Denn definiert man $\tau > 0$ durch $\tau := \min_{\|p\|=1} \|F'(x^*)p\|$ und wählt anschließend $\delta > 0$ so klein, daß

$$\|F(x) - \underbrace{F(x^*)}_{=0} - F'(x^*)(x - x^*)\| \leq \frac{\tau}{2} \|x - x^*\| \quad \text{für alle } x \in B[x^*; \delta],$$

so ist

$$\tau \|x - x^*\| \leq \|F'(x^*)(x - x^*)\| \leq f(x) + \|F(x) - F'(x^*)(x - x^*)\| \leq f(x) + \frac{\tau}{2} \|x - x^*\|,$$

daher

$$\frac{\tau}{2} \|x - x^*\| + \underbrace{f(x^*)}_{=0} \leq f(x) \quad \text{für alle } x \in B[x^*; \delta]$$

und folglich x^* eine lokal stark eindeutige Lösung von (P).

Ist x^* eine lokal stark eindeutige Lösung der diskreten, nichtlinearen Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n$$

und ist $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ in x^* stetig differenzierbar, so existiert eine Konstante $\sigma > 0$ derart, daß

$$\sigma \|p\| \leq f'(x^*; p) \leq \|F(x^*) + F'(x^*)p\| - \|F(x^*)\| \quad \text{für alle } p \in \mathbb{R}^n$$

bzw.

$$(*) \quad \sigma \|p\| + \|F(x^*)\| \leq \|F(x^*) + F'(x^*)p\| \quad \text{für alle } p \in \mathbb{R}^n$$

gilt. Mit anderen Worten ist *notwendig* $p^* := 0$ eine (global) stark eindeutige Lösung der linearisierten Aufgabe

$$(LP_*) \quad \text{Minimiere } f_*(p) := \|F(x^*) + F'(x^*)p\|, \quad p \in \mathbb{R}^n.$$

Aber auch die Umkehrung ist richtig! Denn existiert ein $\sigma > 0$ derart, daß $(*)$ gilt, so bestimme man ein hinreichend kleines $\delta > 0$ mit

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\| \leq \frac{\sigma}{2} \|x - x^*\| \quad \text{für alle } x \in B[x^*; \delta]$$

und erhalte aus $(*)$, daß

$$\sigma \|x - x^*\| + f(x^*) \leq \|F(x^*) + F'(x^*)(x - x^*)\| \leq f(x) + \frac{\sigma}{2} \|x - x^*\|$$

für alle $x \in B[x^*; \delta]$, daß also x^* eine lokal stark eindeutige Lösung von (P) ist. Daher ist x^* genau dann eine lokal stark eindeutige Lösung von (P), wenn $p^* := 0$ (global) stark eindeutige Lösung von (LP_*) ist.

Die bisherigen Untersuchungen zum Gauß-Newton-Verfahren bei der diskreten, nichtlinearen Approximationsaufgabe (P) sind *normunabhängig*, d. h. in den bisherigen Aussagen mußte *nicht* die Norm $\|\cdot\|$ bei der Definition der Zielfunktion $f(x) := \|F(x)\|$ von (P) spezifiziert werden. Daher könnte man den (falschen) Eindruck gewinnen, auch die Konvergenzeigenschaften des Gauß-Newton-Verfahrens seien normunabhängig.

Ist $F(x^*) = 0$ (und Rang $F'(x^*) = n$), so ist nach obigen Überlegungen x^* eine (unabhängig von der gewählten Norm) lokal stark eindeutige Lösung von (P), so daß das gedämpfte Gauß-Newton-Verfahren unter geeigneten Voraussetzungen von mindestens zweiter Ordnung gegen x^* konvergiert.

Für $F(x^*) \neq 0$ bzw. $f(x^*) > 0$ ist das aber wesentlich anders! Um das einzusehen, stellen wir die beiden wohl wichtigsten Spezialfälle, nämlich $\|\cdot\| := \|\cdot\|_2$ (in diesem Falle heißt (P) ein *nichtlineares Ausgleichsproblem*) und $\|\cdot\| := \|\cdot\|_\infty$ (dann heißt (P) ein *diskretes, nichtlineares Tschebyscheffsches Approximationsproblem*), einander gegenüber.

Ist $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ (mit $m \geq n$) in x^* stetig differenzierbar, so ist, wie man leicht nachweist, x^* genau dann eine lokal stark eindeutige Lösung des nichtlinearen Ausgleichsproblems

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$$

wenn $F(x^*) = 0$. Nur in diesem Falle ist also durch Satz 2.13 gesichert, daß eine durch das gedämpfte Gauß-Newton-Verfahren erzeugte, gegen x^* konvergente Folge $\{x_k\}$ sogar quadratisch gegen x^* konvergiert. Das ist nicht verwunderlich. Denn unter der Rang-Voraussetzung von Satz 2.11 lautet das gedämpfte Gauß-Newton-Verfahren

$$x_{k+1} := x_k + t_k p_k \quad \text{mit} \quad p_k := -[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k),$$

wobei p_k natürlich nicht „wörtlich“ nach der angegebenen Formel, sondern mit Hilfe einer QR -Zerlegung von $F'(x_k)$ berechnet wird. Andererseits ist das nichtlineare Ausgleichsproblem (P) äquivalent zu

$$\text{Minimiere } h(x) := \frac{1}{2} \|F(x)\|_2^2, \quad x \in \mathbb{R}^n.$$

Dann ist $\nabla h(x) = F'(x)^T F(x)$ und $\nabla^2 h(x) = F'(x)^T F'(x) + S(x)$ mit

$$S(x) := \sum_{i=1}^m F_i(x) \nabla^2 F_i(x).$$

Das (ungedämpfte) Newton-Verfahren zur Berechnung einer stationären Lösung von (P) bzw. eines stationären Punktes von h lautet daher

$$x_{k+1} := x_k - [F'(x_k)^T F'(x_k) + S(x_k)]^{-1} F'(x_k)^T F(x_k).$$

Hieran erkennt man den Unterschied zum Gauß-Newton-Verfahren sehr deutlich. Nur für $S(x^*) = 0$ wird man hoffen können, daß sich die lokale, quadratische Konvergenz des Newton-Verfahrens auf das Gauß-Newton-Verfahren überträgt. Ferner wird man

erwarten, daß für „kleines“ $S(x^*)$ (wenn also F „nur schwach nichtlinear“ oder $F(x^*)$ „klein“ ist) die Konvergenz des Gauß-Newton-Verfahrens befriedigend sein wird.

Anders sind die Verhältnisse beim diskreten, nichtlinearen Tschebyscheffschen Approximationsproblem. Hier kann es auch eine lokal stark eindeutige Lösung x^* mit $F(x^*) \neq 0$ geben. Dieses in sich interessante Ergebnis formulieren wir im folgenden Satz (siehe z. B. auch G. A. WATSON (1980, S. 32)). \square

Satz 2.14 Gegeben sei die diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^n.$$

Sei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ mit $m \geq n$ in $x^* \in \mathbb{R}^n$ stetig differenzierbar, x^* eine stationäre Lösung von (P) und die Haarsche Bedingung in x^* erfüllt, d. h. jede $n \times n$ -Untermatrix von $F'(x^*)$ sei nichtsingulär. Dann ist x^* eine lokal stark eindeutige Lösung von (P).

Beweis: O. B. d. A. kann $F(x^*) \neq 0$ angenommen werden. Definiert man zu der stationären Lösung x^* von (P) die Indexmenge

$$I(x^*) := \{i \in \{1, \dots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\},$$

so sagt Satz 1.9, daß reelle Zahlen $\lambda_i^*, i \in I(x^*)$, existieren mit

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \quad \sum_{i \in I(x^*)} \lambda_i^* = 1, \quad \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) = 0.$$

Mit anderen Worten liegt der Nullvektor 0 des \mathbb{R}^n in der *konvexen Hülle* der Vektoren $\{\operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) : i \in I(x^*)\}$, er läßt sich also als eine *Konvexitätskombination* dieser Vektoren darstellen. Die Anwendung eines bekannten Satzes von Carathéodory (siehe z. B. G. A. WATSON (1980, S. 12)) liefert, daß sich der Nullvektor 0 des \mathbb{R}^n sogar als Konvexitätskombination von *höchstens* $n+1$ dieser Vektoren darstellen läßt. Daher existieren eine Indexmenge $I^* \subset I(x^*)$ mit höchstens $n+1$ Elementen und reelle Zahlen $\mu_i^*, i \in I^*$, mit

$$(**) \quad \mu_i^* > 0 \quad (i \in I^*), \quad \sum_{i \in I^*} \mu_i^* = 1, \quad \sum_{i \in I^*} \mu_i^* \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) = 0.$$

Hätte I^* weniger als $n+1$ Elemente, so ließe sich der Nullvektor des \mathbb{R}^n als nichttriviale Linearkombination von n Zeilen von $F'(x^*)$ darstellen, was einen Widerspruch dazu bedeutet, daß jede $n \times n$ -Untermatrix von $F'(x^*)$ nichtsingulär ist. Daher enthält I^* genau $n+1$ Elemente. Das Ziel besteht darin, die Existenz einer Konstanten $\sigma > 0$ mit

$$(*) \quad \sigma \|p\|_\infty + \|F(x^*)\|_\infty \leq \|F(x^*) + F'(x^*)p\|_\infty \quad \text{für alle } p \in \mathbb{R}^n$$

zu zeigen, woraus nach obigen Bemerkungen die Behauptung folgt.

Sei $q \in \mathbb{R}^n \setminus \{0\}$. Dann ist $\max_{i \in I^*} \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*)^T q > 0$, denn andernfalls wäre

$$0 \geq \sum_{i \in I^*} \underbrace{\mu_i^* \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*)^T q}_{>0 \leq 0} = \left(\sum_{i \in I^*} \mu_i^* \operatorname{sign}[F_i(x^*)] \nabla F_i(x^*) \right)^T q = 0,$$

so daß $q \neq 0$ auf den $n+1$ Vektoren $\nabla F_i(x^*)$, $i \in I^*$, von denen je n linear unabhängig sind, senkrecht stehen würde, was einen Widerspruch bedeutet. Daher ist

$$0 < \sigma := \min_{\|q\|_\infty=1} \max_{i \in I^*} \text{sign}[F_i(x^*)] \nabla F_i(x^*)^T q.$$

Wir wollen zeigen, daß (*) gilt. Für $p = 0$ ist das trivialerweise der Fall, so daß wir $p \neq 0$ annehmen können. Zu $q := p/\|p\|_\infty$ existiert nach Definition von σ ein $k \in I^*$ mit $\sigma \leq \text{sign}[F_k(x^*)] \nabla F_k(x^*)^T q$. Dann ist

$$\begin{aligned} \|F(x^*) + F'(x^*)p\|_\infty &\geq |F_k(x^*) + \nabla F_k(x^*)^T p| \\ &= \left| \|F(x^*)\|_\infty + \underbrace{\text{sign}[F_k(x^*)] \nabla F_k(x^*)^T p}_{\geq \sigma \|p\|_\infty} \right| \\ &\geq \|F(x^*)\|_\infty + \sigma \|p\|_\infty, \end{aligned}$$

womit die Gültigkeit von (*) gezeigt und schließlich die lokale starke Eindeutigkeit von x^* bewiesen ist. \square

Aufgaben

- Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c) in 7.2.1. Sei $x \in L_0$ und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x , d. h. $\nabla f(x)^T p < 0$. Seien α und β mit $0 < \alpha < \beta < \frac{1}{2}$ vorgegeben. Hiermit definiere man

$$T(x, p) := \{t > 0 : f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p, \quad |\nabla f(x + tp)^T p| \leq -\beta \nabla f(x)^T p\}.$$

Man zeige:

- (a) Es ist $T(x, p) \neq \emptyset$.
- (b) Es existiert eine von x und p unabhängige Konstante $\theta > 0$ mit

$$f(x) - f(x + tp) \geq \theta \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2 \quad \text{für alle } t \in T(x, p).$$

Hinweis: Offenbar ist $T(x, p) \subset T_P(x, p)$, so daß (b) aus Satz 2.3 folgt. Siehe auch M. AL-BAALI (1985).

- Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c) in 7.2.1. Sei $x \in L_0$ und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x , d. h. $\nabla f(x)^T p < 0$. Seien $\alpha \in (0, 1)$, $\sigma > 0$ und $\rho \in (0, 1)$ vorgegeben. Folgendermaßen bestimme man eine Schrittweite $t = t(x, p)$:

- Wähle $\tau \geq -\sigma \frac{\nabla f(x)^T p}{\|p\|_2^2}$, bestimme die kleinste nichtnegative ganze Zahl j mit

$$f(x + \tau \rho^j p) \leq f(x) + \alpha \tau \rho^j \nabla f(x)^T p$$

und setze $t := \tau \rho^j$.

Dann existiert eine von x und p unabhängige Konstante $\theta > 0$ mit

$$f(x) - f(x + tp) \geq \theta \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2.$$

3. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c) in 7.2.1. Sei $x \in L_0$ und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x , d. h. $\nabla f(x)^T p < 0$. Bei vorgegebenem $\alpha \in (0, \frac{1}{2})$ definiere man

$$T_G(x, p) := \{t \in \mathbb{R}_+ : f(x) + (1 - \alpha)t \nabla f(x)^T p \leq f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p\}$$

(Menge der *Goldstein-Schrittweiten*). Analog zu Satz 2.3 zeige man:

- (a) Es ist $T_G(x, p) \neq \emptyset$.
- (b) Es existiert eine von x und p unabhängige Konstante $\theta > 0$ mit

$$f(x) - f(x + tp) \geq \theta \left(\frac{\nabla f(x)^T p}{\|p\|_2} \right)^2 \quad \text{für alle } t \in T_G(x, p).$$

4. Eine Funktion $\phi: [a, b] \rightarrow \mathbb{R}$ heißt *unimodal*, wenn es genau ein $t^* \in (a, b)$ gibt mit $\phi(t^*) = \min_{t \in [a, b]} \phi(t)$, und wenn ϕ auf $[a, t^*]$ monoton fallend und auf $[t^*, b]$ monoton wachsend ist. Zur Lokalisierung des Minimums t^* der auf $[a, b]$ unimodalen Funktion ϕ betrachte man die *Methode vom goldenen Schnitt*:

- $\epsilon > 0$ (gewünschte Genauigkeit) vorgegeben, $F := (\sqrt{5} - 1)/2$.
- Berechne $\begin{cases} s := a + (1 - F)(b - a), & \phi_s := \phi(s), \\ t := a + F(b - a), & \phi_t := \phi(t). \end{cases}$
- Solange $b - a > \epsilon$:

Falls $\phi_s > \phi_t$, dann:

$$a := s, \quad s := t, \quad t := a + F(b - a), \quad \phi_s := \phi_t, \quad \phi_t := \phi(t)$$

Andernfalls:

$$b := t, \quad t := s, \quad s := a + (1 - F)(b - a), \quad \phi_t := \phi_s, \quad \phi_s := \phi(s).$$

- $t^* \approx (a + b)/2$.

Man beweise, daß dieser Algorithmus nach endlich vielen Schritten mit einem Intervall $[a, b]$ abbricht, das t^* enthält.

5. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c) in 7.2.1. Zur Lösung von (P) wende man den Modellalgorithmus mit $p_k := -g_k$ (hierbei sei wieder $g_k := \nabla f(x_k)$) und der konstanten Schrittweite $t_k := 1/\gamma$ an. Dann ist jeder Häufungspunkt der durch das Verfahren erzeugten Folge $\{x_k\}$ eine stationäre Lösung von (P).

Hinweis: Bezeichnet t_k^* die exakte Schrittweite, so zeigt Satz 2.2, daß $1/\gamma \leq t_k^*$. Anschließend zeige man

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2\gamma} \|g_k\|_2^2,$$

und beweise hiermit die Behauptung.

6. Die Voraussetzungen von Satz 2.5 seien erfüllt. Die Zielfunktion f besitze in der Niveaumenge L_0 nur endlich viele stationäre Punkte. Der Modellalgorithmus erzeuge eine Folge $\{x_k\}$ mit $\lim_{k \rightarrow \infty} (x_{k+1} - x_k) = 0$. Dann konvergiert die gesamte Folge $\{x_k\}$ gegen einen der stationären Punkte von f .

Hinweis: Siehe J. M. ORTEGA, W. C. RHEINBOLDT (1970, S. 476).

7. Sei $Q \in \mathbb{R}^{n \times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert λ_{\min} und größtem Eigenwert λ_{\max} . Dann gilt die *Ungleichung von Kantorowitsch*:

$$(x^T Q x)(x^T Q^{-1} x) \leq \frac{(\lambda_{\min} + \lambda_{\max})^2}{4 \lambda_{\min} \lambda_{\max}} (x^T x)^2 \quad \text{für alle } x \in \mathbb{R}^n.$$

Hinweis: Durch eine orthogonale Ähnlichkeitstransformation kann man erreichen, daß o. B. d. A. Q eine Diagonalmatrix ist. Siehe auch D. G. LUENBERGER (1973, S. 151).

8. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit mit kleinstem Eigenwert λ_{\min} und größtem Eigenwert λ_{\max} . Für alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\left(\frac{x^T A x}{\|x\|_2 \|Ax\|_2} \right)^2 \geq \frac{4 \lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2} = \frac{4 \operatorname{cond}_2(A)}{(1 + \operatorname{cond}_2(A))^2}.$$

Hinweis: Man setze $z := A^{1/2}x$ und wende die Ungleichung von Kantorowitsch aus Aufgabe 7 an. Siehe auch D. F. SHANNO (1978a, Lemma 3.1) und als besonders hübsche Anwendung dieser Aufgabe Lemma 4.5.

9. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (K) (a)–(c) in Lemma 2.6. Auf (P) wende man das Gradienten-Verfahren (d. h. $p_k := -g_k$, wobei $g_k := \nabla f(x_k)$) mit exakter Schrittweite (d. h. $t_k := t^*(x_k, p_k)$) an. Mit x^* werde die globale Lösung von (P) bezeichnet.

- (a) Mit Hilfe von Satz 2.2 und Lemma 2.6 zeige man, daß

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{c}{\gamma}\right) [f(x_k) - f(x^*)], \quad k = 0, 1, \dots$$

- (b) Sei $Q \in \mathbb{R}^{n \times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert λ_{\min} und größtem Eigenwert λ_{\max} , ferner sei $f(x) := d^T x + \frac{1}{2} x^T Q x$. Dann sind die Voraussetzungen (K) (a)–(c) mit $c := \lambda_{\min}$ und $\gamma := \lambda_{\max}$ erfüllt. Die aus dem vorigen Teil der Aufgabe resultierende Abschätzung lässt sich zu

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 [f(x_k) - f(x^*)], \quad k = 0, 1, \dots$$

verbessern.

Hinweis: Man zeige

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}{\|g_k\|_2^4}$$

und wende die Ungleichung von Kantorowitsch an.

10. Gegeben sei das nichtlineare Ausgleichsproblem

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n.$$

Es sei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine Abbildung, die auf einer Umgebung U^* einer stationären Lösung $x^* \in \mathbb{R}^n$ von (P) stetig differenzierbar ist. Es sei $\text{Rang } F'(x^*) = n$, ferner existiere eine Konstante $\sigma \in [0, 1)$ mit

$$\|[F'(x) - F'(x^*)]^T F(x^*)\|_2 \leq \sigma \|x - x^*\|_* \quad \text{für alle } x \in U^*,$$

wobei die Norm $\|\cdot\|_*$ durch $\|x\|_* := \|F'(x^*)^T F'(x^*)x\|_2$ definiert ist. Dann gilt:

- (a) Zu jedem $q \in (\sigma, 1)$ gibt es ein $\delta > 0$ derart, daß das Gauß-Newton-Verfahren

$$x_{k+1} = x_k - [F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k)$$

für jedes x_0 mit $\|x_0 - x^*\|_* \leq \delta$ durchführbar ist und bezüglich der Norm $\|\cdot\|_*$ linear gegen x^* konvergiert:

$$\|x_{k+1} - x^*\|_* \leq q \|x_k - x^*\|_*, \quad k = 0, 1, \dots$$

- (b) Ist zusätzlich $F'(\cdot)$ auf U^* in x^* lipschitzstetig und ist $F(x^*) = 0$, so ist das Gauß-Newton-Verfahren lokal mindestens quadratisch konvergent gegen x^* .

Hinweis: Man definiere $\mu := 2 \max(\|F'(x^*)^T\|_2, \|[F'(x^*)^T F'(x^*)]^{-1}\|_2)$ und wähle anschließend $\epsilon > 0$ so klein, daß $(1 + \mu\epsilon)(\sigma + \mu\epsilon) \leq q$, was wegen $\sigma < q$ möglich ist. Zur Abkürzung setze man $G(x) := F'(x)^T F'(x)$ und bestimme $\delta > 0$ so klein, daß die Kugel $B_*(x^*; \delta) := \{x \in \mathbb{R}^n : \|x - x^*\|_* \leq \delta\}$ in U^* enthalten ist und für $x \in B_*(x^*; \delta)$ gilt:

- (i) $G(x)$ ist nichtsingulär, $\|G(x)^{-1}\|_2 \leq \mu$ und $\|F'(x)^T\|_2 \leq \mu$,
- (ii) $\|G(x^*) - G(x)\|_2 \leq \epsilon$, $\|G(x)G(x^*)^{-1} - I\|_2 \leq \epsilon$,
- (iii) $\|F(x^*) - F(x) - F'(x^*)(x^* - x)\|_2 \leq \epsilon \|x^* - x\|_*$.

Für $x \in B_*(x^*; \delta)$ ist daher $x_+ := x - G(x)^{-1} F'(x)^T F(x)$ definiert. Dann ist unter Berücksichtigung von $F'(x^*)^T F(x^*) = 0$ (da x^* stationäre Lösung) nach einfacher Rechnung

$$\begin{aligned} G(x^*)(x_+ - x^*) &= \{I + [G(x^*) - G(x)]G(x)^{-1}\} \{[F'(x^*) - F'(x)]^T F(x^*) \\ &\quad + F'(x)^T [F(x^*) - F(x) - F'(x^*)(x^* - x)]\} \end{aligned}$$

und daher $\|x_+ - x^*\|_* \leq (1 + \mu\epsilon)(\sigma + \mu\epsilon) \|x - x^*\|_* \leq q \|x - x^*\|_*$. Hieraus folgt die Durchführbarkeit und die lineare Konvergenz der Folge $\{x_k\}$ für jedes Startelement $x_0 \in B_*(x^*; \delta)$. Daß unter den zusätzlichen Voraussetzungen (wegen $F(x^*) = 0$ kann $\sigma = 0$ gewählt werden) quadratische Konvergenz vorliegt, erkennt man durch Inspektion. Siehe auch J. E. DENNIS, R. B. SCHNABEL (1983, S. 222 ff.) und für eine „inexakte“ Version, bei der die Normalgleichung $F'(x_k)^T [F(x_k) + F'(x_k)p_k] = 0$ nur näherungsweise erfüllt zu sein braucht, J. E. DENNIS, T. STEIHAUG (1986).

11. Sei das nichtlineare Gleichungssystem $F(x) = 0$ mit einer stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ gegeben. Man programmiere das gedämpfte Newton-Verfahren $x_{k+1} = x_k + t_k p_k$ mit der Newton-Richtung $p_k := -F'(x_k)^{-1} F(x_k)$, wobei t_k die Armijo-Schrittweite ist. Diese wird z. B. durch den folgenden Algorithmus bestimmt:

- Sei $\alpha \in (0, \frac{1}{2})$ gegeben, setze $\rho_{k,0} := 1$.
- Für $j = 0, 1, \dots$:
 - Falls $\|F(x_k + \rho_{k,j} p_k)\|_\infty \leq (1 - \alpha \rho_{k,j}) \|F(x_k)\|_\infty$, dann: $t_k := \rho_{k,j}$, STOP.
 - Andernfalls: Berechne

$$\rho_{k,j}^* := \frac{0.5 \rho_{k,j}^2 \|F(x_k)\|_\infty}{\|F(x_k + \rho_{k,j} p_k)\|_\infty - (1 - \rho_{k,j}) \|F(x_k)\|_\infty},$$

$$\rho_{k,j+1} := \max(0.1 \rho_{k,j}, \rho_{k,j}^*).$$

Anschließend teste man das Programm an den nichtlinearen Gleichungssystemen

$$F(x) := \begin{pmatrix} x_1^2 + x_2^2 - 2 \\ \exp(x_1 - 1) + x_2^3 - 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

und

$$F(x) := \begin{pmatrix} \exp(x_2) \sin x_1 + 3x_1 x_2 + 3\pi x_2^2 \cos x_1 \\ x_2^3 \cos x_1 + x_1 x_2^3 + x_2 - \pi \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hinweis: Nimmt man im ersten Beispiel wie J. E. DENNIS, R. B. SCHNABEL (1983, S. 149 ff.) als Startwert $x_0 := (2, 0.5)^T$ und wählt $\alpha := 0.0001$, so wird man etwa die in der Tabelle 7.1 angegebenen Ergebnisse erhalten. Hieran erkennt man, daß in den

i	y_i
1	5.308
2	7.240
3	9.638
4	12.866
5	17.069
6	23.192
7	31.443
8	38.558
9	50.156
10	62.948
11	75.995
12	91.972

k	x_k	$\ F(x_k)\ _\infty$	t_k
0	2.0000000000000	0.5000000000000	2.2500000000000
1	1.970033236872	0.597367052513	2.237878349810
2	1.840722959221	0.836502483148	2.087997416918
3	1.094807773337	1.229836043369	0.959570440917
4	0.974993835849	1.047391279160	0.124321975246
5	0.998684479869	1.002595252477	0.006491325705
6	0.999995823913	1.000008381805	0.000020969548
7	0.999999999956	1.000000000087	0.0000000000218
8	1.0000000000000	1.0000000000000	0.0000000000000

Tabelle 7.1: Ergebnisse zu Aufgabe 11, Daten zu Aufgabe 13

ersten beiden Schritten kleinere Schrittweiten als 1 gewählt werden. Man sollte sich davon überzeugen, daß das ungedämpfte Newton-Verfahren mit demselben Startwert x_0 nicht konvergiert. Startet man dagegen mit $x_0 = (-16, -16)^T$, so konvergiert das gedämpfte Newton-Verfahren gegen eine weitere Nullstelle $x^* \approx (-0.7137, 1.2209)^T$, während das ungedämpfte Newton-Verfahren gegen die Nullstelle $(1, 1)$ konvergiert. Im zweiten Beispiel starte man bei dem gedämpften Newton-Verfahren mit dem (schlechten) Startwert $x_0 = (10, 5)^T$ und vergleiche die erhaltenen Werte mit denen, die das ungedämpfte Newton-Verfahren liefert.

12. Sei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ mit $m \geq n$ eine stetig differenzierbare Abbildung, $f(x) := \|F(x)\|_2$ und $h(x) := \frac{1}{2} \|F(x)\|_2^2$. Bei gegebenem $x \in \mathbb{R}^n$ sei $p^* \in \mathbb{R}^n$ eine Lösung des in x linearisierten Ausgleichsproblems

$$(LP_x) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|_2, \quad p \in \mathbb{R}^n.$$

Man zeige:

- (a) Es ist $\nabla h(x)^T p^* = f_x(p^*)^2 - f(x)^2$.

Hinweis: Es gilt die Normalgleichung $F'(x)^T [F(x) + F'(x)p^*] = 0$.

- (b) Ist $f_x(p^*) < f(x)$, ist also p^* eine Abstiegsrichtung (für f und h) in x , und sind $\alpha, \rho \in (0, 1)$, so gilt

$$h(x + \rho p^*) \leq h(x) + \alpha \rho \nabla h(x)^T p^* \implies f(x + \rho p^*) \leq f(x) + \alpha \rho [f_x(p^*) - f(x)].$$

13. Sei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ mit $m \geq n$ eine stetig differenzierbare Abbildung. Man programmiere das gedämpfte Gauß-Newton-Verfahren zur Lösung der nichtlinearen Ausgleichsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n.$$

Anschließend teste man das Programm an der Abbildung $F = (F_i): \mathbb{R}^3 \rightarrow \mathbb{R}^{12}$, die durch

$$F_i(x) := \frac{x_1}{1 + x_2 \exp(ix_3)} - y_i, \quad i = 1, \dots, 12,$$

gegeben ist, wobei die Zahlen y_1, \dots, y_{12} der Tabelle 7.1 zu entnehmen sind (siehe J. C. NASH (1979, S. 120)). Es handelt sich hierbei darum, die Parameter bei einem sogenannten logistischen Wachstumsmodell den beobachteten Daten anzupassen.

Hinweis: Das in jedem Iterationsschritt auftretende lineare Ausgleichsproblem

$$(LP_k) \quad \text{Minimiere } f_k(p) := \|F(x_k) + F'(x_k)p\|_2, \quad p \in \mathbb{R}^n$$

mit der (bei vollem Rang von $F'(x_k)$) eindeutigen Lösung p_k sollte, wie mehrfach betont, mit Hilfe einer QR-Zerlegung von $F'(x_k)$ gelöst werden. Hierbei erhält man fast automatisch den Wert $f_k(p_k)$ mit, den man bei der Armijo-Schrittweite sowie bei der Überprüfung einer Abbruchbedingung benötigt.

Mit dem Startwert $x_0 := (200, 30, -0.4)^T$ wie bei J. C. NASH (1979, S. 178) wird man bei Anwendung der Armijo-Schrittweite mit $\alpha := 0.0001$ (die aber von Anfang an gleich 1 ist) etwa die in der Tabelle 7.2 angegebenen Ergebnisse erhalten.

14. Die Abbildung $F = (F_i): \mathbb{R}^5 \rightarrow \mathbb{R}^9$ sei gegeben durch

$$F_i(x) := x_1 + x_2 \exp(x_3 t_i) + x_4 \exp(x_5 t_i) - y_i, \quad i = 1, \dots, 9,$$

wobei die (Zeiten) t_1, \dots, t_9 und die (beobachteten Werte) y_1, \dots, y_9 in Tabelle 7.3 zu finden sind. Man löse das zugehörige nichtlineare Ausgleichsproblem mit dem gedämpften Gauß-Newton-Verfahren.

Hinweis: Nimmt man wie bei H. R. SCHWARZ (1988, S. 318) den Startwert

$$x_0 := (1.75, 1.20, -0.5, 0.8, -2.0)^T,$$

so wird bei der Armijo-Schrittweite von Anfang an die Schrittweite 1 akzeptiert. Bricht man das Verfahren ab, sobald $\|F(x_k)\|_2 - \|F(x_k) + F'(x_k)p_k\|_2 \leq 10^{-12}$, so erhält man etwa die in Tabelle 7.4 angegebenen Ergebnisse.

k	x_k			$\ F(x_k)\ _2$	$\ F(x_k) + F'(x_k)p_k\ _2$
0	200.00000000	30.00000000	-0.40000000	153.57848266	2.73846102
1	141.80746504	31.75257702	-0.34448830	18.10038874	1.80497564
2	171.20291007	40.80614279	-0.31029874	7.20040886	1.61267736
3	195.25942267	48.49540278	-0.31299184	1.62873785	1.60849667
4	196.16144824	49.08592601	-0.31358303	1.60851152	1.60850161
5	196.18593259	49.09159233	-0.31356989	1.60850160	1.60850160
6	196.18625897	49.09163902	-0.31356973	1.60850160	1.60850160
7	196.18626175	49.09163945	-0.31356973	1.60850160	1.60850160

Tabelle 7.2: Numerische Ergebnisse zu Aufgabe 13

t_i	0.0	0.5	1.0	1.5	2.0	3.0	5.0	8.0	10.0
y_i	3.85	2.95	2.63	2.33	2.24	2.05	1.82	1.80	1.75

Tabelle 7.3: Daten zu Aufgabe 14

15. Man programmiere das gedämpfte Gauß-Newton-Verfahren zur Lösung der diskreten, nichtlinearen Tschebyscheffschen Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^n$$

mit der stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$. Zur Berechnung einer Lösung p_k der in jedem Iterationsschritt auftretenden diskreten, linearen Tschebyscheffschen Approximationsaufgabe

$$(LP_k) \quad \text{Minimiere } f_k(p) := \|F(x_k) + F'(x_k)p\|_\infty, \quad p \in \mathbb{R}^n$$

kann das in Abschnitt 6.3 geschilderte Verfahren benutzt werden. Ferner verwende man die Armijo-Schrittweite. Anschließend teste man das Programm an dem folgenden Beispiel (siehe K. MADSEN (1975a)). Man setze $t_i := (i - 11)/10$, $i = 1, \dots, 21$, und

k	x_k					$\ F(x_k)\ _2$
0	1.750000	1.200000	-0.500000	0.800000	-2.000000	0.131151
1	1.761110	1.547618	-0.589502	0.539783	-3.088365	0.118795
2	1.757621	1.408975	-0.552848	0.682683	-3.349923	0.077353
3	1.757684	1.420068	-0.554883	0.671653	-3.375225	0.077097
4	1.757725	1.420782	-0.555157	0.670909	-3.381627	0.077097
5	1.757736	1.420961	-0.555228	0.670721	-3.383124	0.077097
6	1.757739	1.421003	-0.555245	0.670677	-3.383474	0.077097
7	1.757739	1.421013	-0.555249	0.670667	-3.383555	0.077097
8	1.757739	1.421016	-0.555250	0.670665	-3.383574	0.077097

Tabelle 7.4: Ergebnisse zu Aufgabe 14

definiere die Abbildung $F: \mathbb{R}^5 \rightarrow \mathbb{R}^{21}$ durch

$$F_i(x_1, x_2, x_3, x_4, x_5) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2 + x_5 t_i^3} - \exp(t_i), \quad i = 1, \dots, 21.$$

Hinweis: Wird das gedämpfte Gauß-Newton-Verfahren mit $(0.5, 0.5, 0.5, 0.5, 0.5)$ gestartet, verwendet man ferner $\alpha := 0.0001$ in der Armijo-Schrittweite, so erhält man die in Tabelle 7.5 angegebenen Werte. In Abbildung 7.2 ist der Defekt auf dem Inter-

k	x_k					$\ F(x_k)\ _\infty$	t_k
0	0.500000	0.500000	0.500000	0.500000	0.500000	2.318282	0.100000
1	0.541074	0.515928	0.454031	0.036569	-0.212615	1.891197	0.100000
2	0.587661	0.536593	0.302364	-0.060281	-0.177824	1.661910	0.100000
3	0.629268	0.553918	0.189404	-0.112187	-0.126064	1.474334	0.157024
4	0.687734	0.576090	0.054645	-0.159485	-0.057320	1.209851	1.000000
5	1.000519	0.667103	-0.535678	-0.262931	0.230042	1.147024	1.000000
6	0.996555	0.557926	-0.438390	-0.121810	0.089987	0.215880	1.000000
7	1.000431	0.310042	-0.685788	0.188742	-0.017155	0.020716	1.000000
8	0.999909	0.255847	-0.744253	0.242928	-0.036907	0.001174	1.000000
9	0.999878	0.253594	-0.746602	0.245196	-0.037488	0.000123	1.000000
10	0.999878	0.253588	-0.746608	0.245202	-0.037490	0.000122	1.000000
11	0.999878	0.253588	-0.746608	0.245202	-0.037490	0.000122	

Tabelle 7.5: Ergebnisse zum Beispiel in Aufgabe 15

vall $[-1, 1]$ skizziert. Um nicht etwa in den Irrtum zu verfallen, das gedämpfte Gauß-

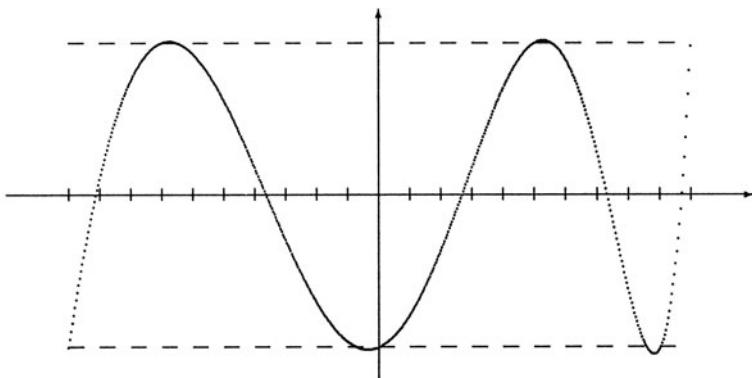


Abbildung 7.2: Der Defekt $d(t) := \frac{x_1^* + x_2^* t}{1 + x_3^* t + x_4^* t^2 + x_5^* t^3} - \exp(t)$

Newton-Verfahren würde immer so gut wie im obigen Beispiel konvergieren, wende

man es auch noch auf die durch

$$F(x_1, x_2) := \begin{pmatrix} x_1^2 + x_2^2 + x_1 x_2 \\ \sin x_1 \\ \cos x_2 \end{pmatrix}$$

gegebene Abbildung $F: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ an (siehe K. MADSEN (1975a) und auch R. GONIN, A. H. MONEY (1989, S. 133 und S. 155 ff.)).

7.3 Quasi-Newton-Verfahren

7.3.1 Das Newton-Verfahren

Wir betrachten die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n,$$

wobei die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ als zweimal stetig differenzierbar vorausgesetzt wird. Stationäre Lösungen von (P) sind Lösungen des i. allg. nichtlinearen Gleichungssystems $\nabla f(x) = 0$. Es liegt nahe, hierauf das Newton-Verfahren anzuwenden, was auf die Iterationsvorschrift

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

führt. Der lokale Konvergenzsatz für das Newton-Verfahren (siehe Satz 3.1 in Abschnitt 2.3) sagt aus:

Satz 3.1 Die Abbildung $f: \mathbb{R}^n \rightarrow \mathbb{R}$ sei auf einer Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Es sei $\nabla f(x^*) = 0$ (also x^* ein stationärer Punkt von f bzw. eine stationäre Lösung von (P)) und $\nabla^2 f(x^*)$ nichtsingulär. $\|\cdot\|$ bezeichne eine beliebige Norm im \mathbb{R}^n bzw. die zugeordnete Matrixnorm. Dann existiert ein $\delta > 0$ derart, daß für jedes $x_0 \in B[x^*; \delta] := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$ die durch das Newton-Verfahren

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k), \quad k = 0, 1, \dots,$$

gewonnene Folge $\{x_k\}$ definiert ist (d. h. $\nabla^2 f(x_k)$ existiert und ist nichtsingulär für $k = 0, 1, \dots$) und superlinear gegen x^* konvergiert, d. h. es gilt

$$\lim_{k \rightarrow \infty} x_k = x^*, \quad \lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Ist zusätzlich $\nabla^2 f(\cdot)$ auf einer (hinreichend kleinen) Kugel um x^* in x^* lipschitzstetig, d. h. existieren $\eta > 0$ und $L > 0$ mit

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq L \|x - x^*\| \quad \text{falls } \|x - x^*\| \leq \eta,$$

so konvergiert $\{x_k\}$ bei hinreichend kleinem $\delta > 0$ für jedes $x_0 \in B[x^*; \delta]$ sogar von mindestens zweiter Ordnung gegen x^* , d. h. es existiert eine Konstante $C > 0$ mit $\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$ für alle hinreichend großen $k \in \mathbb{N}$.

Durch die Einführung von Schrittweiten, also den Übergang zum gedämpften Newton-Verfahren

$$x_{k+1} := x_k - t_k \nabla^2 f(x_k)^{-1} \nabla f(x_k),$$

kann man versuchen, zu einem global konvergenten Verfahren zu kommen. Unter geeigneten Konvexitätsvoraussetzungen, die u. a. sichern, daß $\nabla^2 f(x_k)$ positiv definit und damit die Newton-Richtung $p_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ eine Abstiegsrichtung ist, wird man die Konvergenz des gedämpften Newton-Verfahrens erwarten. Ferner wird man die Schrittweitenstrategie so gestalten wollen, daß nach endlich vielen Schritten, wenn die durch das gedämpfte Newton-Verfahren erzeugten Näherungen erst einmal hinreichend nahe bei einer Lösung liegen, automatisch vom gedämpften zum ungedämpften Newton-Verfahren übergegangen wird. Diese Erwartungen werden durch den folgenden globalen Konvergenzsatz für das Newton-Verfahren bestätigt.

Satz 3.2 Gegeben sei die unrestringierte Optimierungsaufgabe (P). Über die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ wird vorausgesetzt:

- (a) Mit einem $x_0 \in \mathbb{R}^n$ ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.
- (b) f ist auf einer Umgebung von L_0 zweimal stetig differenzierbar und es existieren positive Konstanten $c \leq \gamma$ mit

$$(*) \quad c \|p\|_2^2 \leq p^T \nabla^2 f(x)p \leq \gamma \|p\|_2^2 \quad \text{für alle } x \in L_0, p \in \mathbb{R}^n.$$

Zur Bestimmung der unter diesen Voraussetzungen eindeutig existierenden (globalen) Lösung x^* von (P) betrachte man das gedämpfte Newton-Verfahren

$$x_{k+1} := x_k + t_k p_k \quad \text{mit } p_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k),$$

wobei t_k in jedem Schritt die Armijo-Schrittweite (mit $\alpha \in (0, \frac{1}{2})$) sei. Dann gilt: Bricht das Verfahren nicht vorzeitig mit der Lösung x^* von (P) ab, so erzeugt es eine gegen x^* konvergente Folge $\{x_k\}$. Ferner ist $t_k = 1$ für alle hinreichend großen k , nach endlich vielen Schritten geht das gedämpfte Newton-Verfahren also in das ungedämpfte über.

Beweis: Aus $c \|p\|_2^2 \leq p^T \nabla^2 f(x)p$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ folgt wegen Satz 1.15, daß die Zielfunktion f auf der nach Voraussetzung konvexen Niveaumenge L_0 gleichmäßig konvex ist. Wegen $p^T \nabla^2 f(x)p \leq \gamma \|p\|_2^2$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ ist $\|\nabla^2 f(x)\|_2 \leq \gamma$ für alle $x \in L_0$. Hieraus folgt die Lipschitzstetigkeit von $\nabla f(\cdot)$ auf L_0 mit der Lipschitzkonstanten $\gamma > 0$ (bezüglich der euklidischen Norm). Denn für beliebige $x, y \in L_0$ ist

$$\|\nabla f(x) - \nabla f(y)\|_2 = \left\| \int_0^1 \nabla^2 f(x + s(y-x)) \underbrace{(x-y)}_{\in L_0} ds \right\|_2 \leq \gamma \|x-y\|_2.$$

Zum Nachweis der Konvergenz der Folge $\{x_k\}$ wollen wir Satz 2.7 anwenden. Durch (a) und (b) sind die Voraussetzungen (K) (a)–(c) aus Lemma 2.6 erfüllt, wie wir uns gerade eben überlegt haben. Die Bedingung (*) in Voraussetzung (b) impliziert

$$\delta_k := \min \left[-\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2^2}, \left(\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2 \|p_k\|_2} \right)^2 \right] \geq \min \left[\frac{1}{\gamma}, \frac{c}{\gamma} \right] =: \delta,$$

wie man unschwer nachweist (siehe auch eine Bemerkung im Anschluß an Satz 2.5). Insbesondere ist $\delta(k+1) \leq \sum_{j=0}^k \delta_j$ für $k = 0, 1, \dots$. Wegen Satz 2.7 konvergiert die Folge $\{x_k\}$ gegen die eindeutige (globale) Lösung x^* von (P). Um $t_k = 1$ für alle hinreichend großen k nachzuweisen, müssen wir zeigen, daß

$$f(x_k + p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k \quad \text{für alle hinreichend großen } k.$$

Wegen $\|p_k\|_2 \leq \|\nabla f(x_k)\|_2/c$ konvergiert die Folge $\{p_k\}$ der Newton-Richtungen gegen den Nullvektor. Da außerdem o. B. d. A. x^* im Innern der Niveaumenge L_0 liegt und $\{x_k\}$ gegen x^* konvergiert, liegt die gesamte Verbindungsstrecke zwischen x_k und $x_k + p_k$ für alle hinreichend großen k in L_0 . Mit einem $\theta_k \in (0, 1)$ ist daher für diese k wegen des Mittelwertsatzes

$$\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} - \frac{p_k^T [\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)] p_k}{2 p_k^T \nabla^2 f(x_k) p_k}.$$

Auf der rechten Seite dieser Gleichung konvergiert der zweite Summand gegen Null, denn wegen $x_k + \theta_k p_k \rightarrow x^*$ und $x_k \rightarrow x^*$ ist

$$\frac{|p_k^T [\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)] p_k|}{p_k^T \nabla^2 f(x_k) p_k} \leq \frac{1}{c} \|\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)\|_2 \rightarrow 0.$$

Damit ist

$$\lim_{k \rightarrow \infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha$$

bewiesen, woraus die Behauptung folgt. \square

Es gibt einige Einwände gegen das Newton-Verfahren. Der erste ist theoretischer Art und besteht darin, daß die Anwendung des Newton-Verfahrens fern einer stationären Lösung von (P), also etwa mit einem schlechten Startwert, keinen Sinn macht. Denn hier „zieht“ sozusagen die dem Newton-Verfahren zugrunde liegende Motivation nicht. Denn diese besteht ja darin, das zu lösende nichtlineare Gleichungssystem $\nabla f(x) = 0$ in einer aktuellen Näherung x_k zu linearisieren und die Lösung dieses linearen Gleichungssystems $\nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$ als neue Näherung zu nehmen (oder äquivalent dazu, die Zielfunktion f in x_k zu „quadratisieren“ und einen stationären Punkt der quadratischen Funktion

$$f_k(x) := f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k) (x - x_k)$$

als neue Näherung zu berechnen). Auch wenn in dem zu berechnenden Punkt x^* die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist, also $\nabla f(x^*) = 0$ gilt und $\nabla^2 f(x^*)$ positiv definit ist, wird die Hessesche $\nabla^2 f(x_k)$ für von x^* weit entfernte x_k i. allg. nicht positiv definit sein und damit die Newton-Richtung nicht unbedingt eine Abstiegsrichtung sein. Der zweite Einwand ist praktischer Art. Das Newton-Verfahren verlangt die Berechnung der Hesseschen der Zielfunktion in der aktuellen Näherung, also der zweiten partiellen Ableitungen. Dies ist bei vielen Anwendungen, bei denen schon die Berechnung des Gradienten der Zielfunktion Mühe

bereitet und durch die Bildung von Differenzenquotienten ersetzt werden muß, unzumutbar. Weiter muß beim Newton-Verfahren in jedem Schritt ein lineares Gleichungssystem mit der (symmetrischen) Koeffizientenmatrix $\nabla^2 f(x_k)$ gelöst werden. Die Anzahl der hierzu nötigen arithmetischen Operationen ist im wesentlichen proportional zu n^3 (bei positiv defitem $\nabla^2 f(x_k)$ bietet sich natürlich die Anwendung des Cholesky-Verfahrens an). Ferner wird man kaum hoffen können, Kenntnisse über eine Zerlegung (Cholesky-, *LR*- oder *QR*-Zerlegung) von $\nabla^2 f(x_k)$ nutzbringend auf die Berechnung einer entsprechenden Zerlegung von $\nabla^2 f(x_{k+1})$ anwenden zu können.

7.3.2 Die Broyden-Klasse und das BFGS-Verfahren

Die *Quasi-Newton-Verfahren* versuchen, die Nachteile des Newton-Verfahrens zu vermeiden, ohne die Vorteile (globale Konvergenz durch Einführung von Schrittweiten und automatischer Übergang zum ungedämpften Verfahren bei gleichmäßig konvexer Zielfunktion, lokal superlineare Konvergenz des ungedämpften Verfahrens) aufzugeben. Insbesondere das zu dieser Klasse gehörende *BFGS-Verfahren* (BFGS steht für Broyden-Fletcher-Goldfarb-Shanno, die dieses Verfahren unabhängig voneinander 1970 entdeckten) gilt für glatte, nicht zu hochdimensionale, unrestringierte Optimierungsaufgaben, bei denen neben den Zielfunktionswerten auch der Gradient zur Verfügung steht, als das anerkanntermaßen beste Minimierungsverfahren. In den Grundzügen sehen die zu betrachtenden Quasi-Newton-Verfahren folgendermaßen aus (man vergleiche mit dem Modellalgorithmus zu Beginn von 7.2):

- Gegeben $x_0 \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B_0 \in \mathbb{R}^{n \times n}$.
Ferner sei $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \dots$

Test auf Abbruch: Falls $g_k = 0$, dann: STOP.

Berechne Abstiegsrichtung $p_k := -B_k^{-1}g_k$.

Berechne Schrittweite t_k , etwa die exakte Schrittweite, die Powell- oder die Armijo-Schrittweite.

Berechne neue Näherung $x_{k+1} := x_k + t_k p_k$ und $g_{k+1} := \nabla f(x_{k+1})$.

Berechne symmetrische, positiv definite Matrix $B_{k+1} \in \mathbb{R}^{n \times n}$ durch eine sogenannte *Update-Formel*. In die Berechnung von B_{k+1} gehen i. allg. B_k sowie $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ ein.

Ein Quasi-Newton-Verfahren ist also (neben der Wahl der Schrittweitenstrategie) durch die Update-Formel festgelegt. Um Schreibarbeit zu sparen, nehmen wir nun an, $x := x_k$ sei eine aktuelle Näherung mit $\nabla f(x) \neq 0$, $B := B_k$ sei symmetrisch und positiv definit und damit $p := -B^{-1}\nabla f(x)$ eine Abstiegsrichtung, und mit einer geeigneten Schrittweite $t := t_k$ sei die neue Näherung $x_+ := x + tp$ berechnet. Ferner sei $s := x_+ - x$ und $y := \nabla f(x_+) - \nabla f(x)$. Wie sollte nun die neue, symmetrische und positiv definite Matrix $B_+ := B_{k+1}$ berechnet werden? Hierauf gibt es keine eindeutige Antwort (sonst wären in den letzten dreißig Jahren nicht so viele

Update-Formeln entwickelt worden). Neben der Symmetrie und positiven Definitheit sollte B_+ der sogenannten *Quasi-Newton-Gleichung* $B_+s = y$ (gelegentlich auch *Sekantengleichung* genannt) genügen. Wir geben zwei Motivationen hierfür an. Ist die Zielfunktion quadratisch und gleichmäßig konvex, so ist $\nabla^2 f(\cdot)$ konstant und positiv definit, ferner ist $\nabla^2 f(\cdot)(x_+ - x) = \nabla f(x_+) - \nabla f(x)$. Daher liegt es nahe, von B_+ das Erfülltsein der Quasi-Newton-Gleichung $B_+s = y$ zu erwarten. Ist andererseits diese Gleichung erfüllt und

$$f_+(z) := f(x_+) + \nabla f(x_+)^T(z - x_+) + \frac{1}{2}(z - x_+)^T B_+(z - x_+)$$

eine quadratische Approximation an $f(z)$, so ist $f_+(x_+) = f(x_+)$, $\nabla f_+(x) = \nabla f(x)$ und $\nabla f_+(x_+) = \nabla f(x_+)$, was ein Indiz dafür ist, daß f_+ in der Nähe von x_+ eine gute Approximation an f sein könnte.

Eine *notwendige* Bedingung für die Existenz einer symmetrischen, positiv definiten Matrix B_+ mit $B_+s = y$ ist offenbar $y^T s > 0$. Diese Bedingung ist z. B. unter den Voraussetzungen (K) (a)–(c) aus Lemma 2.6 erfüllt, wenn also insbesondere die Zielfunktion f auf der konvexen Niveaumenge L_0 gleichmäßig konvex (mit einer Konstanten $c > 0$) ist. Denn dann ist

$$y^T s = [\nabla f(x_+) - \nabla f(x)]^T(x_+ - x) \geq c\|x_+ - x\|_2^2 = c\|s\|_2^2 > 0.$$

Aber auch ohne die gleichmäßige Konvexität von f ist i. allg. $y^T s > 0$. Denn ist z. B. $t > 0$ eine Powell-Schrittweite, so ist

$$\nabla f(x_+)^T p \geq \beta \nabla f(x)^T p > \nabla f(x)^T p$$

mit vorgegebenem $\beta \in (0, 1)$ und daher $y^T s = t y^T p > 0$. Ist ferner t die exakte Schrittweite, so ist $\nabla f(x_+)^T p = 0$ und daher $y^T s = -t \nabla f(x)^T p > 0$.

Eine *Update-Formel der Broyden-Klasse* ist bei gegebenem $\phi \geq 0$ durch

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s} + \phi(s^T Bs)vv^T \quad \text{mit } v := \frac{y}{y^T s} - \frac{Bs}{s^T Bs}$$

definiert. Die prominentesten Vertreter dieser Broyden-Klasse ergeben sich für $\phi = 0$ (BFGS-Update-Formel) und $\phi = 1$ (DFP-Update-Formel, wobei DFP für Davidon-Fletcher-Powell steht, die 1959 bzw. 1963 diese Formel angaben und das zugehörige Verfahren untersuchten, das zu der Zeit einen wesentlichen Fortschritt gegenüber dem vorher noch ziemlich konkurrenzlosen Gradientenverfahren bedeutete).

In dem folgenden Satz wird unter der Voraussetzung $y^T s > 0$ u. a. gezeigt, daß die Formeln der Broyden-Klasse der Quasi-Newton-Gleichung genügen und mit B auch B_+ symmetrisch und positiv definit ist.

Satz 3.3 Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ sowie eine symmetrische, positiv definite Matrix $B \in \mathbb{R}^{n \times n}$ vorgegeben. Sei $\phi \geq 0$ und

$$B_+ := B_{\text{BFGS}} + \phi(s^T Bs)vv^T$$

mit

$$B_{\text{BFGS}} := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s}, \quad v := \frac{y}{y^T s} - \frac{Bs}{s^T Bs}.$$

Dann gilt:

1. Die Quasi-Newton-Gleichung $B_+ s = y$ ist erfüllt.
2. Die Matrix B_+ ist symmetrisch und positiv definit.
3. Es ist

$$\det(B_{\text{BFGS}}) = \frac{y^T s}{s^T B s} \det(B)$$

und

$$\begin{aligned} B_{\text{BFGS}}^{-1} &= B^{-1} + \left(1 + \frac{y^T B^{-1} y}{y^T s}\right) \frac{s s^T}{y^T s} - \frac{s (B^{-1} y)^T + (B^{-1} y) s^T}{y^T s} \\ &= \left(I - \frac{s y^T}{y^T s}\right) B^{-1} \left(I - \frac{y s^T}{y^T s}\right) + \frac{s s^T}{y^T s}. \end{aligned}$$

Beweis: Die Gültigkeit der Quasi-Newton-Gleichung erkennt man durch Inspektion. Offenbar ist mit B auch B_+ symmetrisch. Wegen $\phi \geq 0$ ist in der Update-Formel für B_+ der Summand $\phi(s^T B s) v v^T$ positiv semidefinit. Daher zeigen wir für den zweiten Teil des Beweises, daß B_{BFGS} positiv definit ist.

Wie in der schönen Darstellung bei J. E. DENNIS, R. B. SCHNABEL (1983, S. 198 ff.) soll der Beweis so erfolgen, daß wir in einer späteren Bemerkung begründen können, weshalb man verhältnismäßig einfach (nämlich mit einer im wesentlichen zu n^2 proportionalen Anzahl an arithmetischen Operationen) aus einer Cholesky-Zerlegung von B eine solche von B_{BFGS} berechnen kann. Da $B \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist, besitzt B eine Cholesky-Zerlegung $B = LL^T$ mit einer unteren Dreiecksmatrix L , deren Diagonalelemente positiv sind. Wir zeigen, daß $B_{\text{BFGS}} = J_+ J_+^T$ mit einer nichtsingulären Matrix J_+ , woraus die zweite Behauptung folgt. Hierzu definiere man

$$w := \left(\frac{y^T s}{s^T B s}\right)^{1/2} L^T s, \quad J_+ := L + \frac{(y - Lw) w^T}{w^T w}.$$

Wegen

$$\sigma := 1 + \frac{w^T (L^{-1} y - w)}{w^T w} = \frac{w^T L^{-1} y}{w^T w} = \left(\frac{y^T s}{s^T B s}\right)^{1/2} \neq 0$$

ist J_+ nach der Sherman-Morrison-Formel (siehe Lemma 2.14 in Abschnitt 1.2) nicht-singulär, $\det(J_+) = \sigma \det(L)$ und

$$J_+^{-1} = L^{-1} - \frac{(L^{-1} y - w) w^T L^{-1}}{\sigma w^T w}.$$

Ferner bestätigt man nach leichter Rechnung, daß $J_+ J_+^T = B_{\text{BFGS}}$, womit der zweite Teil des Satzes bewiesen ist.

Es ist

$$\det(B_{\text{BFGS}}) = \det(J_+)^2 = \sigma^2 \det(L)^2 = \frac{y^T s}{s^T B s} \det(B).$$

Aus $B_{\text{BFGS}}^{-1} = J_+^{-T} J_+^{-1}$ erhält man durch Einsetzen und Umformen leicht einen Beweis der restlichen Behauptungen. \square

Bemerkung: Durch eine erneute Anwendung der Sherman-Morrison-Formel erhält man aus $B_+ = J_+[I + \phi s^T B s (J_+^{-1} v)(J_+^{-1} v)^T] J_+^T$, daß

$$\begin{aligned} B_+^{-1} &= J_+^{-T} \left[I - \frac{\phi s^T B s}{1 + \phi(s^T B s) \|J_+^{-1} v\|_2^2} (J_+^{-1} v)(J_+^{-1} v)^T \right] J_+^{-1} \\ &= B_{\text{BFGS}}^{-1} - \frac{\phi s^T B s}{1 + \phi(s^T B s)(v^T B_{\text{BFGS}}^{-1} v)} (B_{\text{BFGS}}^{-1} v)(B_{\text{BFGS}}^{-1} v)^T. \end{aligned}$$

Hierbei ist

$$B_{\text{BFGS}}^{-1} v = \frac{1}{y^T s} \left(B^{-1} y - \frac{y^T B^{-1} y}{y^T s} s \right), \quad v^T B_{\text{BFGS}}^{-1} v = \frac{1}{y^T s} \left(\frac{y^T B^{-1} y}{y^T s} - \frac{y^T s}{s^T B s} \right).$$

Daher erhält man z. B. für $\phi = 1$, also die DFP-Update-Formel, daß

$$B_{\text{DFP}}^{-1} = B^{-1} - \frac{(B^{-1} y)(B^{-1} y)^T}{y^T B^{-1} y} + \frac{s s^T}{y^T s}.$$

□

Ausgehend von denselben Startelementen x_0 und B_0 erzeugen alle Verfahren der Broyden-Klasse identische Folgen $\{x_k\}$, wenn in jedem Schritt die exakte Schrittweite $t_k = t^*(x_k, p_k)$ gewählt wird (also die erste positive Nullstelle von $\nabla f(x_k + t p_k)^T p_k$). Dieses bemerkenswerte Ergebnis stammt von L. C. W. DIXON (1972). Im folgenden Satz findet man eine genaue Formulierung. Der Beweis folgt dem bei R. FLETCHER (1987, S. 66).

Satz 3.4 Über die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ der unrestringierten Optimierungsaufgabe (P) wird vorausgesetzt, daß die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ mit einem vorgegebenem $x_0 \in \mathbb{R}^n$ kompakt und f auf einer Umgebung von L_0 stetig differenzierbar ist. Sei $B_0 \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Man betrachte das Quasi-Newton-Verfahren der Broyden-Klasse mit exakter Schrittweite:

- Sei $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: STOP.

Berechne $p_k := -B_k^{-1} g_k$, die exakte Schrittweite $t_k := t^*(x_k, p_k)$ und setze $x_{k+1} := x_k + t_k p_k$. Berechne $g_{k+1} := \nabla f(x_{k+1})$ sowie $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$.

Wähle $\phi_k \geq 0$ und berechne

$$B_{k+1} := F_{k+1} + \phi_k (s_k^T B_k s_k) v_k v_k^T,$$

wobei

$$F_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \quad v_k := \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}.$$

Dann sind die Folgen $\{x_k\}$ und $\{F_k\}$ von der Wahl der Parameter $\{\phi_k\}$ unabhängig.

Beweis: Für $k \in \mathbb{N}$ geht die Wahl von ϕ_{k-1} in die Berechnung von B_k und damit in die von x_{k+1} und F_{k+1} ein. Daher ist zu zeigen, daß x_{k+1} und F_{k+1} von der Wahl von $\phi_0, \dots, \phi_{k-1}$ unabhängig sind, was durch vollständige Induktion geschehen wird.

Wegen $B_k s_k = -t_k g_k$ und $g_{k+1}^T s_k = 0$ (denn t_k ist exakte Schrittweite) ist

$$v_k = \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k} = \frac{g_{k+1} - g_k}{y_k^T s_k} + \frac{g_k}{y_k^T s_k} = \frac{g_{k+1}}{y_k^T s_k}.$$

In der Bemerkung im Anschluß an Satz 3.3 haben wir

$$B_{k+1}^{-1} = F_{k+1}^{-1} - \frac{\phi_k s_k^T B_k s_k}{1 + \phi_k (s_k^T B_k s_k) (v_k^T F_{k+1}^{-1} v_k)} (F_{k+1}^{-1} v_k) (F_{k+1}^{-1} v_k)^T$$

nachgewiesen. Da v_k ein Vielfaches von g_{k+1} und $p_{k+1} = -B_{k+1}^{-1} g_{k+1}$ ist, folgt

$$p_{k+1} = -\delta_{k+1} F_{k+1}^{-1} g_{k+1}, \quad B_{k+1}^{-1} = F_{k+1}^{-1} - \theta_{k+1} s_{k+1} s_{k+1}^T, \quad k = 0, 1, \dots$$

mit (von ϕ_k abhängigen) Konstanten $\delta_{k+1} > 0$ und $\theta_{k+1} \geq 0$. Wegen

$$\begin{aligned} F_{k+2}^{-1} &= \left(I - \frac{s_{k+1} y_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) B_{k+1}^{-1} \left(I - \frac{y_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) + \frac{s_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} \\ &= \left(I - \frac{s_{k+1} y_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) (F_{k+1}^{-1} - \theta_{k+1} s_{k+1} s_{k+1}^T) \left(I - \frac{y_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) + \frac{s_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} \end{aligned}$$

ist

$$(**) \quad F_{k+2}^{-1} = \left(I - \frac{s_{k+1} y_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) F_{k+1}^{-1} \left(I - \frac{y_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} \right) + \frac{s_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}}, \quad k = 0, 1, \dots$$

Nun kommen wir zum Induktionsbeweis. Durch x_0 und B_0 sind x_1 und F_1 festgelegt. Da p_1 wegen (*) ein Vielfaches des von ϕ_0 unabhängigen Vektors $F_1^{-1} g_1$ ist, ist auch x_2 von ϕ_0 unabhängig. Das gilt dann auch für s_1 und y_1 , wegen (**) ist F_2 von ϕ_0 unabhängig. Daher ist die Behauptung für $k = 1$ richtig. Der Schluß von k nach $k + 1$ verläuft genauso. Denn sind x_{k+1} und F_{k+1} von $\phi_0, \dots, \phi_{k-1}$ unabhängig, so ist $x_{k+2} = x_{k+1} + t_{k+1} p_{k+1}$ wegen (*) von ϕ_0, \dots, ϕ_k unabhängig, das gilt dann auch für s_{k+1} und y_{k+1} und wegen (**) auch für F_{k+2} . Der Satz ist damit bewiesen. \square

Bemerkung: Die Kompaktheit der Niveaumenge L_0 wurde in Satz 3.4 nur vorausgesetzt, um die Existenz der exakten Schrittweite zu sichern. Satz 3.4 impliziert z. B., daß aus der Konvergenz des BFGS-Verfahrens mit exakter Schrittweite auch die des DFP-Verfahrens mit exakter Schrittweite folgt. Wie wir in Satz 3.6 sehen werden, zeichnet sich das BFGS-Verfahren vor den anderen Quasi-Newton-Verfahren der Broyden-Klasse dadurch aus, daß es bei gleichmäßig konvexer Zielfunktion für viele, auch inexakte Schrittweitenstrategien global konvergent ist, was etwa für das DFP-Verfahren bisher nicht gezeigt werden konnte (siehe z. B. R. H. BYRD, J. NOCEDAL, Y. YUAN (1987)). Hierin scheint eines der Erfolgsgeheimnisse des BFGS-Verfahrens zu liegen. \square

Wendet man ein Quasi-Newton-Verfahren der Broyden-Klasse auf eine gleichmäßig konvexe, quadratische Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ an und benutzt man in jedem Schritt die exakte Schrittweite, so bricht das Verfahren nach $m \leq n$ Schritten mit der Lösung ab. Wegen Satz 3.4 genügt es, dieses Ergebnis für das BFGS-Verfahren zu beweisen.

Satz 3.5 Auf die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := c^T x + \frac{1}{2} x^T Q x, \quad x \in \mathbb{R}^n$$

mit symmetrischer und positiv definiter Matrix $Q \in \mathbb{R}^{n \times n}$ wende man das BFGS-Verfahren mit exakter Schrittweite an:

- Seien $x_0 \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $H_0 \in \mathbb{R}^{n \times n}$ vorgegeben. Berechne $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP.

Berechne

$$p_k := -H_k g_k, \quad t_k := -\frac{g_k^T p_k}{p_k^T Q p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := \nabla f(x_{k+1}).$$

Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ berechne

$$H_{k+1} := \left(I - \frac{s_k y_k^T}{y_k^T s_k} \right) H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k} \right) + \frac{s_k s_k^T}{y_k^T s_k}.$$

Dann gilt:

1. Das Verfahren bricht nach $m \leq n$ Schritten mit der Lösung $x_m = x^*$ von (P) ab.
2. Es ist $p_i^T Q p_k = 0$ für $0 \leq i < k \leq m-1$.
3. Es ist $p_i^T g_k = 0$ und $H_k y_i = s_i$ für $0 \leq i < k \leq m$.
4. Ist $m = n$, so ist $H_n = Q^{-1}$.

Beweis: Die Durchführbarkeit des Verfahrens ist offenbar gesichert. Denn ist $g_k \neq 0$ und $H_k = B_k^{-1}$ symmetrisch und positiv definit, so ist $p_k := -H_k g_k$ eine Abstiegsrichtung für f in x_k , die exakte Schrittweite t_k ist durch $t_k := -g_k^T p_k / p_k^T Q p_k$ gegeben und wegen $y_k^T s_k = t_k g_k^T H_k g_k > 0$ ist $H_{k+1} = B_{k+1}^{-1}$ nach Satz 3.3 symmetrisch und positiv definit. Solange kein Abbruch erfolgt, werden durch das Verfahren also symmetrische, positiv definite Matrizen H_k und Abstiegsrichtungen p_k erzeugt.

Wir zeigen durch vollständige Induktion nach k : Sind $g_0, \dots, g_k \neq 0$, wird das Verfahren also im k -ten Schritt noch nicht abgebrochen, so gilt

$$(a) \quad p_i^T g_k = 0, \quad (b) \quad H_k y_i = s_i, \quad (c) \quad p_i^T Q p_k = 0 \quad \text{für } 0 \leq i < k.$$

Diese drei Aussagen sind für $k = 0$ trivialerweise richtig. Für den Induktionsschluß nehmen wir an, g_0, \dots, g_{k+1} seien von Null verschieden.

Für $0 \leq i < k$ ist $p_i^T g_{k+1} = p_i^T(g_k + t_k Q p_k) = 0$, wobei die Induktionsvoraussetzungen (a) und (b) benutzt wurden. Beachtet man noch, daß $p_k^T g_{k+1} = 0$, da t_k die exakte Schrittweite ist, so erkennt man, daß (a) für $k+1$ gilt.

Für $0 \leq i < k$ ist $s_k^T y_i = t_k p_k^T Q s_i = t_i t_k p_i^T Q p_k = 0$ und ebenso $y_k^T s_i = 0$, wobei die Induktionsvoraussetzung (c) einging. Benutzt man auch noch die Induktionsvoraussetzung (b), so erhält man aus der Update-Formel $H_{k+1} y_i = s_i$. Wegen der Quasi-Newton-Gleichung ist $H_{k+1} y_k = s_k$, womit (b) auch für $k+1$ nachgewiesen ist.

Für $0 \leq i < k+1$ ist

$$p_i^T Q p_{k+1} = \frac{(Q s_i)^T p_{k+1}}{t_i} = \frac{y_i^T p_{k+1}}{t_i} = -\frac{(H_{k+1} y_i)^T g_{k+1}}{t_i} = -\frac{s_i^T g_{k+1}}{t_i} = -\frac{p_i^T g_{k+1}}{t_i} = 0,$$

wobei (b) und (a) für $k+1$ benutzt wurden. Insgesamt ist auch (c) für $k+1$ bewiesen. Der Induktionsbeweis ist abgeschlossen.

Wegen (c) gilt: Sind $g_0, \dots, g_k \neq 0$, so sind die vom Nullvektor verschiedenen Richtungen p_0, \dots, p_k linear unabhängig. Da es im \mathbb{R}^n höchstens n linear unabhängige Vektoren gibt, bricht das Verfahren nach $m \leq n$ Schritten ab. Damit sind die ersten drei Aussagen des Satzes bewiesen.

Ist $m = n$, bricht das Verfahren also nicht vorzeitig ab, so sind die $n \times n$ -Matrizen

$$S := (s_0 \ \cdots \ s_{n-1}), \quad Y := (y_0 \ \cdots \ y_{n-1}) = QS$$

nichtsingulär. Wegen $QS = Y$ und $H_n Y = S$ ist $H_n = Q^{-1}$. Damit ist der Satz schließlich bewiesen. \square

Es muß zugegeben werden, daß die Update-Formeln der Broyden-Klasse, und damit insbesondere auch die BFGS-Update-Formel, bisher „vom Himmel gefallen“ sind. Es gibt einige Möglichkeiten, diesen Ansatz zu begründen, siehe z. B. Aufgabe 2.

Bemerkung: Will man das BFGS-Verfahren implementieren, so muß in jedem Iterationsschritt die Richtung $p := -B^{-1} \nabla f(x)$ berechnet bzw. das Gleichungssystem $Bp = -\nabla f(x)$ gelöst werden. Zwei Möglichkeiten bieten sich hier an. Die erste besteht darin, die Rekursionsformel (siehe Satz 3.3)

$$B_+^{-1} = \left(I - \frac{sy^T}{y^T s} \right) B^{-1} \left(I - \frac{ys^T}{y^T s} \right) + \frac{ss^T}{y^T s}$$

auszunutzen. Ist B^{-1} bekannt (sehr häufig setzt man am Anfang $B_0 := D_0$ mit einer positiven Diagonalmatrix D_0 , z. B. $D_0 = I$, so daß B_0^{-1} leicht zu erhalten ist), so ist die zur Berechnung von B_+^{-1} benötigte Anzahl arithmetischer Operationen im wesentlichen proportional zu n^2 . Vom Aufwand her ist dieses Vorgehen befriedigend, es hat aber den Nachteil, daß man keine Kontrolle darüber behält, ob B_+ noch „hinreichend positiv definit“ ist. Daher geht man (nur scheinbar aufwendiger) so vor, daß man eine Update-Strategie für die Cholesky-Zerlegung entwickelt. Hierbei merkt man sich nicht die Matrizen B , sondern nur die Faktoren in einer Cholesky-Zerlegung. Genauer könnte das folgendermaßen aussehen:

- Sei $B = LL^T$ mit einer unteren Dreiecksmatrix L , deren Diagonalelemente positiv sind. Ferner sei $y^T s > 0$.

- Berechne

$$w := (y^T s)^{1/2} \frac{L^T s}{\|L^T s\|_2}, \quad J_+^T := L^T + \frac{w(y - Lw)^T}{y^T s}.$$

Dann ist $B_+ = J_+ J_+^T$ (siehe Beweis zu Satz 3.3).

- Berechne eine QR-Zerlegung $J_+^T = Q_+ R_+$, wobei die obere Dreiecksmatrix R_+ positive Diagonalelemente besitzt. Mit $L_+ := R_+^T$ ist dann

$$B_+ = J_+ J_+^T = R_+^T Q_+^T Q_+ R_+ = L_+ L_+^T$$

die gesuchte Cholesky-Zerlegung von B_+ . Besonders schön an dieser Vorgehensweise ist, daß man sich nicht den orthogonalen Anteil Q_+ in einer QR-Zerlegung von J_+^T zu merken braucht. Sind die Diagonalelemente von L_+ „hinreichend positiv“, so kann die neue Richtung $p_+ := -B_+^{-1} \nabla f(x_+)$ aus $L_+ L_+^T p_+ = -\nabla f(x_+)$ durch Vorwärts- und Rückwärtseinsetzen bestimmt werden.

Hierbei bleibt noch offen, wie man die QR-Zerlegung einer nichtsingulären Matrix $A_+ = R + uv^T$ auf effektive Weise berechnet, wobei R eine obere Dreiecksmatrix mit positiven Diagonalelementen ist. Bei der uns interessierenden Anwendung ist $R := L^T$ und z. B. (hier hat man Freiheiten bei der Verteilung der Konstanten)

$$u := \frac{L^T s}{\sqrt{y^T s \|L^T s\|_2}}, \quad v := y - (y^T s) Lu.$$

Das Ziel erreicht man, indem man A_+ sukzessive von links mit höchstens $2(n-1)$ Givens-Rotationen multipliziert. Sei $m := \max\{i \in \{1, \dots, n\} : u_i \neq 0\}$. Zunächst führt man den Vektor u durch sukzessive Multiplikation mit $m-1$ geeigneten Givens-Rotationen $G_{m-1,m}, \dots, G_{12}$, welche der Reihe nach die Komponenten mit den Indizes $m, \dots, 2$ annullieren, in ein Vielfaches $u_1 e_1$ des ersten Einheitsvektors über. Die parallel hierzu durchgeführte Multiplikation der oberen Dreiecksmatrix R mit den Givens-Rotationen $G_{m-1,m}, \dots, G_{12}$ transformiert diese in eine (obere) Hessenberg Matrix, die wir wieder mit R bezeichnen. Wir veranschaulichen uns diesen ersten Schritt im Falle $n=4$ und $m=3$, wobei festbleibende Elemente mit \bullet , sich verändernde mit $*$ bezeichnet werden.

$$\left(\begin{array}{cccc|c} \bullet & * & * & * & \bullet \\ * & \bullet & * & * & \bullet \\ * & * & \bullet & * & \bullet \\ * & * & * & \bullet & \bullet \end{array} \right) \xrightarrow{G_{23}} \left(\begin{array}{cccc|c} \bullet & * & * & * & \bullet \\ * & * & * & * & * \\ * & * & * & * & \bullet \\ * & * & * & * & \bullet \end{array} \right) \xrightarrow{G_{12}} \left(\begin{array}{cccc|c} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{array} \right)$$

Nach Abschluß dieses ersten Schrittes ist $G_{12} \cdots G_{m-1,m} A_+ = R + u_1 e_1 v^T$ mit einer Hessenberg-Matrix R (deren Subdiagonalelemente in den Spalten $m, \dots, n-1$ verschwinden). In einem Zwischenschritt berechnet man $R := R + u_1 e_1 v^T$, wodurch nur die erste Zeile verändert wird. Durch Multiplikation mit weiteren $m-1$ Givens-Rotationen $G_{12}, \dots, G_{m-1,m}$ annulliert man schließlich im letzten Schritt die störenden Subdiagonalelemente in den Spalten $1, \dots, m-1$. Hierbei hat man darauf zu

achten, daß die erzeugten Diagonalelemente positiv sind. Auch diesen Schritt wollen wir uns für $n = 4, m = 3$ veranschaulichen.

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{12}} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{23}} \begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ * & * & * & * \\ * & * & * & * \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix}$$

Damit ist ausführlich beschrieben, wie man aus einer Cholesky-Zerlegung der positiv definiten Matrix B eine der BFGS-Update-Matrix B_+ berechnen kann. Die hierzu benötigte Anzahl arithmetischer Operationen ist offenbar im wesentlichen proportional zu n^2 . \square

7.3.3 Die globale Konvergenz des BFGS-Verfahrens

Wir betrachten wieder die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

und setzen in diesem Unterabschnitt voraus, daß die folgenden Voraussetzungen erfüllt sind (siehe auch Lemma 2.6):

- (K) (a) Mit einem $x_0 \in \mathbb{R}^n$ ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.
- (b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und auf L_0 gleichmäßig konvex, d. h. es existiert eine Konstante $c > 0$ mit

$$\frac{c}{2} \|y - x\|_2^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \quad \text{für alle } x, y \in L_0.$$

- (c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2 \quad \text{für alle } x, y \in L_0.$$

Ziel in diesem Unterabschnitt ist es, den folgenden globalen Konvergenzsatz für das BFGS-Verfahren zu beweisen. Bemerkenswert an diesem ist vor allem, daß in jedem Iterationsschritt die Verwendung von inexakten Schrittweiten erlaubt ist.

Satz 3.6 Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (K) (a)–(c) seien erfüllt. Man betrachte das BFGS-Verfahren:

- Der Startwert $x_0 \in \mathbb{R}^n$ genüge (K), zur Abkürzung sei $g_0 := \nabla f(x_0)$. Gegeben sei eine symmetrische, positiv definite Matrix $B_0 \in \mathbb{R}^{n \times n}$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: STOP, x_k ist Lösung von (P).

Berechne $p_k := -B_k^{-1}g_k$.

Sei t_k die exakte Schrittweite oder die Powell- oder die Armijo-Schrittweite.

Setze $x_{k+1} := x_k + t_k p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$.

Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ sei

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Dann gilt: Das BFGS-Verfahren bricht nach endlich vielen Schritten mit der Lösung x^* von (P) ab oder es liefert eine Folge $\{x_k\}$, die gegen x^* konvergiert. Genauer existieren in diesem Falle Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\|_2 \leq C q^k$ für $k = 0, 1, \dots$

Beweis: Die Durchführbarkeit des Verfahrens ist gesichert. Denn ist $g_k \neq 0$ (andernfalls ist $x_k = x^*$ Lösung von (P)) und B_k symmetrisch und positiv definit, so ist $p_k = -B_k^{-1}g_k$ eine Abstiegsrichtung und daher $s_k \neq 0$. Wegen der gleichmäßigen Konvexität der Zielfunktion f ist folglich $y_k^T s_k \geq c \|s_k\|_2^2 > 0$ und aus Satz 3.3 folgt, daß auch B_{k+1} symmetrisch und positiv definit ist.

Wir nehmen nun an, das Verfahren breche nicht schon nach endlich vielen Schritten mit der Lösung ab. Wir wollen Satz 2.7 anwenden und zeigen hierzu die Existenz einer Konstanten $\delta > 0$ mit $\delta(k+1) \leq \sum_{j=0}^k \delta_j$ für $k = 0, 1, \dots$, wobei

$$\delta_j := \min \left[-\frac{g_j^T p_j}{\|g_j\|_2^2}, \left(\frac{g_j^T p_j}{\|g_j\|_2 \|p_j\|_2} \right)^2 \right].$$

Um dies nachzuweisen, zeigen wir:

$$(a) \text{ Es gibt eine Konstante } c_1 > 0 \text{ mit } -\sum_{j=0}^k \frac{\|g_j\|_2^2}{g_j^T p_j} \leq c_1(k+1) \text{ für } k = 0, 1, \dots$$

Denn: Aus der Update-Formel für B_{k+1} erhält man

$$\text{Spur}(B_{k+1}) = \text{Spur}(B_k) - \frac{\|B_k s_k\|_2^2}{s_k^T B_k s_k} + \frac{\|y_k\|_2^2}{y_k^T s_k} = \text{Spur}(B_0) - \sum_{j=0}^k \frac{\|B_j s_j\|_2^2}{s_j^T B_j s_j} + \sum_{j=0}^k \frac{\|y_j\|_2^2}{y_j^T s_j}.$$

Die Spur der positiv definiten Matrix B_{k+1} ist positiv, ferner ist

$$\frac{\|y_j\|_2^2}{y_j^T s_j} \leq \frac{\gamma^2 \|s_j\|_2^2}{c \|s_j\|_2^2} = \frac{\gamma^2}{c}$$

wegen der Lipschitzstetigkeit des Gradienten und der gleichmäßigen Konvexität der Zielfunktion. Daher ist

$$-\sum_{j=0}^k \frac{\|g_j\|_2^2}{g_j^T p_j} = \sum_{j=0}^k \frac{|B_j p_j|^2}{p_j^T B_j p_j} = \sum_{j=0}^k \frac{\|B_j s_j\|_2^2}{s_j^T B_j s_j} < \text{Spur}(B_0) + \frac{\gamma^2}{c} (k+1) \leq c_1(k+1)$$

mit $c_1 := \text{Spur}(B_0) + \gamma^2/c$. Damit ist (a) nachgewiesen.

- (b) Es gibt eine Konstante $c_2 > 0$ mit $\prod_{j=0}^k \left(\frac{\|g_j\|_2 \|p_j\|_2}{g_j^T p_j} \right)^2 \leq c_2^{k+1}$ für $k = 0, 1, \dots$

Denn: Aus Satz 3.3 und mit Hilfe der Ungleichung vom geometrisch-arithmetischen Mittel (angewandt auf die Eigenwerte von B_{k+1}) folgt

$$\det(B_{k+1}) = \frac{y_k^T s_k}{s_k^T B_k s_k} \det(B_k) = \prod_{j=0}^k \frac{y_j^T s_j}{s_j^T B_j s_j} \det(B_0) \leq \left[\frac{1}{n} \operatorname{Spur}(B_{k+1}) \right]^n.$$

Beim Beweis von (a) haben wir insbesondere erhalten, daß $\operatorname{Spur}(B_{k+1}) \leq c_1(k+1)$, so daß

$$\prod_{j=0}^k \frac{y_j^T s_j}{s_j^T B_j s_j} \leq \left[\frac{1}{n} c_1(k+1) \right]^n \det(B_0^{-1}).$$

Ferner ist wegen der Ungleichung vom geometrisch-arithmetischen Mittel sowie (a):

$$\prod_{j=0}^k \frac{\|B_j s_j\|_2^2}{s_j^T B_j s_j} \leq \left(\frac{1}{k+1} \sum_{j=0}^k \frac{\|B_j s_j\|_2^2}{s_j^T B_j s_j} \right)^{k+1} \leq c_1^{k+1}.$$

Damit wird schließlich

$$\begin{aligned} \prod_{j=0}^k \left(\frac{\|g_j\|_2 \|p_j\|_2}{g_j^T p_j} \right)^2 &= \prod_{j=0}^k \frac{\|B_j s_j\|_2^2 \|s_j\|_2^2}{(s_j^T B_j s_j)^2} \\ &\leq \frac{1}{c^{k+1}} \prod_{j=0}^k \frac{\|B_j s_j\|_2^2}{s_j^T B_j s_j} \prod_{j=0}^k \frac{y_j^T s_j}{s_j^T B_j s_j} \quad (\text{wegen } c \|s_j\|_2^2 \leq y_j^T s_j) \\ &\leq \left(\frac{c_1}{c} \right)^{k+1} \left[\frac{1}{n} c_1(k+1) \right]^n \det(B_0^{-1}) \\ &\leq c_2^{k+1} \end{aligned}$$

mit hinreichend großem $c_2 > 0$. Damit ist auch (b) bewiesen.

- (c) Sind $\alpha_0, \dots, \alpha_k \geq 0$ und $a > 0$ eine Konstante mit $\sum_{j=0}^k \alpha_j \leq a(k+1)$, so gibt es eine Indexmenge $J_k \subset \{0, \dots, k\}$, die mindestens $\frac{2}{3}(k+1)$ Elemente enthält, und für die $\alpha_j \leq 3a$ für alle $j \in J_k$.

Denn: Man definiere $I_k := \{i \in \{0, \dots, k\} : \alpha_i > 3a\}$. Dann ist

$$(k+1)a \geq \sum_{j=0}^k \alpha_j \geq \sum_{i \in I_k} \alpha_i > \operatorname{Anzahl}(I_k) 3a,$$

so daß I_k weniger als $\frac{1}{3}(k+1)$ Elemente enthält. Dann ist $J_k := \{0, \dots, k\} \setminus I_k$ die gesuchte Indexmenge.

- (d) Sind $\beta_0, \dots, \beta_k \geq 1$ und $b > 1$ eine Konstante mit $\prod_{j=0}^k \beta_j \leq b^{k+1}$, so existiert eine Indexmenge $J_k \subset \{0, \dots, k\}$, die mindestens $\frac{2}{3}(k+1)$ Elemente enthält, und für die $\beta_j \leq b^3$ für alle $j \in J_k$.

Denn: Die Behauptung (d) folgt aus (c), indem man $\alpha_j := \ln \beta_j$ und $a := \ln b$ setzt.

Aus (a)–(d) erhält man bei festem $k \in \mathbb{N} \cup \{0\}$ die Existenz einer Indexmenge $J_k \subset \{0, \dots, k\}$, die mindestens $\frac{1}{3}(k+1)$ Elemente enthält, mit

$$-\frac{\|g_j\|_2^2}{g_j^T p_j} \leq 3c_1, \quad \left(\frac{\|g_j\|_2 \|p_j\|_2}{g_j^T p_j} \right)^2 \leq c_2^3 \quad \text{für alle } j \in J_k.$$

Damit ist

$$\sum_{j=0}^k \delta_j = \sum_{j=0}^k \min \left[-\frac{g_j^T p_j}{\|g_j\|_2^2}, \left(\frac{g_j^T p_j}{\|g_j\|_2 \|p_j\|_2} \right)^2 \right] \geq \underbrace{\frac{1}{3} \min \left(\frac{1}{3c_1}, \frac{1}{c_2^3} \right)}_{=: \delta} (k+1).$$

Aus Satz 2.7 folgt die Behauptung. \square

Bemerkung: Der erste globale Konvergenzsatz für das DFP-Verfahren mit exakter Schrittweitenbestimmung stammt von M. J. D. POWELL (1971). Wegen Satz 3.4 erhält man hieraus einen globalen Konvergenzsatz für jedes Verfahren der Broyden-Klasse, wenn in jedem Schritt die exakte Schrittweite benutzt wird. Für das BFGS-Verfahren konnte die globale Konvergenz mit einer inexakten Schrittweite von M. J. D. POWELL (1976) bewiesen werden. Die hier entwickelten Techniken wurden von J. WERNER (1978) benutzt, um eine Verallgemeinerung dieses Ergebnisses zu beweisen, indem weitere inexakte Schrittweiten zugelassen wurden. Sehr ähnlich gehen auch R. H. BYRD, J. NOCEDAL, Y. YUAN (1987) vor, um die globale Konvergenz für die eingeschränkte Broyden-Klasse (hier sind die Parameter ϕ_k aus dem Intervall $[0, 1]$ zu wählen) bei konvexer Zielfunktion zu untersuchen. Es kann die globale Konvergenz gezeigt werden, wenn mit einem $\phi \in [0, 1]$ stets $\phi_k \in [0, \phi]$ gewählt wird, wodurch das DFP-Verfahren ausgeschlossen ist. Die Aussage von Satz 3.6, in dem auch die Armijo-Schrittweite zugelassen ist, findet man (mit einem etwas anderen Beweis) auch bei R. H. BYRD, J. NOCEDAL (1989). \square

7.3.4 Die superlineare Konvergenz des BFGS-Verfahrens

In diesem Unterabschnitt sollen Aussagen über die Konvergenzgeschwindigkeit des BFGS-Verfahren gemacht werden. Genauer soll unter geeigneten Voraussetzungen die superlineare Konvergenz des BFGS-Verfahrens bewiesen werden. Hier muß man zwischen Aussagen des folgenden Typs unterscheiden.

- Die Folge $\{x_k\}$ sei durch das (gedämpfte) BFGS-Verfahren erzeugt, es sei also $x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k)$, wobei mit einer symmetrischen, positiv definiten Matrix B_0 die Folge der symmetrischen, positiv definiten Matrizen $\{B_k\}$ durch die BFGS-Update-Formel berechnet sei. Konvergiert dann die Folge $\{x_k\}$ so „schnell“ gegen eine stationäre Lösung x^* von (P), daß $\sum_{k=0}^{\infty} \|x_k - x^*\|_2 < \infty$ (dies ist sicher dann der Fall, wenn Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\|_2 \leq C q^k$ existieren), und ist darüber hinaus in x^* die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt, ist also $\nabla^2 f(x^*)$ positiv definit, so konvergiert $\{x_k\}$ sogar superlinear gegen x^* , wenn $t_k = 1$ für alle hinreichend

großen k . Dies ist insbesondere dann der Fall, wenn t_k die Armijo-Schrittweite ist.

- Sei $f: \mathbb{R}^n \rightarrow \mathbb{R}$ zweimal stetig differenzierbar auf einer offenen konvexen Menge $D \subset \mathbb{R}^n$, auf der $\nabla^2 f(\cdot)$ auch lipschitzstetig ist. Es existiere ein $x^* \in D$, in dem die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt sind, für welches also $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit ist. Dann existieren positive Konstanten ϵ und δ mit der folgenden Eigenschaft: Ist $\|x_0 - x^*\|_2 \leq \epsilon$ und B_0 eine symmetrische, positiv definite Matrix mit $\|B_0 - \nabla^2 f(x^*)\|_2 \leq \delta$, so ist das ungedämpfte BFGS-Verfahren durchführbar und liefert eine Folge $\{x_k\}$, die in D bleibt und superlinear gegen x^* konvergiert.

Wir werden nur auf die erste Aussage eingehen. Einen Beweis des lokalen Konvergenzsatzes in der zweiten Aussage findet man bei C. G. BROYDEN, J. E. DENNIS, J. J. MORÉ (1973).

Satz 3.7 Gegeben sei die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n.$$

Mit einem $x_0 \in \mathbb{R}^n$ und einer symmetrischen, positiv definiten Matrix $B_0 \in \mathbb{R}^{n \times n}$ sei das gedämpfte BFGS-Verfahren durchführbar und liefere eine Folge $\{x_k\}$ mit $x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k)$, die gegen einen Punkt x^* wenigstens so schnell konvergiert, daß $\sum_{k=0}^{\infty} \|x_k - x^*\|_2 < \infty$ (Satz 3.6 gibt hierfür hinreichende Bedingungen an), und eine Folge $\{B_k\}$ symmetrischer, positiv definiter Matrizen. Ferner seien in x^* die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt, d. h. die Zielfunktion f sei auf einer offenen und konvexen Umgebung U^* von x^* zweimal stetig differenzierbar und es sei $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit. Schließlich sei noch $\nabla^2 f(\cdot)$ auf U^* in x^* lipschitzstetig, d. h. es existiere eine Konstante $L > 0$ mit $\|\nabla^2 f(x) - \nabla^2 f(x^*)\|_2 \leq L \|x - x^*\|_2$ für alle $x \in U^*$. Dann gilt:

1. Die Folgen $\{\|B_k\|_2\}$ und $\{\|B_k^{-1}\|_2\}$ sind beschränkt.

2. Es ist

$$\lim_{k \rightarrow \infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|_2}{\|x_{k+1} - x_k\|_2} = 0.$$

3. Gilt $\lim_{k \rightarrow \infty} t_k = 1$, so konvergiert die Folge $\{x_k\}$ superlinear gegen x^* .

4. Wird im Verfahren stets die Armijo-Schrittweite gewählt, so ist $t_k = 1$ für alle hinreichend großen k . In diesem Falle geht also das gedämpfte BFGS-Verfahren nach endlich vielen Schritten in das ungedämpfte Verfahren über und liefert eine superlinear gegen x^* konvergierende Folge $\{x_k\}$.

Beweis: Der Beweis der ersten beiden Aussagen ist der bei weitem schwierigste Teil. Hier folgen wir der sehr schönen Darstellung bei R. H. BYRD, J. NOCEDAL (1989, Theorem 3.2).

Wegen der vorausgesetzten Konvergenz der Folge $\{x_k\}$ gegen x^* ist $x_k \in U^*$ für alle hinreichend großen k . Wir benutzen die gewohnten Bezeichnungen. So sei etwa $s_k := x_{k+1} - x_k$ und $y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$. Nun definieren wir

$$\tilde{s}_k := \nabla^2 f(x^*)^{1/2} s_k, \quad \tilde{y}_k := \nabla^2 f(x^*)^{-1/2} y_k, \quad \tilde{B}_k := \nabla^2 f(x^*)^{-1/2} B_k \nabla^2 f(x^*)^{-1/2}.$$

Dann ist

$$\tilde{B}_{k+1} = \tilde{B}_k - \frac{(\tilde{B}_k \tilde{s}_k)(\tilde{B}_k \tilde{s}_k)^T}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} + \frac{\tilde{y}_k \tilde{y}_k^T}{\tilde{y}_k^T \tilde{s}_k}$$

und daher (siehe Satz 3.3)

$$\text{Spur}(\tilde{B}_{k+1}) = \text{Spur}(\tilde{B}_k) - \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} + \frac{\|\tilde{y}_k\|_2^2}{\tilde{y}_k^T \tilde{s}_k}, \quad \det(\tilde{B}_{k+1}) = \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} \det(\tilde{B}_k).$$

Der hübsche Trick von Byrd-Nocedal besteht darin, für positiv definites $B \in \mathbb{R}^{n \times n}$ die Funktion

$$\psi(B) := \text{Spur}(B) - \ln \det(B)$$

einzuführen. Dann ist

$$\begin{aligned} \psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \frac{\|\tilde{y}_k\|_2^2}{\tilde{y}_k^T \tilde{s}_k} - \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} - \ln \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} \\ &= \psi(\tilde{B}_k) + \frac{\|\tilde{y}_k\|_2^2}{\tilde{y}_k^T \tilde{s}_k} - 1 - \ln \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{s}_k\|_2^2} + \ln \left(\frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{B}_k \tilde{s}_k\|_2 \|\tilde{s}_k\|_2} \right)^2 \\ &\quad + \left[1 - \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} + \ln \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} \right]. \end{aligned}$$

Wegen der Lipschitzstetigkeit von $\nabla^2 f(\cdot)$ auf U^* in x^* ist

$$\begin{aligned} \frac{\|\tilde{y}_k - \tilde{s}_k\|_2}{\|\tilde{s}_k\|_2} &= \frac{\|\nabla^2 f(x^*)^{-1/2}[y_k - \nabla^2 f(x^*)s_k]\|_2}{\|\nabla^2 f(x^*)^{1/2}s_k\|_2} \\ &\leq \|\nabla^2 f(x^*)^{-1}\|_2 \frac{\|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x^*)(x_{k+1} - x_k)\|_2}{\|x_{k+1} - x_k\|_2} \\ &\leq \|\nabla^2 f(x^*)^{-1}\|_2 L \int_0^1 \|x_k + t(x_{k+1} - x_k) - x^*\|_2 dt \\ &\leq \frac{L \|\nabla^2 f(x^*)^{-1}\|_2}{2} (\|x_k - x^*\|_2 + \|x_{k+1} - x^*\|_2) \\ &=: \epsilon_k \end{aligned}$$

für alle hinreichend großen k . Wir merken uns, daß $\sum_{k=0}^{\infty} \epsilon_k < \infty$, da nach Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\|_2 < \infty$. Für alle hinreichend großen k ist

$$\frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{s}_k\|_2^2} = 1 + \frac{(\tilde{y}_k - \tilde{s}_k)^T \tilde{s}_k}{\|\tilde{s}_k\|_2^2} \geq 1 - \frac{\|\tilde{y}_k - \tilde{s}_k\|_2}{\|\tilde{s}_k\|_2} \geq 1 - \epsilon_k$$

und

$$\frac{\|\tilde{y}_k\|_2^2}{\tilde{y}_k^T \tilde{s}_k} \leq \left(\frac{\|\tilde{s}_k\|_2 + \|\tilde{y}_k - \tilde{s}_k\|_2}{\|\tilde{s}_k\|_2} \right)^2 \frac{\|\tilde{s}_k\|_2^2}{\tilde{y}_k^T \tilde{s}_k} \leq \frac{(1 + \epsilon_k)^2}{1 - \epsilon_k} \leq 1 + c\epsilon_k$$

mit einer hinreichend großen Konstanten $c > 1$. Für alle hinreichend großen k ist $\epsilon_k < \frac{1}{2}$ und daher

$$\ln \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{s}_k\|_2^2} \geq \ln(1 - \epsilon_k) \geq -2\epsilon_k \geq -2c\epsilon_k.$$

Insgesamt ist

$$\psi(\tilde{B}_{k+1}) \leq \psi(\tilde{B}_k) + 3c\epsilon_k + \ln\left(\frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{B}_k \tilde{s}_k\|_2 \|\tilde{s}_k\|_2}\right)^2 + \left[1 - \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} + \ln \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}\right]$$

für alle hinreichend großen k . Nach Aufsummieren unter Berücksichtigung von $\sum_{j=0}^{\infty} \epsilon_j < \infty$ ist mit einer hinreichend großen Konstanten $\hat{c} > 0$ daher

$$(*) \quad \psi(\tilde{B}_k) \leq \hat{c} + \sum_{j=0}^{k-1} \underbrace{\left\{ \ln\left(\frac{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j}{\|\tilde{B}_j \tilde{s}_j\|_2 \|\tilde{s}_j\|_2}\right)^2 + \left[1 - \frac{\|\tilde{B}_j \tilde{s}_j\|_2^2}{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j} + \ln \frac{\|\tilde{B}_j \tilde{s}_j\|_2^2}{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j}\right] \right\}}_{\leq 0} \leq \hat{c}$$

für alle $k \in \mathbb{N} \cup \{0\}$. Aus $(*)$ erkennen wir zweierlei. Einerseits ist die Folge $\{\psi(\tilde{B}_k)\}$ nach oben durch die Konstante \hat{c} beschränkt. Hieraus folgt die Beschränktheit von $\{\|\tilde{B}_k\|_2\}$ und $\{\|\tilde{B}_k^{-1}\|_2\}$. Denn bezeichnen $\tilde{\lambda}_n^{(k)} \leq \dots \leq \tilde{\lambda}_1^{(k)}$ die (positiven) Eigenwerte der positiv definiten Matrix \tilde{B}_k , so ist

$$\hat{c} \geq \psi(\tilde{B}_k) = \text{Spur}(\tilde{B}_k) - \ln \det(\tilde{B}_k) = \sum_{i=1}^n (\underbrace{\tilde{\lambda}_i^{(k)} - \ln \tilde{\lambda}_i^{(k)}}_{\geq 1}).$$

Hieraus wiederum folgt $\hat{c} \geq \tilde{\lambda}_1^{(k)} - \ln \tilde{\lambda}_1^{(k)} \geq \ln \tilde{\lambda}_1^{(k)}$ und $\hat{c} \geq -\ln \tilde{\lambda}_n^{(k)}$, so daß

$$\|\tilde{B}_k\|_2 = \tilde{\lambda}_1^{(k)} \leq \exp(\hat{c}), \quad \|\tilde{B}_k^{-1}\|_2 = \frac{1}{\tilde{\lambda}_n^{(k)}} \leq \exp(\hat{c}) \quad \text{für } k = 0, 1, \dots$$

Mit $\{\|\tilde{B}_k\|_2\}$ und $\{\|\tilde{B}_k^{-1}\|_2\}$ sind auch die Folgen $\{\|B_k\|_2\}$ und $\{\|B_k^{-1}\|_2\}$ beschränkt. Andererseits folgt aus $(*)$ unter Berücksichtigung von $\psi(\tilde{B}_k) > 0$, daß

$$\sum_{j=0}^{k-1} \underbrace{\left\{ \ln\left(\frac{\|\tilde{B}_j \tilde{s}_j\|_2 \|\tilde{s}_j\|_2}{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j}\right)^2 + \left[\frac{\|\tilde{B}_j \tilde{s}_j\|_2^2}{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j} - 1 - \ln \frac{\|\tilde{B}_j \tilde{s}_j\|_2^2}{\tilde{s}_j^T \tilde{B}_j \tilde{s}_j}\right] \right\}}_{\geq 0} \geq 0 \quad \text{für alle } k \in \mathbb{N},$$

woraus

$$\lim_{k \rightarrow \infty} \ln\left(\frac{\|\tilde{B}_k \tilde{s}_k\|_2 \|\tilde{s}_k\|_2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}\right)^2 = 0, \quad \lim_{k \rightarrow \infty} \left[\frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} - 1 - \ln \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} \right] = 0$$

folgt. Daher ist

$$\lim_{k \rightarrow \infty} \frac{\|\tilde{B}_k \tilde{s}_k\|_2 \|\tilde{s}_k\|_2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} = 1, \quad \lim_{k \rightarrow \infty} \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} = 1.$$

Damit wird

$$\begin{aligned} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|_2^2}{\|x_{k+1} - x_k\|_2^2} &= \frac{\|\nabla^2 f(x^*)^{1/2}(\tilde{B}_k - I)\tilde{s}_k\|_2^2}{\|\nabla^2 f(x^*)^{-1/2}\tilde{s}_k\|_2^2} \\ &\leq \|\nabla^2 f(x^*)\|_2^2 \frac{\|(\tilde{B}_k - I)\tilde{s}_k\|_2^2}{\|\tilde{s}_k\|_2^2} \end{aligned}$$

$$\begin{aligned}
&= \|\nabla^2 f(x^*)\|_2^2 \frac{\|\tilde{B}_k \tilde{s}_k\|_2^2 - 2\tilde{s}_k^T \tilde{B}_k \tilde{s}_k + \|\tilde{s}_k\|_2^2}{\|\tilde{s}_k\|_2^2} \\
&= \|\nabla^2 f(x^*)\|_2^2 \left(\underbrace{\frac{\|\tilde{B}_k \tilde{s}_k\|_2^2}{\|\tilde{s}_k\|_2^2}}_{\rightarrow 1} - 2 \underbrace{\frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\|_2^2}}_{\rightarrow 1} + 1 \right) \\
&\rightarrow 0 \quad \text{für } k \rightarrow \infty,
\end{aligned}$$

womit schließlich die ersten beiden Aussagen des Satzes bewiesen sind.

Die dritte Aussage des Satzes, daß nämlich aus $\lim_{k \rightarrow \infty} t_k = 1$ die superlineare Konvergenz der Folge $\{x_k\}$ gegen x^* folgt, kann leicht mit dem Satz von Dennis-Moré über die Charakterisierung der superlinearen Konvergenz bei Quasi-Newton-Verfahren (siehe Satz 3.2 in Abschnitt 2.3) bewiesen werden. Um die Darstellung in sich abgeschlossen zu machen, geben wir einen auf die vorliegende Situation zugeschnittenen ad hoc Beweis an.

Wir können annehmen, daß die offene und konvexe Umgebung U^* von x^* so klein ist, daß eine Konstante $c > 0$ mit

$$c \|p\|_2^2 \leq p^T \nabla^2 f(x)p \quad \text{für alle } x \in U^*, p \in \mathbb{R}^n$$

existiert. Für alle hinreichend großen k ist

$$\begin{aligned}
\nabla f(x_{k+1}) &= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x^*)(x_{k+1} - x_k) \\
&\quad - \left[\frac{1}{t_k} B_k - \nabla^2 f(x^*) \right] (x_{k+1} - x_k) \\
&= \int_0^1 [\nabla^2 f(x_k + s(x_{k+1} - x_k)) - \nabla^2 f(x^*)] (x_{k+1} - x_k) ds \\
&\quad - \left[\frac{1}{t_k} B_k - \nabla^2 f(x^*) \right] (x_{k+1} - x_k).
\end{aligned}$$

Die Lipschitzstetigkeit von $\nabla^2 f(\cdot)$ in x^* und die Dreiecksungleichung liefern

$$\begin{aligned}
\frac{\|\nabla f(x_{k+1})\|_2}{\|x_{k+1} - x_k\|_2} &\leq \frac{L}{2} (\|x_{k+1} - x^*\|_2 + \|x_k - x^*\|_2) + \left| \frac{1}{t_k} - 1 \right| \|B_k\|_2 \\
&\quad + \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|_2}{\|x_{k+1} - x_k\|_2}
\end{aligned}$$

für alle hinreichend großen k . Wegen der vorausgesetzten Konvergenz von $\{x_k\}$ gegen x^* , wegen $\lim_{k \rightarrow \infty} t_k = 1$ und der beiden schon bewiesenen Aussagen des Satzes konvergieren alle drei Terme auf der rechten Seite gegen Null. Berücksichtigt man nun noch, daß

$$(x_{k+1} - x^*)^T [\nabla f(x_{k+1}) - \underbrace{\nabla f(x^*)}_{=0}] \geq c \|x_{k+1} - x^*\|_2^2$$

und daher $\|\nabla f(x_{k+1})\|_2 \geq c \|x_{k+1} - x^*\|_2$ für alle hinreichend großen k , so erhält man

$$\frac{1}{1 + \|x_k - x^*\|_2 / \|x_{k+1} - x^*\|_2} = \frac{\|x_{k+1} - x^*\|_2}{\|x_{k+1} - x^*\|_2 + \|x_k - x^*\|_2} \leq \frac{\|\nabla f(x_{k+1})\|_2}{c \|x_{k+1} - x_k\|_2} \rightarrow 0,$$

woraus die superlineare Konvergenz der Folge $\{x_k\}$ gegen x^* folgt.

Für den letzten Teil des Satzes ist zu zeigen, daß mit $p_k := -B_k^{-1}\nabla f(x_k)$ und vorgegebenem $\alpha \in (0, \frac{1}{2})$ die Ungleichung

$$f(x_k + p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k$$

für alle hinreichend großen k erfüllt ist. Wie beim Beweis von Satz 3.2 zeigen wir

$$\lim_{k \rightarrow \infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha,$$

woraus die Behauptung folgt. Hierzu bemerken wir zunächst, daß die Richtungsfolge $\{p_k\}$ wegen der Beschränktheit von $\{\|B_k^{-1}\|_2\}$ (siehe erster Teil des Satzes) gegen den Nullvektor konvergiert. Daher ist die gesamte Verbindungsstrecke zwischen x_k und $x_k + p_k$ für alle hinreichend großen k in U^* enthalten. Aus dem Mittelwertsatz folgt die Existenz von $\theta_k \in (0, 1)$ mit

$$\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} - \frac{p_k^T [\nabla^2 f(x_k + \theta_k p_k) - B_k] p_k}{2 p_k^T B_k p_k}.$$

Nun ist

$$\begin{aligned} \frac{|p_k^T [\nabla^2 f(x_k + \theta_k p_k) - B_k] p_k|}{p_k^T B_k p_k} &\leq \|B_k^{-1}\|_2 \left[\underbrace{\|\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x^*)\|_2}_{\rightarrow 0} \right. \\ &\quad \left. + \underbrace{\frac{\|[B_k - \nabla^2 f(x^*)] p_k\|_2}{\|p_k\|_2}}_{\rightarrow 0} \right], \end{aligned}$$

und hieraus folgt wegen der Beschränktheit von $\{\|B_k^{-1}\|_2\}$ die Behauptung. Insgesamt ist der Satz damit bewiesen. \square

Bemerkungen: Nur für den Beweis der ersten beiden Aussagen von Satz 3.7 geht ein, daß es sich bei dem betrachteten Verfahren um das BFGS-Verfahren handelt. Hierin besteht die Hauptarbeit, der Beweis der beiden restlichen Aussagen ist dann einfach und vom BFGS-Verfahren unabhängig.

Die Voraussetzungen des globalen Konvergenzsatzes 3.2 für das Newton-Verfahren seien erfüllt. Aus dem globalen Konvergenzsatz 3.6 für das (gedämpfte) BFGS-Verfahren folgt die Konvergenz einer durch dieses Verfahren erzeugten Folge $\{x_k\}$ gegen die (einheitige) globale Lösung x^* , ferner ist $\sum_{k=0}^{\infty} \|x_k - x^*\|_2 < \infty$ gesichert. Ist $\nabla^2 f(\cdot)$ auch noch in x^* lipschitzstetig, so folgt aus Satz 3.7 die superlineare Konvergenz von $\{x_k\}$, wenn $\lim_{k \rightarrow \infty} t_k = 1$ gilt. Dies ist für die Armijo-Schrittweite der Fall, kann aber auch für die exakte Schrittweite und die Powell-Schrittweite nachgewiesen werden. Letzteres wird in Aufgabe 5 präzisiert. \square

Aufgaben

- Unter den Voraussetzungen von Satz 3.3 sei B_+ die mit einem Parameter $\phi \geq 0$ aus B gewonnene Update-Matrix der Broyden-Klasse. Man zeige

$$\det(B_+) = \left[(1 - \phi) \frac{y^T s}{s^T B s} + \phi \frac{y^T B^{-1} y}{y^T s} \right] \det(B).$$

2. Seien die symmetrische Matrix $H \in \mathbb{R}^{n \times n}$ und $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ gegeben. Hiermit sei

$$E_+ := \left(1 + \frac{y^T H y}{y^T s}\right) \frac{ss^T}{y^T s} - \frac{s(Hy)^T + (Hy)s^T}{y^T s}.$$

Ferner sei $M \in \mathbb{R}^{n \times n}$ eine nichtsinguläre Matrix mit $M^T M s = y$. Dann ist E_+ die eindeutige Lösung der Aufgabe

$$\text{Minimiere } \|M E M^T\|_F \text{ auf } K := \{E \in \mathbb{R}^{n \times n} : E = E^T, E y = s - Hy\}.$$

Hinweis: Auf $\mathbb{R}^{n \times n}$ definiere man ein inneres Produkt $(\cdot, \cdot)_M$ durch

$$(A, B)_M := \text{Spur}(MA^T M^T MBM^T).$$

Die durch dieses innere Produkt induzierte Norm ist

$$\|A\|_M := (A, A)_M^{1/2} = [\text{Spur}(MA^T M^T MAM^T)]^{1/2} = \|MAM^T\|_F.$$

Daher ist zu zeigen, daß $E_+ \in K$ die (bekanntlich eindeutig existierende) Projektion der Nullmatrix $0 \in \mathbb{R}^{n \times n}$ auf den affin linearen Teilraum $K \subset \mathbb{R}^{n \times n}$ bezüglich der durch $(\cdot, \cdot)_M$ erzeugten Norm $\|\cdot\|_M$ ist. Dies wiederum ist gleichwertig damit, daß E_+ auf dem linearen Teilraum $L := \{F \in \mathbb{R}^{n \times n} : F = F^T, Fy = 0\}$ bezüglich des inneren Produktes $(\cdot, \cdot)_M$ senkrecht steht.

Die Aussage der Aufgabe impliziert insbesondere: Ist $B \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, sind $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und ist $M \in \mathbb{R}^{n \times n}$ nichtsingulär mit $M^T M s = y$, so erhält man die Inverse der BFGS-Update-Matrix B_{BFGS} gerade als Lösung der Aufgabe, $\|M(B_+^{-1} - B^{-1})M^T\|_F$ unter allen symmetrischen Matrizen B_+ mit $B_+^{-1}y = s$ zu minimieren. Mit anderen Worten erhält man B_{BFGS}^{-1} , indem man B^{-1} bezüglich einer geeigneten, gewichteten Frobenius-Norm auf die Menge der symmetrischen Matrizen B_+^{-1} mit $B_+^{-1}y = s$ projiziert.

3. Analog zu Aufgabe 2 zeige man: Ist $B \in \mathbb{R}^{n \times n}$ symmetrisch, $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und $M \in \mathbb{R}^{n \times n}$ eine nichtsinguläre Matrix mit $MM^T s = y$, so ist

$$E_+ := \left(1 + \frac{s^T Bs}{y^T s}\right) \frac{yy^T}{y^T s} - \frac{y(Bs)^T + (Bs)y^T}{y^T s}$$

die eindeutige Lösung der Aufgabe

$$\text{Minimiere } \|M^{-1} E M^{-T}\|_F \text{ auf } K := \{E \in \mathbb{R}^{n \times n} : E = E^T, Es = y - Bs\}.$$

Hinweis: Es ist $B + E_+ = B_{\text{DFP}}$ gerade die DFP-Update-Matrix ($\phi = 1$ in der Broyden-Klasse). Die Aussage der Aufgabe ist daher, daß man B_{DFP} als Projektion von B auf die Menge der symmetrischen Matrizen B_+ mit $B_+s = y$ (Quasi-Newton-Gleichung) bezüglich einer geeigneten, gewichteten Frobenius-Norm erhält.

4. Sei $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n \times n}$ und $c \in \mathbb{R}^n$. Dann ist das ungedämpfte BFGS-Verfahren (hier wird also stets die konstante Schrittweite $t_k = 1$ benutzt) mit beliebigen Startwerten $x_0 \in \mathbb{R}^n$ und (symmetrischer, positiv definiter Matrix) $B_0 \in \mathbb{R}^{n \times n}$ durchführbar und liefert eine Folge $\{x_k\}$, die superlinear gegen das eindeutige Minimum $x^* := -Q^{-1}c$ von f konvergiert.

Hinweis: Man benutze die üblichen Bezeichnungen. Da die Zielfunktion f quadratisch ist, ist $y = Qs$. Mit $\tilde{B} := Q^{-1/2} B Q^{-1/2}$ und $\tilde{s} := Q^{1/2}s$ zeige man:

(a) Es ist

$$\tilde{B}_+ = \tilde{B} - \frac{(\tilde{B}\tilde{s})(\tilde{B}\tilde{s})^T}{\tilde{s}^T \tilde{B}\tilde{s}} + \frac{\tilde{s}\tilde{s}^T}{\|\tilde{s}\|_2^2}$$

und daher $\|\tilde{B}_+\|_2 \leq \max(1, \|\tilde{B}\|_2)$. Folglich ist $\{\|\tilde{B}_k\|_2\}$ beschränkt.

(b) Es ist

$$\tilde{B}_+^{-1} - I = \left(I - \frac{\tilde{s}\tilde{s}^T}{\|\tilde{s}\|_2^2}\right)(\tilde{B}^{-1} - I)\left(I - \frac{\tilde{s}\tilde{s}^T}{\|\tilde{s}\|_2^2}\right).$$

(c) Es ist

$$\begin{aligned} \text{Spur}[(\tilde{B}_+^{-1} - I)^2] &= \text{Spur}[(\tilde{B}^{-1} - I)^2] - 2 \frac{\|(\tilde{B}^{-1} - I)\tilde{s}\|_2^2}{\|\tilde{s}\|_2^2} + \frac{[\tilde{s}^T(\tilde{B}^{-1} - I)\tilde{s}]^2}{\|\tilde{s}\|_2^4} \\ &\leq \text{Spur}[(\tilde{B}^{-1} - I)^2] - \frac{\|(\tilde{B}^{-1} - I)\tilde{s}\|_2^2}{\|\tilde{s}\|_2^2} \end{aligned}$$

und folglich

$$\sum_{k=0}^{\infty} \frac{\|(\tilde{B}_k^{-1} - I)\tilde{s}_k\|_2^2}{\|\tilde{s}_k\|_2^2} \leq \text{Spur}[(\tilde{B}_0^{-1} - I)^2], \quad \lim_{k \rightarrow \infty} \frac{\|(\tilde{B}_k^{-1} - I)\tilde{s}_k\|_2}{\|\tilde{s}_k\|_2} = 0.$$

(d) Es ist $x_{k+1} - x^* = Q^{-1/2}\tilde{B}_k(\tilde{B}_k^{-1} - I)\tilde{s}_k$ und daher

$$\begin{aligned} \frac{\|x_{k+1} - x^*\|_2}{\|x_k - x^*\|_2 + \|x_{k+1} - x^*\|_2} &\leq \frac{\|x_{k+1} - x^*\|_2}{\|\tilde{s}_k\|_2} \\ &= \frac{\|Q^{-1/2}\tilde{B}_k(\tilde{B}_k^{-1} - I)\tilde{s}_k\|_2}{\|Q^{-1/2}\tilde{s}_k\|_2} \\ &\leq \|Q^{-1/2}\|_2 \|Q^{1/2}\|_2 \|\tilde{B}_k\|_2 \frac{\|(\tilde{B}_k^{-1} - I)\tilde{s}_k\|_2}{\|\tilde{s}_k\|_2}. \end{aligned}$$

Weil hier die rechte Seite gegen Null konvergiert, ergibt sich die superlineare Konvergenz der Folge $\{x_k\}$ gegen x^* .

Ein entsprechendes Ergebnis gilt auch für das DFP-Verfahren (siehe J. E. DENNIS, J. J. MORÉ (1977, S. 84)).

5. Die Abbildung $f: \mathbb{R}^n \rightarrow \mathbb{R}$ sei auf einer offenen, konvexen Umgebung U^* eines Punktes x^* zweimal stetig differenzierbar. Es sei $\nabla f(x^*) = 0$, die Hessesche $\nabla^2 f(x^*)$ sei positiv definit. Sei $\{x_k\}$ eine gegen x^* konvergente Folge, $\{B_k\}$ eine Folge symmetrischer, positiv definiter Matrizen und $p_k := -B_k^{-1}\nabla f(x_k)$. Die ersten beiden Aussagen von Satz 3.7 seien gültig, d. h. die Folgen $\{\|B_k\|_2\}$ und $\{B_k^{-1}\|_2\}$ seien beschränkt und es sei

$$\lim_{k \rightarrow \infty} \frac{\|[B_k - \nabla^2 f(x^*)]p_k\|_2}{\|p_k\|_2} = 0.$$

Dann gilt:

- (a) Ist $t_k = t^*(x_k, p_k)$ die exakte Schrittweite, also die erste positive Lösung von $\nabla f(x_k + tp_k)^T p_k = 0$, so ist $\lim_{k \rightarrow \infty} t_k = 1$.

(b) Sind $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ vorgegeben, so ist

$$f(x_k + p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k, \quad \nabla f(x_k + p_k)^T p_k \geq \beta \nabla f(x_k)^T p_k$$

für alle hinreichend großen k . Wird also bei der Powell-Schrittweite t_k zunächst immer getestet, ob die Schrittweite $t = 1$ den beiden geforderten Ungleichungen genügt, so ist auch für diese $t_k = 1$ für alle hinreichend großen k .

6. In den letzten zwanzig Jahren sind neben den Verfahren der Broyden-Klasse, insbesondere also dem DFP- und dem BFGS-Verfahren, sehr viele weitere Update-Formeln vorgeschlagen worden. So wurde z. B. von S. S. OREN, D. G. LUENBERGER (1974) und S. S. OREN (1974) die folgende, sogenannte Oren-Luenberger-Klasse von Update-Formeln untersucht.

Sei $H \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit (hier hat man sich H als eine Approximation für $\nabla^2 f(x)^{-1}$ vorzustellen) und $y, s \in \mathbb{R}^n$ mit $y^T s > 0$. Mit zwei reellen Parametern ρ, θ sei

$$H_+ := \rho \left(H - \frac{(Hy)(Hy)^T}{y^T Hy} + \theta w w^T \right) + \frac{ss^T}{y^T s}, \quad w := (y^T Hy)^{1/2} \left(\frac{s}{y^T s} - \frac{Hy}{y^T Hy} \right).$$

Man zeige:

- (a) Es gilt die Quasi-Newton-Gleichung $H_+ y = s$.
- (b) H_+ ist genau dann positiv definit, wenn $\rho > 0$ und

$$\theta > -\frac{(y^T s)^2}{(s^T H^{-1} s)(y^T Hy) - (y^T s)^2}.$$

Insbesondere ist H_+ positiv definit, wenn $\rho > 0$ und $\theta \geq 0$.

- (c) Welche schon bekannten Update-Formeln erhält man für $\rho = 1$ und $\theta = 0$ bzw. $\theta = 1$?
- (d) Sei H_+ positiv definit. Zur Abkürzung sei

$$\epsilon := s^T H^{-1} s, \quad \sigma := y^T s, \quad \tau := y^T Hy.$$

Mit

$$\Lambda := -\frac{\rho}{2} + \frac{\sigma \epsilon + \rho \theta (\epsilon \tau - \sigma^2)}{2 \sigma^2}, \quad \Delta := \frac{\rho (\epsilon \tau - \sigma^2) [(1 - \theta) \sigma + \rho \theta \tau]}{\tau \sigma^2}$$

ist dann für alle $p \in \mathbb{R}^n$:

$$\min(\rho + \Lambda - (\Delta + \Lambda^2)^{1/2}, \rho) p^T H p \leq p^T H_+ p \leq \max(\rho + \Lambda + (\Delta + \Lambda^2)^{1/2}, \rho) p^T H p.$$

Daher ist $\text{cond}_2(H_+) \leq W(\rho, \theta) \text{cond}_2(H)$ mit

$$W(\rho, \theta) := \frac{\max(\rho + \Lambda + (\Delta + \Lambda^2)^{1/2}, \rho)}{\min(\rho + \Lambda - (\Delta + \Lambda^2)^{1/2}, \rho)}.$$

Hinweis: Bei den sogenannten *optimal konditionierten* Oren-Luenberger-Verfahren, siehe S. S. OREN, E. SPEDICATO (1976), bestimmt man (ρ, θ) so, daß $W(\rho, \theta)$ minimal ist. Es sei aber betont, daß diese Verfahren zwar theoretisch interessante Eigenschaften besitzen, sich aber gegen das BFGS-Verfahren in der Praxis nicht haben durchsetzen können.

7. Ist $f: \mathbb{R}^n \rightarrow \mathbb{R}$ eine quadratische, gleichmäßig konvexe Funktion mit einem (eindeutigen) Minimum bei $x^* \in \mathbb{R}^n$, ist also $f(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*) + f(x^*)$ mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n \times n}$, ist ferner $H \in \mathbb{R}^{n \times n}$ eine symmetrische, positiv definite Matrix, $p := -H\nabla f(x)$ die hierdurch gegebene Abstiegsrichtung in einer nichtstationären, aktuellen Näherung x und bezeichnet $t^* := -\nabla f(x)^T p / p^T Q p$ die exakte Schrittweite in x in Richtung p , so gilt

$$f(x + t^* p) - f(x^*) \leq \left(\frac{\text{cond}_2(Q^{1/2} H Q^{1/2}) - 1}{\text{cond}_2(Q^{1/2} H Q^{1/2}) + 1} \right)^2 [f(x) - f(x^*)].$$

Hinweis: Mit $z := H^{1/2} \nabla f(x)$ zeige man

$$\begin{aligned} \frac{f(x) - f(x^*)}{f(x) - f(x + t^* p)} &= \frac{[\nabla f(x)^T Q^{-1} \nabla f(x)] [\nabla f(x)^T H Q H \nabla f(x)]}{[\nabla f(x)^T H \nabla f(x)]^2} \\ &= \frac{[z^T (H^{1/2} Q H^{1/2})^{-1} z] [z^T H^{1/2} Q H^{1/2} z]}{(z^T z)^2}, \end{aligned}$$

wende die Ungleichung von Kantorowitsch (siehe Aufgabe 7 in Abschnitt 7.2) an und benutze, daß $H^{1/2} Q H^{1/2}$ und $Q^{1/2} H Q^{1/2}$ als ähnliche Matrizen dieselben Eigenwerte besitzen.

8. Ein Nachteil der Update-Formeln der Broyden-Klasse, also insbesondere des BFGS-Verfahrens, besteht darin, daß eine „Dünnbesetzung“ durch die Update-Formeln nicht vererbt wird. Genauer: Angenommen, für die Hessische $\nabla^2 f(\cdot)$ der Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ gilt $\nabla^2 f(x)_{ij} = 0$ für alle $(i, j) \in I$ und alle x , wobei $I \subset \{1, \dots, n\} \times \{1, \dots, n\}$. Für die erweiterte Rosenbrock-Funktion

$$f(x) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 100(x_4 - x_3^2)^2 + (1 - x_3)^2$$

in Aufgabe 9 (b) ist z. B. $I = \{(1, 3), (3, 1), (1, 4), (4, 1), (2, 3), (3, 2), (2, 4), (4, 2)\}$, d. h. die Hessische $\nabla^2 f(x)$ hat die Form

$$\nabla^2 f(x) = \begin{pmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{pmatrix}.$$

Es erscheint vorteilhaft, wenn die Matrizen der Folge $\{B_k\}$, die etwa durch das BFGS-Verfahren erzeugt werden, dieselbe „Dünnbesetzungssstruktur“ wie die Hessische besitzen. Aber selbst dann, wenn $B = B^T$ eine symmetrische Matrix mit $B_{ij} = 0$ für alle $(i, j) \in I$ ist, und $y, s \in \mathbb{R}^n$ Vektoren mit $y^T s > 0$ sind, wird für die BFGS-Update-Matrix

$$(*) \quad B_{\text{BFGS}} := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s}$$

diese Struktur verloren gehen. Es liegt daher nahe, die folgende Aufgabe zu untersuchen:

Sei $I \subset \{1, \dots, n\} \times \{1, \dots, n\}$ symmetrisch, d. h. mit $(i, j) \in I$ ist auch $(j, i) \in I$, ferner sei $(i, i) \notin I$ für $i = 1, \dots, n$. Sei $B \in \mathbb{R}^{n \times n}$ symmetrisch und $B_{ij} = 0$ für alle $(i, j) \in I$, weiter seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ vorgegeben und $B_{\text{BFGS}} \in \mathbb{R}^{n \times n}$ durch $(*)$ definiert.

Zu bestimmen ist die symmetrische Matrix B^* , welche die vorgeschriebene „Dünnbesetztheits-Struktur“ besitzt, der Quasi-Newton-Gleichung genügt, und bezüglich der Frobenius-Norm $\|\cdot\|_F$ einen minimalen Abstand zur BFGS-Update-Matrix B_{BFGS} besitzt, also eine Lösung der Aufgabe

$$\text{Minimiere } \|B^* - B_{\text{BFGS}}\|_F \text{ unter der Nebenbedingung } B^* \in K$$

mit

$$K := \{B^* \in \mathbb{R}^{n \times n} : B^* \text{ symmetrisch}, (B^*)_{ij} = 0 \text{ für alle } (i, j) \in I, B^* s = y\}.$$

Hinweis: Aufgaben dieser Art wurden von P.H. L. TOINT (1977) und D. F. SHANNO (1980) gelöst. Die folgenden Hinweise sollen einen Einblick in die Methoden und Ergebnisse dieser beiden Arbeiten geben. Die Aufgabe besteht darin, diese Hinweise auszuarbeiten.

Zur Abkürzung setze man $G := B_{\text{BFGS}}$, definiere

$$M := \{E \in \mathbb{R}^{n \times n} : E \text{ symmetrisch}, E_{ij} = -G_{ij} \text{ für alle } (i, j) \in I, Es = 0\}$$

und betrachte die Aufgabe

$$(P) \quad \text{Minimiere } \|E\|_F \text{ unter der Nebenbedingung } E \in M.$$

Ist E^* Lösung von (P), so ist $B^* := G + E^*$ Lösung des Ausgangsproblems. Ferner wissen wir: Ist der affine Raum $M \subset \mathbb{R}^{n \times n}$ nicht leer, so besitzt (P) eine eindeutige Lösung E^* und diese ist charakterisiert durch Spur $(E^* F) = 0$ für alle $F \in L$ mit

$$L := \{F \in \mathbb{R}^{n \times n} : F = F^T, F_{ij} = 0 \text{ für alle } (i, j) \in I, Fs = 0\}.$$

Für $i = 1, \dots, n$ seien die Vektoren $s(i) \in \mathbb{R}^n$ durch

$$s(i)_j := \begin{cases} s_j & \text{für } (i, j) \notin I, \\ 0 & \text{für } (i, j) \in I \end{cases} \quad (j = 1, \dots, n)$$

gegeben. Anschließend definiere man die Matrix $Q = (Q_{ij}) \in \mathbb{R}^{n \times n}$ durch

$$Q_{ij} := s(i)_j s(j)_i + \|s(i)\|_2^2 \delta_{ij}, \quad i, j = 1, \dots, n.$$

Dann ist Q eine symmetrische Matrix mit $Q_{ij} = 0$ für alle $(i, j) \in I$. Ferner gilt: Sind $s(1), \dots, s(n)$ vom Nullvektor verschieden, was im folgenden angenommen wird, so ist Q positiv definit. Man sollte versuchen, dieses nicht schwierige Resultat selbst zu beweisen, man kann aber auch P.H. L. TOINT (1977, Theorem 1) oder P. KOSMOL (1989, S. 160) konsultieren. Definiert man den Vektor $r = (r_i) \in \mathbb{R}^n$ durch $r_i := \sum_{j:(i,j) \in I} G_{ij} s_j$, $i = 1, \dots, n$, so ist daher das lineare Gleichungssystem $Q\lambda = r$ eindeutig durch den Vektor $\lambda = (\lambda_i) \in \mathbb{R}^n$ lösbar.

Man definiere die Matrix $E^* = (E_{ij}^*) \in \mathbb{R}^{n \times n}$ durch

$$E_{ij}^* := \begin{cases} -G_{ij} & \text{für } (i, j) \in I, \\ \lambda_i s_j + \lambda_j s_i & \text{für } (i, j) \notin I. \end{cases}$$

Dann ist $E^* \in M$, also E^* zulässig für (P). Hierzu bleibt nur die Gültigkeit von $E^*s = 0$ zu zeigen. Für $i = 1, \dots, n$ ist aber

$$(E^*s)_i = -r_i + \sum_{j:(i,j) \notin I} (\lambda_i s_j + \lambda_j s_i) s_j = -r_i + (Q\lambda)_i = 0.$$

Für beliebiges $F \in L$ ist

$$\text{Spur}(E^*F) = \sum_{i,j=1}^n E_{ij}^* F_{ij} = \sum_{i,j=1}^n (\lambda_i s_j + \lambda_j s_i) F_{ij} = \lambda^T F s + s^T F \lambda = 0,$$

also steht E^* (bezüglich des inneren Produktes auf $\mathbb{R}^{n \times n}$, welches die Frobenius-Norm erzeugt) senkrecht auf dem linearen Teilraum L , der parallel zum affinen Teilraum M liegt. Damit ist die Aussage bewiesen.

Ein gewichtiger Nachteil der vorgestellten Methode, die BFGS-Update-Matrix B_{BFGS} auf den affinen Teilraum K zu projizieren, um eine Matrix $B^* = B_{\text{BFGS}} + E^*$ zu erhalten, die symmetrisch ist, die vorgeschriebene „Dünnbesetztheits-Struktur“ besitzt, der Quasi-Newton-Gleichung genügt und bezüglich der Frobenius-Norm einen minimalem Abstand besitzt, besteht darin, daß B^* nicht mehr positiv definit zu sein braucht.

9. Man programmiere das gedämpfte Newton-Verfahren zur Lösung der unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

und teste das Programm an folgenden Zielfunktionen:

- (a) $f(x) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ (Rosenbrock-Funktion).
- (b) $f(x) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 100(x_4 - x_3^2)^2 + (1 - x_3)^2$ (Erweiterte Rosenbrock-Funktion).
- (c) $f(x) := x_1^4 + (x_1 + x_2)^2 + (\exp x_2 - 1)^2$ (siehe J. E. DENNIS, R. B. SCHNABEL (1983, S. 210)).
- (d) $f(x) := 100(x_1^2 - x_2)^2 + (1 - x_1)^2 + 90(x_3^2 - x_4)^2 + (1 - x_3)^2 + 10.1[(1 - x_2)^2 + (1 - x_4)^2] + 19.8(1 - x_2)(1 - x_4)$ (Wood-Funktion, siehe J. E. DENNIS, R. B. SCHNABEL (1983, S. 363)).

Hinweis: Man sollte jeweils die Anzahl der benötigten Funktionsauswertungen und die Anzahl der benötigten Iterationen zählen, die erforderlich sind, um eine geeignete Abbruchbedingung zu erfüllen. Bei der durchgeföhrten Rechnung wurde die Armijo-Schrittweite (in der Fassung von Han) mit $\alpha := 0.0001$ benutzt und als Abbruchbedingung $\|\nabla f(x)\|_\infty \leq 10^{-12}$ gewählt. In Tabelle 7.6 bedeuten Iter bzw. Fkt die Anzahl der benötigten Iterationen bzw. Zielfunktionsauswertungen. Bei Abbruch ist die Lösung jeweils auf zwölf Stellen genau berechnet. Insbesondere bei der Wood-Funktion sollte man mit den Startwerten einmal etwas spielen. Startet man z. B. mit $x_0 := (-1.0, 1.0, -1.0, 1.0)^T$ so wird man wahrscheinlich Schwierigkeiten bekommen, da in der Nähe einer „fast stationären“ Lösung liegt, in der die Hessesche einen negativen Eigenwert besitzt. Man sollte sich hiervon überzeugen, indem man mit $x_0 := (-1, 1, -1, 1)^T$ als Startwert das Gauß-Newton-Verfahren zur Minimierung von $\|\nabla f(x)\|_2$ anwendet, und in dem auf diese Weise erhaltenen Punkt x^* die Eigenwerte von $\nabla^2 f(x^*)$ (z. B. mit Hilfe des Jacobi-Verfahrens) berechnet.

Problem	Startwert	Iter	Fkt	Lösung
a)	(-1.2, 1.0)	26	35	(1, 1)
b)	(-1.2, 1.0, -1.2, 1.0)	26	35	(1, 1, 1, 1)
c)	(1.0, 1.0)	8	9	(0,0)
	(-1.0, 3.0)	13	14	
d)	(-1.5, -1.0, -3.0, -1.0)	36	49	(1,1,1,1)
	(-3.1, 8.2, 5.5, -3.5)	18	19	

Tabelle 7.6: Ergebnisse zu Aufgabe 9

10. Man programmiere das (gedämpfte) BFGS-Verfahren zur Lösung der unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

und teste das Programm an den Zielfunktionen aus Aufgabe 9. Hierbei sollte man das Programm so gestalten, daß die Cholesky-Zerlegung der Matrizen B_k „upgated“ wird.

Hinweis: Für die Startmatrix $B_0 = L_0 L_0^T$ wurden drei verschiedene Möglichkeiten ausprobiert.

- (a) $L_0 = I$, es wird also mit $B_0 = I$ gestartet. Am Anfang wird daher ein Gradientenschritt gemacht.
- (b) $L_0 = \sqrt{|f(x_0)|} I$, d. h. es wird $B_0 = |f(x_0)| I$ gesetzt (siehe J. E. DENNIS, R. B. SCHNABEL (1983, S. 209)).
- (c) $\nabla^2 f(x_0) = L_0 L_0^T$, d. h. am Anfang wird eine Cholesky-Zerlegung der Hesseschen $\nabla^2 f(x_0)$ gemacht (die natürlich nur dann existiert, wenn $\nabla^2 f(x_0)$ positiv definit ist).

Die Werte in der folgenden Tabelle beziehen sich auf eine Rechnung, bei der als Abbruchbedingung wieder $\|\nabla f(x)\|_\infty \leq 10^{-12}$ benutzt wurde. Ferner wurde stets $L_0 := \sqrt{|f(x_0)|} I$ (die anderen Möglichkeiten ergaben i. allg. keine besseren Ergebnisse) gesetzt. Wieder wurde mit der Armijo-Schrittweite mit $\alpha := 0.0001$ gerechnet. Ein Update erfolgte aber nur dann, wenn $y_k^T s_k = [\nabla f(x_{k+1}) - \nabla f(x_k)]^T (x_{k+1} - x_k) > 0$. Wieder war die Lösung bei Abbruch auf zwölf Stellen genau berechnet.

7.4 Verfahren der konjugierten Gradienten

Die in Abschnitt 7.3 untersuchten Quasi-Newton-Verfahren haben den Nachteil, daß sie Speicherplatz für eine Approximation $B_k \in \mathbb{R}^{n \times n}$ der Hesseschen $\nabla^2 f(x_k)$ der Zielfunktion f in der aktuellen Näherung x_k benötigen. Dies kann für großes n ein Problem werden. Zwar haben hochdimensionale Aufgaben i. allg. eine spezielle Struktur, so daß etwa die Hessesche dünn besetzt ist bzw. viele Nullen enthält, aber

Problem	Startwert	Iter	Fkt	Lösung
a)	(-1.2, 1.0)	38	46	(1, 1)
b)	(-1.2, 1.0, -1.2, 1.0)	54	65	(1, 1, 1, 1)
c)	(1.0, 1.0) (-1.0, 3.0)	13 20	14 21	(0, 0)
d)	(-1.5, -1.0, -3.0, -1.0) (-3.1, 8.2, 5.5, -3.5) (-1.0, 1.0, -1.0, 1.0)	44 114 77	47 133 96	(1, 1, 1, 1)

Tabelle 7.7: Ergebnisse zu Aufgabe 10

die auf diesen Fall zugeschnittenen Quasi-Newton-Verfahren (siehe z. B. Aufgabe 8 in 7.3) haben sich noch nicht durchsetzen können. Hier können Verfahren der konjugierten Gradienten in Betracht kommen, da bei ihnen lediglich Speicherplatz für einige Vektoren zur Verfügung stehen muß und sie trotzdem unrestrictierte Optimierungsaufgaben mit einer gleichmäßig konvexen, quadratischen Zielfunktion in endlich vielen Schritten lösen. Wir wollen in diesem Abschnitt einen Einblick in diese Klasse von Verfahren geben.

7.4.1 Quadratische Zielfunktionen

Wir betrachten die Aufgabe

$$(P) \quad \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x, \quad x \in \mathbb{R}^n,$$

wobei $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ stets als symmetrisch und positiv definit vorausgesetzt wird. (P) besitzt genau eine Lösung x^* , die durch $\nabla f(x^*) = c + Qx^* = 0$ charakterisiert ist. Daher können die Verfahren dieses Unterabschnittes auch zur Lösung linearer Gleichungssysteme mit symmetrischer, positiv definiter Koeffizientenmatrix eingesetzt werden.

In dem folgenden Satz wird das auf M. R. HESTENES, E. STIEFEL (1952) zurückgehende Verfahren der konjugierten Gradienten zur Lösung von (P) angegeben. Zuvor definieren wir aber

Definition 4.1 Ist $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, so heißen Vektoren $p_0, \dots, p_k \in \mathbb{R}^n$, $k < n$, *konjugiert bezüglich Q* oder auch *Q-konjugiert*, wenn sie vom Nullvektor verschieden sind und $p_i^T Q p_j = 0$ für $0 \leq i < j \leq k$ gilt.

Satz 4.2 Zur Lösung von (P) betrachte man das folgende Verfahren:

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := c + Qx_0$ und setze $p_0 := -g_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP, x_m ist die Lösung von (P).

Berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T Q p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := g_k + t_k Q p_k.$$

Berechne

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \quad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Dann gilt:

1. Das Verfahren bricht nach $m \leq n$ Schritten ab.
2. Es ist $p_i^T g_k = 0$ für $0 \leq i < k \leq m$.
3. Es ist $g_k^T p_k = -\|g_k\|_2^2$ für $0 \leq k \leq m$.
4. Es ist $g_i^T g_k = 0$ für $0 \leq i < k \leq m$.
5. Die Richtungen p_0, \dots, p_{m-1} sind Q -konjugiert.
6. Es ist $\text{span}\{p_0, \dots, p_k\} = \text{span}\{g_0, \dots, g_k\}$ für $0 \leq k < m$.

Beweis: Wir zeigen durch vollständige Induktion nach k : Sind $g_0, \dots, g_k \neq 0$, wird das Verfahren also im k -ten Schritt noch nicht abgebrochen, so gilt

- (a) $p_i^T g_k = 0$ für $0 \leq i < k$,
- (b) $g_k^T p_k = -\|g_k\|_2^2$.
- (c) $g_i^T g_k = 0$ für $0 \leq i < k$,
- (d) p_0, \dots, p_k sind Q -konjugiert,
- (e) $\text{span}\{p_0, \dots, p_k\} = \text{span}\{g_0, \dots, g_k\}$.

Diese fünf Aussagen sind für $k = 0$ (beachte: $p_0 := -g_0$) trivialerweise richtig. Für den Induktionsschluß nehmen wir an, g_0, \dots, g_{k+1} seien von Null verschieden.

Für $0 \leq i < k$ ist $p_i^T g_{k+1} = p_i^T (g_{k+1} + t_k Q p_k) = 0$, wobei die Induktionsvoraussetzungen (a) und (d) benutzt wurden. Ferner ist $p_k^T g_{k+1} = 0$ nach Definition der (exakten) Schrittweite t_k . Damit ist der Induktionsschluß für (a) vollzogen.

Es ist $g_{k+1}^T p_{k+1} = g_{k+1}^T (-g_{k+1} + \beta_k p_k) = -\|g_{k+1}\|_2^2$ wegen des gerade eben für $k+1$ bewiesenen Teils (a). Damit ist auch (b) für $k+1$ bewiesen.

Für $1 \leq i < k$ ist $g_i^T g_{k+1} = (\beta_{i-1} p_{i-1} - p_i)^T g_{k+1} = 0$, wobei der für $k+1$ schon bewiesene Teil (a) benutzt wurde, aus welchem auch $g_0^T g_{k+1} = -p_0^T g_{k+1} = 0$ folgt. Berücksichtigt man nun noch, daß $g_k^T Q p_k = (\beta_{k-1} p_{k-1} - p_k)^T Q p_k = -p_k^T Q p_k$, so erhält man

$$g_k^T g_{k+1} = g_k^T (g_k + t_k Q p_k) = \|g_k\|_2^2 - \frac{g_k^T p_k}{p_k^T Q p_k} g_k^T Q p_k = \|g_k\|_2^2 - \frac{\|g_k\|_2^2}{p_k^T Q p_k} p_k^T Q p_k = 0,$$

wobei auch noch die Induktionsvoraussetzung (b) verwandt wurde. Damit ist auch (c) für $k+1$ bewiesen.

Wegen der schon für $k+1$ bewiesenen Aussage (b) ist mit g_{k+1} auch p_{k+1} vom Nullvektor verschieden. Für $0 \leq i < k$ ist

$$p_i^T Q p_{k+1} = p_i^T Q (-g_{k+1} + \beta_k p_k) = -g_{k+1}^T Q p_i = \frac{1}{t_i} g_{k+1}^T (g_i - g_{i+1}) = 0.$$

Da schließlich

$$p_{k+1}^T Q p_k = (-g_{k+1} + \beta_k p_k)^T \frac{1}{t_k} (g_{k+1} - g_k) = \frac{1}{t_k} (-\|g_{k+1}\|_2^2 + \beta_k \|g_k\|_2^2) = 0,$$

ist auch (d) für $k+1$ richtig.

Wegen $p_{k+1} = -g_{k+1} + \beta_k p_k$ und der Induktionsvoraussetzung (e) folgt sofort, daß auch $\text{span} \{p_0, \dots, p_{k+1}\} = \text{span} \{g_0, \dots, g_{k+1}\}$. Damit ist der Induktionsbeweis abgeschlossen.

Insbesondere ist bewiesen worden, daß das Verfahren Q -konjugierte Richtungen erzeugt, solange es nicht abbriicht. Da Q -konjugierte Richtungen linear unabhängig sind, kann es von ihnen nicht mehr als n geben. Der Satz ist daher vollständig bewiesen. \square

Bemerkungen: Die exakte Schrittweite t_k kann im Verfahren der konjugierten Gradienten auch durch $t_k := \|g_k\|_2^2 / p_k^T Q p_k$ (siehe die dritte Behauptung in Satz 4.2) berechnet werden.

Ein Vorteil des Verfahrens der konjugierten Gradienten zur Lösung der unrestriktierten Optimierungsaufgabe (P) (mit der quadratischen, gleichmäßig konvexen Zielfunktion $f(x) := c^T x + \frac{1}{2} x^T Q x$) bzw. des äquivalenten linearen Gleichungssystems $c + Qx = 0$ (mit der symmetrischen, positiv definiten Koeffizientenmatrix Q) gegenüber Eliminationsmethoden besteht darin, daß die Matrix Q in jedem Iterationsschritt nur dadurch eingreift, daß ihre *Wirkung* auf die aktuelle Richtung p_k zu bestimmen ist. Genau das macht das Verfahren attraktiv für hochdimensionale, dünn besetzte Aufgaben vom angegebenen Typ. \square

Wendet man das Gradientenverfahren mit exakter Schrittweitenstrategie auf die Optimierungsaufgabe (P) mit der Zielfunktion $f(x) := c^T x + \frac{1}{2} x^T Q x$ an, wobei Q eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert λ_{\min} und größtem Eigenwert λ_{\max} ist, so konnte in Aufgabe 9b in 7.2 mit Hilfe der Ungleichung von Kantorowitsch gezeigt werden, daß

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 [f(x_k) - f(x^*)], \quad k = 0, 1, \dots,$$

wobei natürlich $x^* := -Q^{-1}c$ die Lösung von (P) bedeutet. Berücksichtigt man $\text{cond}_2(Q) = \lambda_{\max}/\lambda_{\min}$ und führt man die Norm $\|\cdot\|_Q$ durch $\|x\|_Q := \sqrt{x^T Q x}$ auf dem \mathbb{R}^n ein, so erkennt man, daß dieses Ergebnis in

$$\|x_{k+1} - x^*\|_Q \leq \left(\frac{\text{cond}_2(Q) - 1}{\text{cond}_2(Q) + 1} \right) \|x_k - x^*\|_Q, \quad k = 0, 1, \dots$$

übergeht. Hieraus wiederum folgt

$$\frac{\|x_k - x^*\|_Q}{\|x_0 - x^*\|_Q} \leq \left(\frac{\text{cond}_2(Q) - 1}{\text{cond}_2(Q) + 1} \right)^k, \quad k = 0, 1, \dots$$

Hierdurch wird die Aussage quantifiziert: Je kleiner die Kondition von Q , desto besser die Konvergenz des Gradientenverfahrens.

Eine entsprechende Aussage gilt auch für das in Satz 4.2 angegebene Verfahren der konjugierten Gradienten. Hier kann gezeigt werden (siehe J. STOER, R. BULIRSCH (1990, S. 301)), daß

$$\frac{\|x_k - x^*\|_Q}{\|x_0 - x^*\|_Q} \leq 2 \left(\frac{\sqrt{\text{cond}_2(Q)} - 1}{\sqrt{\text{cond}_2(Q)} + 1} \right)^k, \quad k = 0, 1, \dots$$

Dies ist eine bessere Abschätzung als die entsprechende für das Gradientenverfahren. Sie besagt, daß für das Verfahren der konjugierten Gradienten „gute“ Konvergenz zu erwarten ist, wenn die Kondition von Q klein ist. Durch eine sogenannte *Präkonditionierung* kann man versuchen, ein äquivalentes Problem zu lösen, bei welchem die Hessesche der Zielfunktion eine kleinere Kondition besitzt. Die Idee hierzu besteht darin, mit einer symmetrischen, positiv definiten Matrix B zu dem transformierten Problem

$$(P_B) \quad \text{Minimiere } \phi(y) := (B^{-1/2}c)^T y + \frac{1}{2} y^T B^{-1/2} Q B^{-1/2} y, \quad y \in \mathbb{R}^n$$

überzugehen. Dann ist

$$\phi(y) = f(B^{-1/2}y), \quad \nabla \phi(y) = B^{-1/2} \nabla f(B^{-1/2}y), \quad \nabla^2 \phi(y) = B^{-1/2} Q B^{-1/2}.$$

Ist y^* die Lösung von (P_B) , so ist $x^* := B^{-1/2}y^*$ die Lösung von (P) . Nun wende man auf (P_B) das Verfahren der konjugierten Gradienten an. Dies führt auf:

- Wähle $y_0 \in \mathbb{R}^n$, berechne $h_0 := B^{-1/2}(c + QB^{-1/2}y_0)$ und setze $q_0 := -h_0$.
- Für $k = 0, 1, \dots$:

Falls $h_k = 0$, dann: $m := k$, STOP, y_m ist die Lösung von (P_B) , daher ist $x_m := B^{-1/2}y_m$ die Lösung von (P) .

Berechne

$$t_k := \frac{\|h_k\|_2^2}{q_k^T B^{-1/2} Q B^{-1/2} q_k}, \quad y_{k+1} := y_k + t_k q_k$$

sowie

$$h_{k+1} := h_k + t_k B^{-1/2} Q B^{-1/2} q_k.$$

Berechne

$$\beta_k := \frac{\|h_{k+1}\|_2^2}{\|h_k\|_2^2}, \quad q_{k+1} := -h_{k+1} + \beta_k q_k.$$

In dieser Form ist der Algorithmus natürlich nicht brauchbar, da man nicht bereit ist, $B^{-1/2}$ zu berechnen. Setzt man aber $x_k := B^{-1/2}y_k$, $g_k := B^{1/2}h_k$ und $p_k := B^{-1/2}q_k$, so erhält man das *Verfahren der konjugierten Gradienten mit Prädiktionskonditionierung*:

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := c + Qx_0$ und $p_0 := -B^{-1}g_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP, x_m ist die Lösung von (P).

Berechne Qp_k und anschließend

$$t_k := \frac{g_k^T B^{-1} g_k}{p_k^T Q p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := g_k + t_k Q p_k.$$

Berechne $B^{-1}g_{k+1}$ und anschließend

$$\beta_k := \frac{g_{k+1}^T B^{-1} g_{k+1}}{g_k^T B^{-1} g_k}, \quad p_{k+1} := -B^{-1}g_{k+1} + \beta_k p_k.$$

Der Unterschied in der Komplexität zum gewöhnlichen Verfahren der konjugierten Gradienten besteht darin, daß in jedem Schritt $B^{-1}g_k$ zu berechnen, also ein lineares Gleichungssystem mit B als Koeffizientenmatrix zu lösen ist.

Da die Konvergenzgeschwindigkeit des Verfahrens der konjugierten Gradienten mit Prädiktionskonditionierung durch die Kondition der Matrix $B^{-1/2}QB^{-1/2}$ bestimmt ist, wird man an die symmetrische, positiv definite Matrix B die Forderung stellen, daß einerseits Gleichungssysteme mit B als Koeffizientenmatrix „leicht“ lösbar sind, und andererseits $\text{cond}_2(B^{-1/2}QB^{-1/2})$ möglichst klein ist. Nun sind $B^{-1/2}QB^{-1/2}$ und $B^{-1}Q$ ähnlich, die Eigenwerte dieser Matrizen stimmen also überein. Daher wird $\text{cond}_2(B^{-1/2}QB^{-1/2})$ klein sein, wenn der Spektralradius $\rho(I - B^{-1}Q)$ von $I - B^{-1}Q$ klein ist. Genau diese Problemstellung trat bei Iterationsverfahren für lineare Gleichungssysteme (siehe Abschnitt 2.4) auf, wobei die Situation hier dadurch komplizierter wird, daß mit Q auch B symmetrisch und positiv definit zu sein hat. Geht man wie in 2.4 von der Zerlegung $Q = Q_D + Q_L + Q_R$ aus, wobei Q_D die aus den Diagonalen von Q gewonnene Diagonalmatrix und Q_L bzw. Q_R den durch Nullen aufgefüllten strikten unteren bzw. oberen Anteil von Q bedeuten (wegen der Symmetrie von Q ist $Q_R = Q_L^T$), so sollte man die beiden folgenden Möglichkeiten für die Wahl von B in Betracht ziehen:

- (a) Setze $B := Q_D$. Dies entspricht der Zerlegung $Q = B + (Q - B)$, die dem Gesamtschrittverfahren zugrunde liegt.
- (b) Mit einem $\omega \in (0, 2)$ setze man

$$B := (Q_D + \omega Q_L) Q_D^{-1} (Q_D + \omega Q_L)^T.$$

Dann ist B symmetrisch und positiv definit. Ferner lassen sich lineare Gleichungssysteme mit B als Koeffizientenmatrix offenbar leicht durch Vorwärts- und Rückwärtseinsetzen lösen, wobei erfreulicherweise mit Q auch die hierfür benötigten Faktoren $L := Q_D + \omega Q_L$ und L^T dünn besetzt sind. Die Zerlegung $Q = B + (Q - B)$ liegt einer symmetrischen Version des SOR-Verfahrens zugrunde, dem sogenannten SSOR-Verfahren.

Für weitere Hinweise zur Prækonditionierung des Verfahrens der konjugierten Gradienten sei auf G. H. GOLUB, C. F. VAN LOAN (1989, S. 527 ff.) und O. AXELSSON (1985) verwiesen.

7.4.2 Das Fletcher-Reeves-Verfahren

Von R. FLETCHER, C. M. REEVES (1964) stammt eine erste Verallgemeinerung des Verfahrens der konjugierten Gradienten auf unrestringierte Optimierungsaufgaben

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

mit nicht notwendig quadratischer Zielfunktion f . Wir erinnern an die Voraussetzungen (V) (a)–(c) aus 7.2.1 und die (gleichmäßigen) Konvexitätsvoraussetzungen (K) (a)–(c) aus Lemma 2.6, die z. B. auch dem globalen Konvergenzsatz für das BFGS-Verfahren zugrunde lagen. Im folgenden Satz wird das Fletcher-Reeves-Verfahren mit exakter Schrittweitenstrategie angegeben und unter den Voraussetzungen (V) (a)–(c) bzw. (K) (a)–(c) eine globale Konvergenzaussage bewiesen. Auf einen entsprechenden Konvergenzsatz mit einer inexakten Schrittweitenstrategie wird in Aufgabe 5 eingegangen (siehe auch M. AL-BAALI (1985)).

Satz 4.3 Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Zielfunktion $f: \mathbb{R}^n \rightarrow \mathbb{R}$ genüge den Voraussetzungen (V) (a)–(c). Das Verfahren der konjugierten Gradienten von Fletcher-Reeves ist durch den folgenden Algorithmus gegeben:

- Setze $g_0 := \nabla f(x_0)$ und $p_0 := -g_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: STOP, x_k ist stationäre Lösung von (P).

Sei $t_k := t^*(x_k, p_k)$ die exakte Schrittweite in x_k in Richtung p_k , also die erste positive Nullstelle von $\nabla f(x_k + tp_k)^T p_k$.

Berechne

$$x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := \nabla f(x_{k+1})$$

sowie

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \quad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Dann gilt: Bricht das Verfahren nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$, wenigstens ein Häufungspunkt von $\{x_k\}$ ist also eine stationäre Lösung von (P). Sind sogar die Konvexitätsvoraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

Beweis: Zunächst beachten wir: Da im Verfahren stets die exakte Schrittweite gewählt wird, ist $g_k^T p_k = -\|g_k\|_2^2 < 0$ für $g_k \neq 0$, also p_k eine Abstiegsrichtung in x_k . Das Verfahren breche nicht vorzeitig mit einer stationären Lösung ab. Im Widerspruch zur Behauptung nehmen wir an, es sei $\liminf_{k \rightarrow \infty} \|g_k\|_2 > 0$. Dann

existiert ein $\epsilon > 0$ mit $\|g_k\| \geq \epsilon$ für alle k . Wegen Satz 2.2 gibt es eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{g_k^T p_k}{\|p_k\|_2} \right)^2 = \theta \frac{\|g_k\|_2^4}{\|p_k\|_2^2} = \frac{\theta}{\alpha_k} \quad \text{mit } \alpha_k := \frac{\|p_k\|_2^2}{\|g_k\|_2^4}.$$

Für $k \geq 1$ ist

$$\alpha_k = \frac{\|p_k\|_2^2}{\|g_k\|_2^4} = \frac{\|g_k\|_2^2 + \beta_{k-1}^2 \|p_{k-1}\|_2^2}{\|g_k\|_2^4} = \frac{1}{\|g_k\|_2^2} + \alpha_{k-1}.$$

Durch Zurückspulen erhält man

$$\alpha_k = \sum_{j=1}^k \frac{1}{\|g_j\|_2^2} + \alpha_0 = \sum_{j=0}^k \frac{1}{\|g_j\|_2^2} \leq \frac{k+1}{\epsilon^2}, \quad k = 0, 1, \dots,$$

und hieraus

$$(*) \quad f(x_k) - f(x_{k+1}) \geq \frac{\theta \epsilon^2}{k+1}, \quad k = 0, 1, \dots$$

Die harmonische Reihe ist bekanntlich divergent. Daher folgt aus $(*)$, daß $\{f(x_k)\}$ nicht nach unten beschränkt ist, was einen Widerspruch zur vorausgesetzten Kompaktheit der Niveaumenge L_0 darstellt.

Nun seien sogar die Voraussetzungen (K) (a)–(c) erfüllt. Ein $\epsilon > 0$ sei vorgegeben. Wegen der unter den schwächeren Voraussetzungen (V) (a)–(c) bewiesenen Aussage existiert ein $k_0 \in \mathbb{N}$ mit $\|g_{k_0}\| \leq c\epsilon$. Eine Anwendung von Lemma 2.6 liefert für alle $k \geq k_0$ die Ungleichungskette

$$\frac{c}{2} \|x_k - x^*\|_2^2 \leq f(x_k) - f(x^*) \leq f(x_{k_0}) - f(x^*) \leq \frac{1}{2c} \|g_{k_0}\|_2^2 \leq \frac{c\epsilon^2}{2}.$$

Daher ist $\|x_k - x^*\| \leq \epsilon$ für alle $k \geq k_0$, auch der zweite Teil ist bewiesen. \square

Bemerkungen: Spezialisiert man das Fletcher-Reeves-Verfahren auf eine quadratische Zielfunktion, so erhält man genau das in 7.4.1 untersuchte Verfahren der konjugierten Gradienten. Insbesondere bricht das Verfahren in diesem Falle nach $m \leq n$ Schritten ab.

Es gibt einige Varianten zum Fletcher-Reeves-Verfahren. Diese unterscheiden sich im wesentlichen in der Definition von β_k , wodurch $p_{k+1} := -g_{k+1} + \beta_k p_k$ bestimmt wird, reduzieren sich für eine quadratische Zielfunktion aber stets auf das Verfahren der konjugierten Gradienten aus 7.4.1. So setzen z. B. E. POLAK, G. RIBIÈRE (1969)

$$\beta_k := \frac{g_{k+1}^T (g_{k+1} - g_k)}{\|g_k\|_2^2}.$$

Konvergenzaussagen zum Polak-Ribièvre-Verfahren werden in Aufgabe 4 gemacht.

Aussagen⁵ über die Konvergenzgeschwindigkeit der Verfahren der konjugierten Gradienten sind schwierig zu beweisen. Macht man im Polak-Ribière-Verfahren alle n Schritte einen sogenannten restart, indem man wieder mit der negativen Gradientenrichtung in der aktuellen Näherung beginnt, so kann z. B. unter geeigneten Voraussetzungen die n -Schritt superlineare Konvergenz des Polak-Ribière-Verfahrens mit asymptotisch exakter Schrittweitenstrategie nachgewiesen werden (siehe G. P. MCCORMICK, K. RITTER (1974)).

Ein restart, etwa alle n Schritte oder noch häufiger, sollte bei einer Implementation eines Verfahrens der konjugierten Gradienten auf alle Fälle eingebaut werden. In der einfachsten Version wählt man in einem solchen Fall $p_k := -g_k$. Raffiniertere Methoden werden von M. J. D. POWELL (1977) untersucht. \square

7.4.3 Ein gedächtnisloses BFGS-Verfahren

In diesem Unterabschnitt gehen wir auf ein von D. F. SHANNO (1978a, 1978b) angegebenes Verfahren zur Lösung der unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n$$

ein, das man auch als ein gedächtnisloses BFGS-Verfahren bezeichnen könnte. In diesem Verfahren werden die Abstiegsrichtungen sehr ähnlich wie beim BFGS-Verfahren erzeugt, allerdings mit dem entscheidenden Unterschied, daß keine Update-Matrix gespeichert wird. Zunächst soll dieses Verfahren motiviert werden, anschließend wird ein globaler Konvergenzsatz bewiesen.

Ein Schritt des BFGS-Verfahrens sieht bekanntlich folgendermaßen aus (hierbei schreiben wir H statt B^{-1} und entsprechend H_+ statt B_+^{-1}):

- Sei $x \in \mathbb{R}^n$ eine aktuelle Näherung, $g := \nabla f(x) \neq 0$, $H \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $p := -Hg$.
- Mit einer geeigneten Schrittweite $t > 0$ sei $x_+ := x + tp$. Mit $g_+ := \nabla f(x_+)$, $s := x_+ - x$ und $y := g_+ - g$ berechne man die neue Matrix H_+ aus H (unter der Voraussetzung $y^T s > 0$) durch die Update-Formel

$$H_+ := H + \left(1 + \frac{y^T H y}{y^T s}\right) \frac{s s^T}{y^T s} - \frac{s (H y)^T + (H y) s^T}{y^T s}$$

(siehe Satz 3.3). Die neue Richtung p_+ ist dann durch $p_+ := -H_+ g_+$ gegeben.

Ist hier $H = \rho I$ mit $\rho > 0$, so wird die neue Richtung $p_+ := -H_+ g_+$ aus

$$(*) \quad p_+ := -\rho g_+ - \left(1 + \rho \frac{\|y\|_2^2}{y^T s}\right) \frac{s^T g_+}{y^T s} s + \frac{\rho}{y^T s} [(y^T g_+) s + (s^T g_+) y]$$

⁵G. P. MCCORMICK (1983, S. 173) schreibt völlig zutreffend: "Theorems concerning the convergence and rate of convergence of the conjugate gradient method on nonquadratic minimization problems have been forthcoming only in recent years. There is considerable confusion in many cases about which version of the conjugate gradient method is referenced, what conditions are placed on the function minimized, what the conclusions of the theorems are, and the validity of the proofs."

berechnet. Bei der Richtungsstrategie von Shanno wird $\rho := y^T s / \|y\|_2^2$ gesetzt. Eine Motivation hierfür könnte in folgendem bestehen: Ist $f(x) := c^T x + \frac{1}{2} x^T Q x$, wobei $Q = \alpha I$ ein positives Vielfaches der Einheitsmatrix ist, so erhält man

$$\rho = \frac{y^T s}{\|y\|_2^2} = \frac{s^T Q s}{\|Q s\|_2^2} = \frac{1}{\alpha}.$$

In diesem Falle ist also $H^{-1} = Q$ die Hessesche von f . Einsetzen von $\rho := y^T s / \|y\|_2^2$ in (*) und Umordnen ergibt die Vorschrift

$$p_+ := -\frac{y^T s}{\|y\|_2^2} g_+ - \left(2 \frac{s^T g_+}{y^T s} - \frac{y^T g_+}{\|y\|_2^2} \right) s + \frac{s^T g_+}{\|y\|_2^2} y.$$

Genau dies ist die Richtungsstrategie bei D. F. SHANNO (1978a), wobei am Anfang, wie stets bei den Verfahren der konjugierten Gradienten, $p_0 := -g_0$ gesetzt wird. Man beachte: Wird beim Übergang von x zu $x_+ := x + tp$ die exakte Schrittweite gewählt, so ist $s^T g_+ = 0$ und $y^T s = -t g^T p$, so daß sich die neue Richtung p_+ auf

$$p_+ := \frac{y^T s}{\|y\|_2^2} \left(-g_+ - \frac{y^T g_+}{g^T p} p \right)$$

reduziert. Die Verwandtschaft zum Polak-Ribiére-Verfahren ist offensichtlich.

Nun geben wir den angekündigten globalen Konvergenzsatz für das Shanno-Verfahren an (siehe D. F. SHANNO (1978a, Theorem 4), dort wird allerdings nur die Powell-Schrittweite betrachtet).

Satz 4.4 Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (K) (a)–(c) seien erfüllt. Das Shanno-Verfahren mit nicht notwendig exakter Schrittweitenstrategie ist durch den folgenden Algorithmus gegeben:

- Setze $g_0 := \nabla f(x_0)$ und $p_0 := -g_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: STOP, x_k ist die Lösung von (P).

Sei t_k die exakte Schrittweite, eine Powell- oder eine Armijo-Schrittweite in x_k in Richtung p_k .

Berechne

$$x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := \nabla f(x_{k+1}),$$

setze

$$s_k := x_{k+1} - x_k, \quad y_k := g_{k+1} - g_k$$

und berechne

$$p_{k+1} := -\frac{y_k^T s_k}{\|y_k\|_2^2} g_{k+1} - \left(2 \frac{s_k^T g_{k+1}}{y_k^T s_k} - \frac{y_k^T g_{k+1}}{\|y_k\|_2^2} \right) s_k + \frac{s_k^T g_{k+1}}{\|y_k\|_2^2} y_k.$$

Dann gilt: Bricht das Verfahren nicht nach endlich vielen Schritten mit der Lösung x^* von (P) ab, so liefert es eine Folge $\{x_k\}$, die gegen x^* konvergiert. Genauer existieren in diesem Falle Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\|_2 \leq C q^k$ für $k = 0, 1, \dots$

Beweis: Zunächst müssen wir uns Klarheit darüber verschaffen, daß das Shanno-Verfahren überhaupt durchführbar ist, insbesondere darüber, daß die berechneten Richtungen Abstiegsrichtungen sind.

Es sei $g_k^T p_k < 0$, also p_k eine Abstiegsrichtung in x_k . Wenn nicht der erfreuliche, aber unwahrscheinliche Fall vorliegt, daß der Startpunkt x_0 schon die Lösung ist, so ist das für $k = 0$ wegen $p_0 := -g_0$ der Fall. Dann ist aber $y_k^T s_k \geq c \|s_k\|_2^2 > 0$ wegen der gleichmäßigen Konvexität von f . Wie wir uns bei der Motivation für das Shanno-Verfahren überlegten, ist $p_{k+1} = -H_{k+1} g_{k+1}$ mit

$$H_{k+1} := \rho_k I + \left(1 + \rho_k \frac{\|y_k\|_2^2}{y_k^T s_k}\right) \frac{s_k s_k^T}{y_k^T s_k} - \frac{\rho_k}{y_k^T s_k} (s_k y_k^T + y_k s_k^T), \quad \rho_k := \frac{y_k^T s_k}{\|y_k\|_2^2}.$$

Beachtet man nun $\rho_k > 0$ und die Verwandtschaft mit dem BFGS-Verfahren, so erkennt man (siehe Satz 3.3), daß H_{k+1} symmetrisch und positiv definit ist, daß ferner die Inverse zu H_{k+1} durch

$$H_{k+1}^{-1} = \frac{1}{\rho_k} I - \frac{s_k s_k^T}{\rho_k \|s_k\|_2^2} + \frac{y_k y_k^T}{y_k^T s_k}$$

gegeben ist. Insbesondere gilt: Ist $g_{k+1} \neq 0$, so ist $g_{k+1}^T p_{k+1} = -g_{k+1}^T H_{k+1} g_{k+1} < 0$ und damit auch p_{k+1} eine Abstiegsrichtung.

Für den eigentlichen Konvergenzbeweis wenden wir Satz 2.7 an und setzen, wie dort,

$$\delta_k := \min \left[-\frac{g_k^T p_k}{\|g_k\|_2^2}, \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2 \right].$$

Wegen $p_k = -H_k g_k$ (für $k = 0$ ist natürlich $H_0 := I$) ist

$$\delta_k = \min \left[\frac{g_k^T H_k g_k}{\|g_k\|_2^2}, \left(\frac{g_k^T H_k g_k}{\|g_k\|_2 \|H_k g_k\|_2} \right)^2 \right] \geq \min \left[\lambda_{\min}(H_k), \frac{\lambda_{\min}(H_k)}{\lambda_{\max}(H_k)} \right],$$

wobei $\lambda_{\max}(H_k) = \|H_k\|_2$ größter und $\lambda_{\min}(H_k) = 1/\|H_k^{-1}\|_2$ kleinster Eigenwert von H_k ist (siehe auch eine Bemerkung im Anschluß an Satz 2.5). Daher existiert ein $\delta > 0$ mit $\delta_k \geq \delta$ für $k = 0, 1, \dots$, wenn die Folgen $\{\|H_k\|_2\}$ und $\{\|H_k^{-1}\|_2\}$ beschränkt sind. Genau das bleibt daher zu zeigen.

Für die Abschätzung von $\|H_k\|_2$ und $\|H_k^{-1}\|_2$ benutzen wir, daß für $u, v \in \mathbb{R}^n$ nach einfacher Rechnung $\|uv^T\|_2 = \|u\|_2 \|v\|_2$ gilt. Ferner beachten wir, daß wegen der gleichmäßigen Konvexität der Zielfunktion f die Ungleichungen

$$c \|s_k\|_2^2 \leq y_k^T s_k \leq \|y_k\|_2 \|s_k\|_2$$

gelten. Berücksichtigt man noch die Lipschitzstetigkeit des Gradienten von f , so ist

$$c \|s_k\|_2 \leq \|y_k\|_2 \leq \gamma \|s_k\|_2.$$

Daher ist

$$\|H_{k+1}\|_2 \leq \frac{y_k^T s_k}{\|y_k\|_2^2} + 2 \frac{\|s_k\|_2^2}{y_k^T s_k} + 2 \frac{\|s_k\|_2}{\|y_k\|_2} \leq \frac{5}{c}$$

und

$$\|H_{k+1}^{-1}\|_2 \leq \frac{\|y_k\|_2^2}{y_k^T s_k} + \frac{\|y_k\|_2^2}{y_k^T s_k} + \frac{\|y_k\|_2^2}{y_k^T s_k} \leq \frac{3\gamma^2}{c}.$$

Insgesamt ist der Satz damit bewiesen. \square

Zum Schluß dieses Unterabschnittes soll ein weiterer Konvergenzsatz (siehe D. F. SHANNO (1978a, Theorem 5)) zum Shanno-Verfahren bewiesen werden. In diesem wird, wie in Satz 4.3, auf die Voraussetzungen (K) (a)–(c) verzichtet, statt dessen wird (V) (a)–(c) aus 7.2.1 vorausgesetzt. Das folgende Lemma wird sich als wichtig herausstellen.

Lemma 4.5 Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ gegeben und hiermit die symmetrische, positiv definite Matrix $H_+ \in \mathbb{R}^{n \times n}$ durch

$$H_+ := \frac{y^T s}{\|y\|_2^2} I + 2 \frac{ss^T}{y^T s} - \frac{sy^T + ys^T}{\|y\|_2^2}$$

definiert. Für alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\frac{x^T H_+ x}{\|x\|_2 \|H_+ x\|_2} \geq \frac{y^T s}{\|y\|_2 \|s\|_2}.$$

Beweis: O. B. d. A. sind y und s linear unabhängig, da andernfalls s ein positives Vielfaches von y ist, wofür die Aussage offensichtlich richtig ist. Dann ist $y^T s / \|y\|_2^2$ ein $(n-2)$ -facher Eigenwert von H_+ mit zugehörigen Eigenvektoren aus $\text{span}\{y, s\}^\perp$. Die beiden restlichen Eigenwerte seien aus Gründen, die gleich klar sein werden, mit $\lambda_{\min}(H_+)$ und $\lambda_{\max}(H_+)$ bezeichnet, es sei $\lambda_{\min}(H_+) \leq \lambda_{\max}(H_+)$. Unser erstes Ziel besteht im Nachweis von

$$(*) \quad \lambda_{\min}(H_+) \leq \frac{y^T s}{\|y\|_2^2} \leq \lambda_{\max}(H_+).$$

Bezeichnet man mit $\{x_{\min}, x_2, \dots, x_{n-1}, x_{\max}\}$ ein Orthonormalsystem von Eigenvektoren zu $\lambda_{\min}(H_+)$, dem $(n-2)$ -fachen Eigenwert $y^T s / \|y\|_2^2$ und $\lambda_{\max}(H_+)$, so ist

$$\text{span}\{x_2, \dots, x_{n-1}\} = \text{span}\{y, s\}^\perp, \quad \text{span}\{x_{\min}, x_{\max}\} = \text{span}\{y, s\}.$$

Insbesondere läßt sich y eindeutig in der Form $y = \alpha x_{\min} + \beta x_{\max}$ darstellen. Benutzt man nun noch die Quasi-Newton-Gleichung $H_+ y = s$, so erhält man

$$\frac{y^T s}{\|y\|_2^2} = \frac{y^T H_+ y}{\|y\|_2^2} = \frac{\alpha^2 \lambda_{\min}(H_+) \|x_{\min}\|_2^2 + \beta^2 \lambda_{\max}(H_+) \|x_{\max}\|_2^2}{\alpha^2 \|x_{\min}\|_2^2 + \beta^2 \|x_{\max}\|_2^2},$$

woraus (*) sofort folgt.

Der Trick im zweiten Schritt des Beweises besteht darin, die Spur und die Determinante von H_+ auf zweierlei Weise auszurechnen. Nämlich einmal als Summe bzw. Produkt aller Eigenwerte von H_+ , zum anderen sozusagen direkt. Daher ist

$$\lambda_{\min}(H_+) + \lambda_{\max}(H_+) + (n-2) \frac{y^T s}{\|y\|_2^2} = \text{Spur}(H_+) = n \frac{y^T s}{\|y\|_2^2} + 2 \left(\frac{\|s\|_2^2}{y^T s} - \frac{y^T s}{\|y\|_2^2} \right)$$

und

$$\lambda_{\min}(H_+) \lambda_{\max}(H_+) \left(\frac{y^T s}{\|y\|_2^2} \right)^{n-2} = \det(H_+) = \frac{\|s\|_2^2}{y^T s} \left(\frac{y^T s}{\|y\|_2^2} \right)^{n-1}.$$

Hier haben wir bei der letzten Gleichung Satz 3.3 benutzt, indem wir beachteten, daß $H_+ = B_+^{-1}$, wobei B_+ die aus $B := (\|y\|_2^2 / y^T s) I$ resultierende BFGS-Update-Matrix ist. Ohne die Eigenwerte $\lambda_{\min}(H_+)$ und $\lambda_{\max}(H_+)$ explizit ausgerechnet zu haben, wissen wir jetzt, daß

$$\lambda_{\min}(H_+) + \lambda_{\max}(H_+) = 2 \frac{\|s\|_2^2}{y^T s}, \quad \lambda_{\min}(H_+) \lambda_{\max}(H_+) = \frac{\|s\|_2^2}{\|y\|_2^2}.$$

Nun wenden wir Aufgabe 8 aus 7.2 an, die leicht mit Hilfe der Ungleichung von Kantorowitsch (siehe Aufgabe 7 in 7.2) gelöst werden kann. Hiernach ist

$$\frac{x^T H_+ x}{\|x\|_2 \|H_+ x\|_2} \geq \frac{2 \sqrt{\lambda_{\min}(H_+) \lambda_{\max}(H_+)}}{\lambda_{\min}(H_+) + \lambda_{\max}(H_+)} = \frac{y^T s}{\|y\|_2 \|s\|_2}$$

für alle $x \in \mathbb{R}^n \setminus \{0\}$. Nimmt man hin, daß ausnahmsweise eine nicht bewiesene Aussage (die Ungleichung von Kantorowitsch) benutzt wurde, so ist das Lemma bewiesen. \square

Satz 4.6 Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (V) (a)–(c) aus 7.2.1 seien erfüllt. Man betrachte das in Satz 4.4 angegebene Shanno-Verfahren mit der einzigen Modifikation, daß in jedem Iterationsschritt t_k eine Powell-Schrittweite ist, so daß insbesondere $g_{k+1}^T p_k \geq \beta g_k^T p_k$ mit einem vorgegebenem $\beta \in (0, 1)$. Dann gilt: Das Shanno-Verfahren ist ein durchführbares Abstiegsverfahren, bricht es also nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\} \subset L_0$ und eine zugehörige Folge von Abstiegsrichtungen. Ist ferner $\lim_{k \rightarrow \infty} \|x_{k+1} - x_k\|_2 = 0$, so ist $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$, so daß wenigstens ein Häufungspunkt der Folge $\{x_k\}$ stationäre Lösung von (P) ist.

Beweis: Wir benutzen die Bezeichnungen aus Satz 4.4 und dessen Beweis. Zunächst beachten wir: Ist $g_k^T p_k < 0$, also p_k eine Abstiegsrichtung in x_k , und $t_k > 0$ eine zugehörige Powell-Schrittweite, so ist

$$y_k^T s_k = t_k y_k^T p_k = t_k (g_{k+1}^T p_k - g_k^T p_k) \geq -t_k (1 - \beta) g_k^T p_k > 0.$$

Damit ist

$$H_{k+1} := \frac{y_k^T s_k}{\|y_k\|_2^2} I + 2 \frac{s_k s_k^T}{y_k^T s_k} - \frac{s_k y_k^T + y_k s_k^T}{\|y_k\|_2^2}$$

symmetrisch und positiv definit und folglich $p_{k+1} := -H_{k+1} g_{k+1}$ eine Abstiegsrichtung in x_{k+1} , falls $g_{k+1} \neq 0$. Damit ist der erste Teil des Satzes bewiesen.

Nun nehmen wir an, das Shanno-Verfahren breche nicht vorzeitig ab und es sei $\lim_{k \rightarrow \infty} \|s_k\|_2 = 0$. Im Widerspruch zur Behauptung sei $\liminf_{k \rightarrow \infty} \|g_k\|_2 > 0$. Dann existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für $k = 0, 1, \dots$. Wegen Satz 2.3 gibt es eine Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{g_k^T p_k}{\|p_k\|_2} \right)^2 = \theta \|g_k\|_2^2 \delta_k, \quad \delta_k := \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2,$$

für $k = 0, 1, \dots$. Damit wird $f(x_k) - f(x_{k+1}) \geq \theta\epsilon^2\delta_k$ für $k = 0, 1, \dots$, Aufsummieren liefert $\sum_{k=0}^{\infty} \delta_k < \infty$. Andererseits erhält man mit Hilfe von Lemma 4.5, daß

$$\delta_{k+1} = \left(\frac{g_{k+1}^T p_{k+1}}{\|g_{k+1}\|_2 \|p_{k+1}\|_2} \right)^2 = \left(\frac{g_{k+1}^T H_{k+1} g_{k+1}}{\|g_{k+1}\|_2 \|H_{k+1} g_{k+1}\|_2} \right)^2 \geq \left(\frac{y_k^T s_k}{\|y_k\|_2 \|s_k\|_2} \right)^2.$$

Wegen $s_k = t_k p_k$ sowie $y_k^T s_k \geq -t_k(1-\beta)g_k^T p_k$ und $\|y_k\|_2 \leq \gamma \|s_k\|_2$ ergibt sich

$$\delta_{k+1} \geq (1-\beta)^2 \left(\frac{g_k^T p_k}{\|y_k\|_2 \|p_k\|_2} \right)^2 \geq \left(\frac{\|g_k\|_2 (1-\beta)}{\gamma \|s_k\|_2} \right)^2 \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2.$$

Damit ist

$$\delta_{k+1} \geq \delta_k \left(\frac{\epsilon(1-\beta)}{\gamma \|s_k\|_2} \right)^2, \quad k = 0, 1, \dots$$

Wegen $\lim_{k \rightarrow \infty} \|s_k\|_2 = 0$ ist $\|s_k\|_2 \leq \epsilon(1-\beta)/\gamma$ für alle hinreichend großen k . Für diese k ist $\delta_{k+1} \geq \delta_k$. Also ist $\{\delta_k\}$ keine Nullfolge und damit $\sum_{k=0}^{\infty} \delta_k = \infty$. Der gewünschte Widerspruch ist da, der Satz ist bewiesen. \square

Bemerkungen: Störend an den Voraussetzungen in Satz 4.6 ist natürlich die a priori nicht überprüfbare Bedingung $\lim_{k \rightarrow \infty} \|x_{k+1} - x_k\|_2 = 0$. Andererseits zeigt Aufgabe 4, daß auch für das Polak-Ribière-Verfahren für den Beweis einer entsprechenden Aussage dieselbe Bedingung gestellt wird (siehe auch M. J. D. POWELL (1977, Theorem 1)).

Von D. F. SHANNO (1978a, 1978b) wird auch noch auf eine restart-Version des obigen Verfahrens eingegangen. Hierzu werden Konvergenzaussagen analog denen in den Sätzen 4.4 und 4.6 bewiesen. \square

Aufgaben

- Gegeben sei die quadratische Zielfunktion $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n \times n}$. Seien $p_0, \dots, p_{n-1} \in \mathbb{R}^n$ konjugiert bezüglich Q . Man betrachte das folgende Verfahren zur Minimierung von $f(x)$ auf dem \mathbb{R}^n :

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := c + Qx_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP. Andernfalls berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T Q p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := c + t_k Q p_k.$$

Durch vollständige Induktion nach k zeige man: Sind $g_0, \dots, g_k \neq 0$, ist das Verfahren im k -ten Schritt also noch nicht abgebrochen, so ist x_{k+1} die Lösung der Aufgabe

$$(P_k) \quad \text{Minimiere } f(x), \quad x \in x_0 + \text{span}\{p_0, \dots, p_k\}.$$

Wegen $x_0 + \text{span}\{p_0, \dots, p_{n-1}\} = \mathbb{R}^n$ bricht das Verfahren also nach $m \leq n$ Schritten ab.

Hinweis: Nach Konstruktion ist klar, daß $x_{k+1} \in x_0 + \text{span}\{p_0, \dots, p_k\}$. Man zeige, daß $g_{k+1}^T p_i = 0$ für $0 \leq i \leq k$ und überlege sich, daß dies gleichwertig mit der Behauptung ist.

2. Gegeben sei die quadratische Zielfunktion $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n \times n}$. Zur Lösung der zugehörigen unrestringierten Optimierungsaufgabe betrachte man die folgende Modifikation des Verfahrens der konjugierten Gradienten, welche sich von diesem dadurch unterscheidet, daß nicht notwendig am Anfang $p_0 := -g_0$ gewählt wird.

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := c + Qx_0$. O. B. d. A. sei $g_0 \neq 0$. Wähle $p_0 \in \mathbb{R}^n$ mit $g_0^T p_0 < 0$.
- Für $k = 0, 1, \dots$:

Berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T Q p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := c + t_k Q p_k.$$

Falls $g_{k+1} = 0$, dann: $m := k + 1$, STOP. Andernfalls berechne

$$p_{k+1} := \begin{cases} -g_1 - \frac{g_1^T(g_1 - g_0)}{g_0^T p_0} p_0 & \text{für } k = 0, \\ -g_{k+1} + \frac{g_{k+1}^T g_0}{g_0^T p_0} p_0 + \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} p_k & \text{für } k = 1, 2, \dots \end{cases}$$

Durch vollständige Induktion nach k zeige man: Ist $k \geq 1$ und sind $g_1, \dots, g_k \neq 0$, so gilt:

- (a) Es ist $p_i^T g_k = 0$ für $0 \leq i < k$.
- (b) Es ist $g_k^T p_k = -\|g_k\|_2^2$.
- (c) Es ist $g_i^T g_k = 0$ für $1 \leq i < k$.
- (d) Es ist $p_i^T Q p_k = 0$ für $0 \leq i < k$, d. h. die Richtungen p_0, \dots, p_k sind Q -konjugiert.

Insbesondere bricht das Verfahren nach $m \leq n$ Schritten ab.

Hinweis: Siehe E. M. L. BEALE (1972). Von M. J. D. POWELL (1977) wird die Idee aufgegriffen, konjugierte Richtungen zu erzeugen, ohne von $p_0 := -g_0$ auszugehen. Hiermit werden von Powell Verfahren der konjugierten Gradienten mit restart entwickelt.

3. Man entwickle ein Verfahren der konjugierten Gradienten zur Lösung des linearen Ausgleichsproblems

$$(LA) \quad \text{Minimiere } \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = n$ und $b \in \mathbb{R}^m$. Um den Fall einer dünn besetzten Matrix A effektiv behandeln zu können, vermeide man es, am Anfang die (i. allg. voll besetzte) Matrix $Q = A^T A$ zu bilden.

Hinweis: Das lineare Ausgleichsproblem (LA) ist äquivalent zur unrestringierten Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := -(A^T b)^T x + \frac{1}{2} x^T A^T A x.$$

Wegen $\text{Rang}(A) = n$ ist $A^T A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Eine „naive“ Anwendung des Verfahrens der konjugierten Gradienten würde das folgende Verfahren ergeben:

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := A^T(Ax_0 - b)$ und setze $p_0 := -g_0$.

- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP. Andernfalls berechne

$$t_k := \frac{\|g_k\|_2^2}{\|Ap_k\|_2^2}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := g_k + t_k A^T Ap_k.$$

Berechne

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \quad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Mit $r_k := Ax_k - b$ ist $g_k = A^T r_k$. Setzt man noch $q_k := Ap_k$, so erhält man den folgenden Algorithmus (siehe auch J. STOER, R. BULIRSCH (1990, S. 299)):

- Wähle $x_0 \in \mathbb{R}^n$, berechne $r_0 := Ax_0 - b$, $g_0 := A^T r_0$ und setze $p_0 := -g_0$.

- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: $m := k$, STOP. Andernfalls berechne $q_k := Ap_k$ und

$$t_k := \frac{\|g_k\|_2^2}{\|q_k\|_2^2}, \quad x_{k+1} := x_k + t_k p_k, \quad r_{k+1} := r_k + t_k q_k.$$

Berechne

$$g_{k+1} := A^T r_{k+1}, \quad \beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \quad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

4. Das Verfahren der konjugierten Gradienten von Polak-Ribi  re unterscheidet sich von dem Fletcher-Reeves-Verfahren nur darin, da   $\beta_k := g_{k+1}^T(g_{k+1} - g_k)/\|g_k\|_2^2$ (statt $\beta_k := \|g_{k+1}\|_2^2/\|g_k\|_2^2$) gesetzt wird. Man betrachte das dann definierte Polak-Ribi  re-Verfahren mit exakter Schrittweitenstrategie zur L  sung von

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n.$$

Man zeige:

- Sind die Voraussetzungen (K) (a)–(c) erf  llt, so liefert das Verfahren, wenn es nicht vorzeitig mit der L  sung x^* von (P) abbricht, eine Folge $\{x_k\}$, die gegen x^* konvergiert.
- Die Voraussetzungen (V) (a)–(c) seien erf  llt, das Verfahren breche nicht vorzeitig mit einer station  ren L  sung von (P) ab. F  r die durch das Verfahren erzeugte Folge $\{x_k\}$ gelte $\lim_{k \rightarrow \infty} \|x_{k+1} - x_k\|_2 = 0$. Dann ist $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$, so da   die Folge $\{x_k\}$ wenigstens eine station  re L  sung von (P) als H  ufungspunkt besitzt.

Hinweis: Man setze

$$\delta_k := \left(\frac{g_k^T p_k}{\|g_k\|_2 \|p_k\|_2} \right)^2 = \frac{\|g_k\|_2^2}{\|p_k\|_2^2}.$$

F  r den ersten Teil der Aufgabe zeige man die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \geq \delta$, $k = 0, 1, \dots$, und wende Satz 2.7 an. Dies kann folgenderma  en geschehen:

Zunächst ist wegen der Cauchy-Schwarzschen Ungleichung sowie der Lipschitzstetigkeit des Gradienten

$$|\beta_k| \leq t_k \gamma \frac{\|g_{k+1}\|_2 \|p_k\|_2}{\|g_k\|_2^2}.$$

Nun schätzt man die exakte Schrittweite t_k mit Hilfe der vorausgesetzten gleichmäßigen Konvexität von f nach oben ab:

$$0 = g_{k+1}^T p_k = g_k^T p_k + (g_{k+1} - g_k)^T p_k \geq g_k^T p_k + ct_k \|p_k\|_2^2.$$

Der Reihe nach erhält man

$$t_k \leq -\frac{g_k^T p_k}{c \|p_k\|_2^2} = \frac{\|g_k\|_2^2}{c \|p_k\|_2^2}, \quad |\beta_k| \leq \frac{\gamma \|g_{k+1}\|_2}{c \|p_k\|_2}, \quad \|p_{k+1}\|_2 \leq \left(1 + \frac{\gamma}{c}\right) \|g_{k+1}\|_2.$$

Hieraus folgt die Behauptung (a). Für den zweiten Teil der Aufgabe mache man einen Widerspruchsbeweis. Wäre $\liminf_{k \rightarrow \infty} \|g_k\|_2 > 0$, so existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k . Mit Hilfe von Satz 2.2 folgt $\sum_{k=0}^{\infty} \delta_k < \infty$. Andererseits erhält man mit $s_k := x_{k+1} - x_k$ die Abschätzung

$$\|p_{k+1}\|_2^2 = \|g_{k+1}\|_2^2 + \beta_k^2 \|p_k\|_2^2 \leq \left(1 + \frac{\gamma^2 \|s_k\|_2^2 \|p_k\|_2^2}{\|g_k\|_2^4}\right) \|g_{k+1}\|_2^2$$

und hieraus (wegen $\|g_k\|_2 \geq \epsilon$ und $\lim_{k \rightarrow \infty} \|s_k\|_2 = 0$)

$$\frac{1}{\delta_{k+1}} = \frac{\|p_{k+1}\|_2^2}{\|g_{k+1}\|_2^2} \leq 1 + \frac{\gamma^2}{\epsilon^2} \|s_k\|_2^2 \frac{\|p_k\|_2^2}{\|g_k\|_2^2} \leq 1 + \frac{\|p_k\|_2^2}{\|g_k\|_2^2} = 1 + \frac{1}{\delta_k}$$

für alle hinreichend großen k . Daher ist $\sum_{k=0}^{\infty} \delta_k = \infty$, insgesamt hat man einen Widerspruch erhalten. (Mit einem etwas anderen Beweis findet man den zweiten Teil der Aufgabe auch bei M. J. D. POWELL (1977, Theorem 1).)

5. Unter den Voraussetzungen (V) (a)–(c) aus 7.2.1 betrachte man zur Lösung der unrestringierten Optimierungsaufgabe (P) das Fletcher-Reeves-Verfahren mit einer inexakten Schrittweitenstrategie (siehe M. AL-BAALI (1985)):

- Für die Schrittweitenstrategie seien α und β mit $0 < \alpha < \beta < \frac{1}{2}$ vorgegeben.
- Setze $g_0 := \nabla f(x_0)$ und $p_0 := -g_0$.
- Für $k = 0, 1, \dots$:

Falls $g_k = 0$, dann: STOP, x_k ist stationäre Lösung von (P).
Bestimme $t_k > 0$ mit

$$f(x_k + t_k p_k) \leq f(x_k) + \alpha t_k g_k^T p_k, \quad |\nabla f(x_k + t_k p_k)^T p_k| \leq -\beta g_k^T p_k.$$

Setze bzw. berechne

$$x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := \nabla f(x_{k+1})$$

sowie

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \quad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Man zeige:

(a) Ist im k -ten Schritt noch kein Abbruch erfolgt, ist also $g_0, \dots, g_k \neq 0$, so ist

$$-\frac{1}{1-\beta} < -\sum_{j=0}^k \beta^j \leq \frac{g_k^T p_k}{\|g_k\|_2^2} \leq -2 + \sum_{j=0}^k \beta^j < -\frac{1-2\beta}{1-\beta}.$$

Wegen $\beta \in (0, \frac{1}{2})$ ist daher p_k eine Abstiegsrichtung in x_k . Mit Hilfe von Aufgabe 1 aus 7.2 folgt die Existenz einer Schrittweite $t_k > 0$ mit den geforderten Eigenschaften.

(b) Bricht das Verfahren nicht vorzeitig mit einer stationären Lösung ab, so erzeugt es eine Folge $\{x_k\}$ mit $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$. Sind sogar die Voraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

Hinweis: Den ersten Teil der Aufgabe zeige man durch vollständige Induktion nach k . Im zweiten Teil argumentiere man wie beim Beweis von Satz 4.3 und nehme im Widerspruch zur Behauptung an, es gäbe ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k . Eine Anwendung von Aufgabe 1 in 7.2 liefert die Existenz einer von k unabhängigen Konstanten $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{g_k^T p_k}{\|p_k\|_2} \right)^2, \quad k = 0, 1, \dots$$

Wegen der rechten Ungleichung in (a) ist

$$(*) \quad f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{1-2\beta}{1-\beta} \right)^2 \frac{1}{\alpha_k} \quad \text{mit } \alpha_k := \frac{\|p_k\|_2^2}{\|g_k\|_2^2}.$$

Für $k = 1, 2, \dots$ ist dann mit der linken Ungleichung in (a)

$$\alpha_k \leq \left(\frac{1+\beta}{1-\beta} \right) \frac{1}{\|g_k\|_2^2} + \alpha_{k-1} \leq \left(\frac{1+\beta}{1-\beta} \right) \sum_{j=0}^k \frac{1}{\|g_j\|_2^2} \leq \frac{1}{\epsilon^2} \left(\frac{1+\beta}{1-\beta} \right) (k+1)$$

und daher wegen (*)

$$f(x_k) - f(x_{k+1}) \geq \theta \epsilon^2 \left(\frac{1-2\beta}{1-\beta} \right)^2 \left(\frac{1-\beta}{1+\beta} \right) \frac{1}{k+1}, \quad k = 0, 1, \dots$$

Hieraus erhält man einen Widerspruch zur Annahme.

Sind sogar die Voraussetzungen (K) (a)–(c) erfüllt, so folgt genau wie im Beweis zu Satz 4.3 die Konvergenz der Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* .

6. Man programmiere das Verfahren der konjugierten Gradienten zur Lösung der unrestriktiven Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := c^T x + \frac{1}{2} x^T Q x, \quad x \in \mathbb{R}^n,$$

wobei $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist. Anschließend teste man das Programm für $m = 5, 10, 15, 20$ und $n := m^2$ an der Matrix

$$Q := \begin{pmatrix} Q_m & -I_m & \cdots & 0 \\ -I_m & Q_m & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_m \\ 0 & \cdots & -I_m & Q_m \end{pmatrix} \in \mathbb{R}^{n \times n}$$

mit der $m \times m$ -Einheitsmatrix I_m und der Tridiagonalmatrix

$$Q_m := \begin{pmatrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 4 \end{pmatrix} \in \mathbb{R}^{m \times m}.$$

Der Vektor c sei durch $c := -h^2 e$ mit $h := 1/(m+1)$ und $e := (1, \dots, 1)^T \in \mathbb{R}^n$ gegeben.

Hinweis: Natürlich sollte das Programm so geschrieben werden, daß nicht etwa die Matrix Q als gespeichert angenommen wird. Vielmehr ist bei einem vorgegebenem Vektor $p \in \mathbb{R}^n$ nur der Vektor $q := Qp$ als bekannt anzusehen.

7.5 Trust-Region-Verfahren

7.5.1 Einführung

Gegeben sei wieder die unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x), \quad x \in \mathbb{R}^n.$$

Durch den Modellalgorithmus in 7.2 wurde deutlich, daß die bisher behandelten Verfahren zur näherungsweisen Lösung von (P) sich aus einer Richtungs- und einer Schrittweitenstrategie zusammensetzen. Von einer aktuellen Näherung ausgehend wurde in eine Abstiegsrichtung gegangen, in diese wurde ein Schritt gemacht, dessen Länge durch die gewählte Schrittweitenstrategie festgelegt ist.

Die Idee bei den Trust-Region-Verfahren ist etwas anders. Sie besteht im wesentlichen darin, die Zielfunktion lokal auf einer Δ -Kugel (bezüglich einer geeigneten Norm) um eine aktuelle Näherung durch ein einfacheres „Modell“ zu ersetzen, etwa einer linearen oder quadratischen Approximation der Zielfunktion. Dann bestimmt man ein Minimum des Modells bzw. der vereinfachten Zielfunktion auf dieser Δ -Kugel. Gleichzeitig wird hierbei geprüft, ob eine notwendige Optimalitätsbedingung erster oder sogar zweiter Ordnung erfüllt ist. Wird eine Verminderung des Zielfunktionswertes entweder nicht erreicht, oder ist diese eher enttäuschend gering, so hat man dem Modell auf einer zu großen Kugel um die aktuelle Näherung „vertraut“, diese wird daher verkleinert und auf dieser verkleinerten Kugel erneut ein Minimum der Modell-Funktion bestimmt. Andernfalls wird dieses Minimum als neue aktuelle Näherung akzeptiert und der Radius Δ wird vergrößert, wenn ein verschärfter Test auf hinreichende Verminderung des Zielfunktionswertes erfolgreich bestanden wird. Das ist, sehr lax ausgedrückt, die Idee der Trust-Region-Verfahren.

Nun soll diese Idee genauer gefaßt werden. Bei gegebenem $x \in \mathbb{R}^n$ sei ein „einfaches Modell“ $f_x: \mathbb{R}^n \rightarrow \mathbb{R}$ für die (i. allg. komplizierte) Funktion $p \mapsto f(x + p)$ vorgegeben. Eine Minimalforderung an die Modell-Funktion f_x wird $f_x(0) = f(x)$ sein. Ist z. B. $f \in C^2(\mathbb{R}^n)$, so liegt es nahe,

$$f_x(p) := f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p$$

zu setzen, wobei eventuell noch $\nabla^2 f(x)$ durch eine symmetrische (nicht notwendig positiv definite) Matrix $B \in \mathbb{R}^{n \times n}$ ersetzt sein kann. Ist dagegen $f(x) := \|F(x)\|$ mit einer stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ und einer Norm $\|\cdot\|$ auf dem \mathbb{R}^m , so könnte man die Modell-Funktion f_x durch

$$f_x(p) := \|F(x) + F'(x)p\|$$

definieren. Entsprechendes gilt für andere „halbglatte“ Zielfunktionen $f = g \circ F$ mit einer stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ und einer konvexen Funktion $g: \mathbb{R}^m \rightarrow \mathbb{R}$.

Ist eine solche Modell-Funktion f_x für jedes $x \in \mathbb{R}^n$ gegeben, so könnte ein Schritt einer einfachen Version des Trust-Region-Verfahrens zur Lösung von (P) folgendermaßen aussehen, wobei $\|\cdot\|$ eine feste Vektornorm auf dem \mathbb{R}^n bedeute:

- (0) Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0, 1)$ und $\sigma_2 > 1$.
- (1) Gegeben seien eine aktuelle Näherung x für eine (stationäre, lokale, globale) Lösung von (P) und ein $\Delta > 0$.
- (2) Berechne eine Lösung $p^* \in \mathbb{R}^n$ der Aufgabe

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p), \quad \|p\| \leq \Delta.$$

(Wir werden später zu zeigen haben, daß dieses Hilfsproblem unter geeigneten Voraussetzungen „einfach“ zu lösen ist.)

- (3) Falls $f(x) = f_x(p^*)$, dann: STOP.
(Ist die Modell-Funktion f_x richtig gewählt, so wird x in diesem Falle wenigstens eine stationäre Lösung von (P) sein.)
- (4) Berechne

$$r_x := \frac{f(x) - f(x + p^*)}{f(x) - f_x(p^*)}.$$
 Falls $r_x \geq \rho_1$, dann setze $x_+ := x + p^*$. Andernfalls setze $x_+ := x$.
 - (a) Falls $r_x < \rho_1$, dann wähle $\Delta_+ \in (0, \sigma_1 \Delta]$.
 - (b) Falls $r_x \in [\rho_1, \rho_2)$, dann wähle $\Delta_+ \in [\sigma_1 \Delta, \Delta]$.
 - (c) Falls $r_x \geq \rho_2$, dann wähle $\Delta_+ \in [\Delta, \sigma_2 \Delta]$.
- (5) Setze $(x, \Delta) := (x_+, \Delta_+)$ und gehe nach (2).

Einige Bemerkungen zu den Tests in (4) sind angebracht. Da der Schritt (3) jeweils passiert wurde, also $f(x) \neq f_x(p^*)$ gilt, und von der Modell-Funktion $f_x(0) = f(x)$ angenommen wurde, ist notwendig $f(x) > f_x(p^*)$. Entscheidend für den Test ist die Größe r_x , der Quotient aus der tatsächlichen und der durch das Modell vorhergesagten Verminderung des Zielfunktionswertes. Je näher r_x bei 1 liegt, desto besser „stimmt“ das Modell.

Ist $r_x < \rho_1$, so hat sich keine Verminderung eingestellt oder diese ist, verglichen mit der vorhergesagten, zu gering. Das wird darauf zurückgeführt, daß dem Modell auf einer zu großen Kugel vertraut wurde, diese wird daher entsprechend verkleinert und ein erneuter Versuch unternommen.

Ist sogar der verschärzte Test $r_x \geq \rho_2$ erfolgreich, so stimmen die tatsächliche und die vorhergesagte Verminderung hinreichend gut überein, so daß im nächsten Schritt dem Modell auf einer i. allg. größeren Kugel vertraut wird. Für $r_x \in [\rho_1, \rho_2]$ ist man mit der neuen Näherung $x_+ = x + p^*$ zufrieden, vergrößert den Bereich aber nicht.

Beispiel: Unter der Voraussetzung $f \in C^1(\mathbb{R}^n)$ wollen wir einmal das denkbar einfachste Trust-Region-Verfahren betrachten. Hier sei die Modell-Funktion f_x durch $f_x(p) := f(x) + \nabla f(x)^T p$ gegeben, ferner sei $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm. Das Hilfsproblem $(P_{x,\Delta})$ lautet in diesem Falle also:

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := f(x) + \nabla f(x)^T p, \quad \|p\|_2 \leq \Delta.$$

Sei p^* eine Lösung von $(P_{x,\Delta})$. Ist $f(x) = f_x(p^*)$, so bedeutet dies, daß auch der Nullvektor eine Lösung des Hilfsproblems ist. In diesem Falle ist daher $\nabla f(x)^T p \geq 0$ für alle p mit $\|p\|_2 \leq \Delta$. Dies wiederum impliziert $\nabla f(x) = 0$, daß x also eine stationäre Lösung von (P) ist. Für $\nabla f(x) \neq 0$ ist dagegen die Lösung von $(P_{x,\Delta})$ offenbar durch

$$p^* := -\Delta \frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

gegeben. Also ist p^* bis auf den positiven Faktor Δ die negative, normierte Gradientenrichtung. \square

7.5.2 Glatte, unrestringierte Optimierungsaufgaben

In diesem Unterabschnitt wird die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) als zweimal stetig differenzierbar vorausgesetzt. Ziel ist es, eine Trust-Region-Modifikation des Newton-Verfahrens anzugeben und zu untersuchen.

Ein Nachteil des in 7.3.1 behandelten Newton-Verfahrens mit Schrittweitenstrategie besteht darin, daß die Newton-Richtung $p := -\nabla^2 f(x)^{-1} \nabla f(x)$ keine Abstiegsrichtung zu sein braucht, wenn die symmetrische Matrix $\nabla^2 f(x)$, die Hessesche der Zielfunktion in der aktuellen Näherung x , nicht positiv definit ist. Hier können Trust-Region-Verfahren helfen, und in der Tat ist das, neben der Entwicklung von Verfahren für nichtlineare Ausgleichsprobleme, einer der Ausgangspunkte ihrer Entwicklung (siehe z. B. D. C. SØRENSEN (1982)). Wie in der Einführung schon erwähnt wurde, liegt es hier nahe, als Modell $f_x(p)$ eine quadratische Approximation an $f(x + p)$ zu wählen, also

$$f_x(p) := f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p$$

zu setzen. Ferner wird als Norm $\|\cdot\|$ stets die euklidische Norm $\|\cdot\|_2$ gewählt, so daß das im Trust-Region-Verfahren auftretende Hilfsproblem mit einem gegebenem $\Delta > 0$ die Form

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := f(x) + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta$$

hat, wobei zur Abkürzung $g := \nabla f(x)$ und $B := \nabla^2 f(x)$ gesetzt wurde. Man beachte, daß B zwar eine symmetrische, aber nicht notwendig positiv definite Matrix ist. Klar ist, daß dieses Hilfsproblem eine globale Lösung p^* besitzt, da eine stetige Funktion auf einer nichtleeren, kompakten Menge ihr Minimum annimmt. Da B aber nicht als positiv definit vorausgesetzt wurde, so daß die Zielfunktion im Hilfsproblem nicht notwendig gleichmäßig konvex ist, kann $(P_{x,\Delta})$ auch von einer globalen Lösung verschiedene lokale Lösungen besitzen.

In dem folgenden Lemma (siehe D. C. SORENSEN (1982, Lemma 2.4 und Lemma 2.8) oder auch R. FLETCHER (1987, S. 101)) werden notwendige und hinreichende Bedingungen dafür angegeben, daß ein $p^* \in \mathbb{R}^n$ eine globale Lösung von $(P_{x,\Delta})$ ist. Ferner wird die durch das Modell vorhergesagte Verminderung nach unten abgeschätzt.

Lemma 5.1 *Man betrachte die Aufgabe*

$$(P_\Delta) \quad \text{Minimiere } \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

wobei $\Delta > 0$, $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ gegeben sind. Dann ist ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ genau dann eine globale Lösung von (P_Δ) , wenn ein $\lambda^* \geq 0$ mit

- (a) $(B + \lambda^* I)p^* = -g$,
- (b) $\lambda^*(\|p^*\|_2 - \Delta) = 0$,
- (c) $B + \lambda^* I$ ist positiv semidefinit

existiert. Darüberhinaus ist p^* eindeutige globale Lösung von (P_Δ) , wenn $B + \lambda^* I$ sogar positiv definit ist. Ferner ist $\phi(p^*) = f$ genau dann, wenn $g = 0$ und B positiv semidefinit ist. Schließlich ist

$$f - \phi(p^*) \geq \frac{1}{2} \|g\|_2 \min \left(\Delta, \frac{\|g\|_2}{\|B\|_2} \right).$$

Beweis: Zunächst nehmen wir an, p^* sei eine globale Lösung von (P_Δ) und zeigen die Existenz eines λ^* mit (a)–(c).

Die Existenz eines $\lambda^* \geq 0$ derart, daß (a) und (b) gelten, folgt durch eine Anwendung der Lagrangeschen Multiplikatorenregel bzw. der notwendigen Optimalitätsbedingungen erster Ordnung für restringierte Optimierungsaufgaben. Insbesondere ist $\lambda^* = 0$, wenn $\|p^*\|_2 < \Delta$. Die notwendigen Optimalitätsbedingungen zweiter Ordnung (wiederum für restringierte Optimierungsaufgaben) sichern darüberhinaus, daß für $\|p^*\|_2 < \Delta$ die Matrix $B = B + \lambda^* I$ positiv semidefinit ist und für $\|p^*\|_2 = \Delta$ immer noch $h^T(B + \lambda^* I)h \geq 0$ für alle $h \in \mathbb{R}^n$ mit $(p^*)^T h = 0$. Dieser Teil des Beweises wird als Aufgabe gestellt (siehe Aufgabe 1, dort sind ausführliche Hinweise zur Lösung angegeben). Bis hierhin ging lediglich ein, daß p^* insbesondere auch eine lokale Lösung von (P_Δ) ist.

Nun zeigen wir, daß $B + \lambda^* I$ positiv semidefinit ist. Wegen des ersten Teiles des Beweises genügt es hierzu, den Fall $\|p^*\|_2 = \Delta$ zu betrachten und nachzuweisen, daß $h^T(B + \lambda^* I)h \geq 0$ für alle $h \in \mathbb{R}^n$ mit $(p^*)^T h \neq 0$. Ein solches h sei gegeben. Mit

$$t := -\frac{2(p^*)^T h}{\|h\|_2^2}$$

ist dann $t \neq 0$ und nach Konstruktion $\|p^* + th\|_2 = \|p^*\|_2$. Zur Abkürzung setze man $p := p^* + th$. Da p^* eine globale Lösung von (P_Δ) ist, erhalten wir unter Benutzung von (a) sowie der Definition von p , daß

$$0 \leq \phi(p) - \phi(p^*) = \underbrace{(g + Bp^*)^T}_{=-\lambda^* p^*} \underbrace{(p - p^*)}_{=th} + \frac{1}{2} t^2 h^T B h = \frac{1}{2} t^2 h^T (B + \lambda^* I) h,$$

womit auch (c) nachgewiesen ist.

Seien nun $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ und $\lambda^* \geq 0$ mit (a)–(c) gegeben. Für ein beliebiges $p \in \mathbb{R}^n$ mit $\|p\|_2 \leq \Delta$ ist

$$\begin{aligned} \phi(p) - \phi(p^*) &= \underbrace{(g + Bp^*)^T}_{=-\lambda^* p^*} (p - p^*) + \frac{1}{2} (p - p^*)^T B (p - p^*) \\ &= -\lambda^* (p^*)^T (p - p^*) + \frac{1}{2} \underbrace{(p - p^*)^T (B + \lambda^* I) (p - p^*)}_{\geq 0} - \frac{\lambda^*}{2} \|p - p^*\|_2^2 \\ &\geq \frac{\lambda^*}{2} (\|p^*\|_2^2 - \|p\|_2^2) \quad (\text{wegen (c)}) \\ &= \frac{\lambda^*}{2} (\Delta^2 - \|p\|_2^2) \quad (\text{wegen (b)}) \\ &\geq 0, \end{aligned}$$

und daher p^* eine globale Lösung von (P_Δ) . Ist $B + \lambda^* I$ sogar positiv definit, so entnimmt man der obigen Gleichungs-Ungleichungskette, daß $p = p^*$ aus $\phi(p) = \phi(p^*)$ folgt, und das bedeutet die Eindeutigkeit der globalen Lösung.

Ist $\phi(p^*) = f$, so ist auch $p^{**} := 0$ eine Lösung von (P_Δ) . Aus (a)–(c) folgt sofort, daß notwendig $g = 0$ und B positiv semidefinit ist. Die Umkehrung ist trivial.

Für den Beweis der noch zu beweisenden Ungleichung können wir o. B. d. A. $g \neq 0$ annehmen. Für ein beliebiges p mit $\|p\|_2 \leq \Delta$ ist wegen der Optimalität von p^* offenbar

$$(*) \quad f - \phi(p^*) \geq f - \phi(p) = -g^T p - \frac{1}{2} p^T B p \geq -g^T p - \frac{1}{2} \|p\|_2^2 \|B\|_2.$$

Ist $\Delta \|B\|_2 \leq \|g\|_2$, so ist wegen (*) (setze $p := -(\Delta/\|g\|_2)g$)

$$f - \phi(p^*) \geq \Delta \|g\|_2 - \frac{1}{2} \Delta^2 \|B\|_2 \geq \frac{1}{2} \Delta \|g\|_2.$$

Ist dagegen $\Delta \|B\|_2 > \|g\|_2$, so setze man $p := -(1/\|B\|_2)g$ und erhalte aus (*)

$$f - \phi(p^*) \geq \frac{\|g\|_2^2}{\|B\|_2} - \frac{\|g\|_2^2}{2\|B\|_2} = \frac{\|g\|_2^2}{2\|B\|_2},$$

insgesamt ist die behauptete Abschätzung bewiesen. \square

Wenden wir das eben bewiesene Lemma 5.1 auf die uns hier interessierende Situation an, so erkennen wir, daß ein Abbruch in Schritt (3) des Trust-Region-Verfahrens nur dann erfolgt, wenn $\nabla f(x) = 0$ und $\nabla^2 f(x)$ positiv semidefinit ist, also die notwendigen Optimalitätsbedingungen zweiter Ordnung in x erfüllt sind.

Nun geben wir den globalen Konvergenzsatz für eine Trust-Region-Modifikation des Newton-Verfahrens an (siehe D. C. SORENSEN (1982) und J. J. MORE (1983, S. 274 ff.), an dessen Beweisgang wir uns halten werden).

Satz 5.2 Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Niveau-menge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ sei kompakt, die Zielfunktion f sei auf einer offenen Obermenge von L_0 zweimal stetig differenzierbar und der Gradient $\nabla f(\cdot)$ auf L_0 lipschitzstetig. Man betrachte das folgende Verfahren.

- Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0, 1)$ und $\sigma_2 > 1$.
- Sei $x_0 \in \mathbb{R}^n$ gegeben, berechne $g_0 := \nabla f(x_0)$, $B_0 := \nabla^2 f(x_0)$. Ferner sei ein $\Delta_0 > 0$ vorgegeben.
- Für $k = 0, 1, \dots$:

Berechne eine globale Lösung p_k der Aufgabe

$$(P_k) \quad \text{Minimiere } f_k(p) := f(x_k) + g_k^T p + \frac{1}{2} p^T B_k p, \quad \|p\|_2 \leq \Delta_k.$$

Falls $f(x_k) = f_k(p_k)$, dann: STOP. In x_k sind die notwendigen Optimalitätsbedingungen zweiter Ordnung erfüllt.

Berechne

$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$

Falls $r_k \geq \rho_1$, dann setze $x_{k+1} := x_k + p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$, $B_{k+1} := \nabla^2 f(x_{k+1})$. In diesem Falle nennen wir den Iterationsschritt k erfolgreich.

Andernfalls setze $x_{k+1} := x_k$ sowie $g_{k+1} := g_k$, $B_{k+1} := B_k$.

- Falls $r_k < \rho_1$, dann wähle $\Delta_{k+1} \in (0, \sigma_1 \Delta_k]$.
- Falls $r_k \in [\rho_1, \rho_2)$, dann wähle $\Delta_{k+1} \in [\sigma_1 \Delta_k, \Delta_k]$.
- Falls $r_k \geq \rho_2$, dann wähle $\Delta_{k+1} \in [\Delta_k, \sigma_2 \Delta_k]$.

Das Verfahren breche nicht vorzeitig ab. Dann liefert es eine Folge $\{x_k\}$ mit:

- Es ist $\lim_{k \rightarrow \infty} \|g_k\|_2 = 0$, insbesondere ist jeder Häufungspunkt der Folge $\{x_k\}$ eine stationäre Lösung von (P).
- Die Folge $\{x_k\}$ besitzt mindestens einen Häufungspunkt x^* , in dem die notwendigen Optimalitätsbedingungen zweiter Ordnung erfüllt sind, für den also $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv semidefinit ist.

3. Ist x^* ein Häufungspunkt der Folge $\{x_k\}$ und ist $\nabla^2 f(x^*)$ positiv definit, so konvergiert die Folge $\{x_k\}$ gegen x^* , nach endlich vielen Schritten sind alle Iterationsschritte erfolgreich, es ist $\|p_k\|_2 < \Delta_k$ für alle hinreichend großen k und $\inf_{k=0,1,\dots} \Delta_k > 0$. Ist darüberhinaus $\nabla^2 f(\cdot)$ auf einer Kugel um x^* in x^* lipschitzstetig, so konvergiert die Folge $\{x_k\}$ von mindestens zweiter Ordnung gegen x^* .

Beweis: Als Vorbereitung zeigen wir zunächst, daß $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$. Angenommen, das sei nicht der Fall. Dann existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k . Wir zeigen, daß einerseits $\sum_{k=0}^{\infty} \Delta_k < \infty$, insbesondere die Folge $\{\Delta_k\}$ gegen Null konvergiert, und andererseits $\lim_{k \rightarrow \infty} r_k = 1$ ist. Da letzteres zeigt, daß $r_k \geq \rho_2$ und damit $\Delta_{k+1} \geq \Delta_k$ für alle hinreichend k gilt, wird man einen Widerspruch erreicht haben.

Gibt es nur endlich viele erfolgreiche Schritte, so ist $\Delta_{k+1} \leq \sigma_1 \Delta_k$ für alle hinreichend großen k , so daß $\sum_{k=0}^{\infty} \Delta_k < \infty$ wegen $\sigma_1 \in (0, 1)$ trivialerweise richtig ist. Ist der Iterationsschritt k erfolgreich, so ist wegen der Abschätzung in Lemma 5.1

$$f(x_k) - f(x_{k+1}) \geq \rho_1 [f(x_k) - f_k(p_k)] \geq \frac{\rho_1}{2} \|g_k\|_2 \min\left(\Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2}\right).$$

Wegen $\|g_k\|_2 \geq \epsilon$ und mit $\beta := \max_{x \in L_0} \|\nabla^2 f(x)\|_2$ ist daher für jeden erfolgreichen Iterationsschritt k :

$$(*) \quad f(x_k) - f(x_{k+1}) \geq \frac{\rho_1 \epsilon}{2} \min\left(\Delta_k, \frac{\epsilon}{\beta}\right).$$

Nun nehmen wir an, daß es unendlich viele erfolgreiche Iterationsschritte gibt, diese seien in der Indexmenge E zusammengefaßt. Da $\{f(x_k)\}$ eine monoton nicht wachsende, nach unten beschränkte Folge ist, erhält man aus $(*)$

$$\frac{\rho_1 \epsilon}{2} \sum_{k \in E} \min\left(\Delta_k, \frac{\epsilon}{\beta}\right) \leq \sum_{k \in E} [f(x_k) - f(x_{k+1})] \leq \sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})] < \infty$$

und hieraus $\sum_{k \in E} \Delta_k < \infty$, da E unendlich viele Elemente enthält. Nun betrachten wir die k , die zwischen zwei aufeinanderfolgenden erfolgreichen Iterationsschritten $i < j$ liegen. Dann ist $\Delta_{i+1} \leq \sigma_2 \Delta_i$ (da i erfolgreich) und $\Delta_{k+1} \leq \sigma_1 \Delta_k$ für $k = i+1, \dots, j-1$. Hieraus folgt durch Zurückspulen und Summieren

$$\sum_{k=i+1}^{j-1} \Delta_k \leq \frac{\sigma_2}{1 - \sigma_1} \Delta_i.$$

Insgesamt ist damit $\sum_{k=0}^{\infty} \Delta_k < \infty$ bewiesen. Aus

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|_2 \leq \sum_{k=0}^{\infty} \|p_k\|_2 \leq \sum_{k=0}^{\infty} \Delta_k < \infty$$

folgt, daß $\{x_k\}$ eine Cauchy-Folge ist und damit gegen ein x^* konvergiert. Nun wird $\lim_{k \rightarrow \infty} r_k = 1$ nachgewiesen. Für alle hinreichend großen k ist

$$|r_k - 1| = \left| \frac{f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k)}{f(x_k) - f_k(p_k)} \right|$$

$$\begin{aligned}
&\leq \frac{2}{\epsilon \Delta_k} \left| f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k) \right| \\
&\quad (\text{erneute Anwendung der Abschätzung in Lemma 5.1}) \\
&\leq \frac{1}{\epsilon \|p_k\|_2} |p_k^T [\nabla^2 f(x_k) - \nabla^2 f(x_k + \theta_k p_k)] p_k| \quad \text{mit } \theta_k \in (0, 1).
\end{aligned}$$

Wegen $p_k \rightarrow 0$ folgt $r_k \rightarrow 1$. Das aber ist, wie wir am Anfang schon festgestellt haben, ein Widerspruch zu $\Delta_k \rightarrow 0$. Damit ist $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$ bewiesen.

Die erste Aussage des Satzes, daß $\{\|g_k\|_2\}$ gegen Null konvergiert, wird durch Widerspruch bewiesen. Ist das nicht der Fall, so gibt es ein $\epsilon > 0$ und eine Teilfolge $\{x_{k(i)}\} \subset \{x_k\}$ mit $\|g_{k(i)}\|_2 \geq 2\epsilon$ für $i = 1, 2, \dots$. Im vorbereitenden Teil haben wir bewiesen, daß $\liminf_{k \rightarrow \infty} \|g_k\|_2 = 0$. Daher gibt es unendlich viele k mit $\|g_k\|_2 < \epsilon$. Für $i = 1, 2, \dots$ existiert daher ein $l(i) > k(i)$ mit

$$\|g_k\|_2 \geq \epsilon \quad (k(i) \leq k < l(i)), \quad \|g_{l(i)}\|_2 < \epsilon \quad (i = 1, 2, \dots).$$

Hieraus erkennen wir: Ist $k(i) \leq k < l(i)$ und der Iterationsschritt k erfolgreich, so ist wegen $\|x_{k+1} - x_k\|_2 \leq \Delta_k$ und der Abschätzung in Lemma 5.1

$$f(x_k) - f(x_{k+1}) \geq \rho_1 [f(x_k) - f_k(p_k)] \geq \frac{\rho_1 \epsilon}{2} \min \left(\|x_{k+1} - x_k\|_2, \frac{\epsilon}{\beta} \right),$$

wobei wie oben wieder $\beta := \max_{x \in L_0} \|\nabla^2 f(x)\|_2$ gesetzt wurde. Da die Folge $\{f(x_k)\}$ konvergiert, schließen wir hieraus, daß

$$f(x_k) - f(x_{k+1}) \geq \frac{\rho_1 \epsilon}{2} \|x_{k+1} - x_k\|_2, \quad k(i) \leq k < l(i),$$

für alle hinreichend großen i (für nicht erfolgreiche k ist diese Ungleichung trivial). Daher ist

$$\frac{\rho_1 \epsilon}{2} \|x_{k(i)} - x_{l(i)}\|_2 \leq \frac{\rho_1 \epsilon}{2} \sum_{k=k(i)}^{l(i)-1} \|x_{k+1} - x_k\|_2 \leq \sum_{k=k(i)}^{l(i)-1} [f(x_k) - f(x_{k+1})] = f(x_{k(i)}) - f(x_{l(i)})$$

für alle hinreichend großen i . Da die Folge $\{f(x_k)\}$ konvergiert, insbesondere also eine Cauchy-Folge ist, konvergiert hier die rechte Seite gegen Null. Daher konvergiert $\{\|x_{k(i)} - x_{l(i)}\|_2\}$ und wegen der vorausgesetzten Lipschitzstetigkeit von $\nabla f(\cdot)$ auf L_0 auch $\{\|g_{k(i)} - g_{l(i)}\|_2\}$ gegen Null. Andererseits ist

$$\|g_{k(i)} - g_{l(i)}\|_2 \geq \|g_{k(i)}\|_2 - \|g_{l(i)}\|_2 \geq 2\epsilon - \epsilon = \epsilon,$$

womit wir den gewünschten Widerspruch erhalten und den ersten Teil des Satzes bewiesen haben.

Mit $\lambda_{\min}[A]$ sei der kleinste Eigenwert der symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$ bezeichnet. Wir zeigen im Anschluß durch Widerspruch, daß

$$\lambda^* := \limsup_{k \rightarrow \infty} \lambda_{\min}[B_k] \geq 0.$$

Wenn das gelungen ist, so wissen wir die Existenz einer Teilfolge $\{x_{k(i)}\} \subset \{x_k\}$ mit $\lim_{i \rightarrow \infty} \lambda_{\min}[\nabla^2 f(x_{k(i)})] = \lambda^* \geq 0$. Ein Häufungspunkt x^* der Folge $\{x_{k(i)}\}$ ist wegen des ersten Teiles des Satzes notwendig stationär (d. h. $\nabla f(x^*) = 0$), wegen $\lambda_{\min}[\nabla^2 f(x^*)] = \lambda^* \geq 0$ ist $\nabla^2 f(x^*)$ positiv semidefinit. Damit wird dann auch der zweite Teil des Satzes bewiesen sein.

Angenommen, es gibt ein $\epsilon > 0$ mit $\lambda_{\min}[B_k] \leq -\epsilon$ für alle hinreichend großen k . Sei q_k ein durch $\|q_k\|_2 = \Delta_k$ und $g_k^T q_k \leq 0$ normierter Eigenvektor zum Eigenwert $\lambda_{\min}[B_k]$. Da p_k eine globale Lösung von (P_k) ist, erhalten wir

$$f_k(p_k) \leq f_k(q_k) = f(x_k) + \underbrace{g_k^T q_k}_{\leq 0} + \frac{1}{2} q_k^T B_k q_k \leq f(x_k) - \frac{\epsilon}{2} \Delta_k^2$$

und damit

$$(*) \quad f(x_k) - f_k(p_k) \geq \frac{\epsilon}{2} \Delta_k^2$$

für alle hinreichend großen k . Hieraus folgt $\sum_{k=0}^{\infty} \Delta_k^2 < \infty$ mit derselben Argumentation wie zu Beginn des Beweises beim Nachweis von $\sum_{k=0}^{\infty} \Delta_k < \infty$. Mit $\{\Delta_k\}$ ist wegen $\|p_k\|_2 \leq \Delta_k$ auch $\{\|p_k\|_2\}$ eine Nullfolge. Hieraus folgern wir, ähnlich wie im vorbereitenden Teil des Beweises, daß $\lim_{k \rightarrow \infty} r_k = 1$. Denn wegen $(*)$ ist für alle hinreichend großen k

$$\begin{aligned} |r_k - 1| &= \left| \frac{f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k)}{f(x_k) - f_k(p_k)} \right| \\ &\leq \frac{2}{\epsilon \|p_k\|_2^2} \left| f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k) \right| \\ &= \frac{1}{\epsilon \|p_k\|_2^2} |p_k^T [\nabla^2 f(x_k) - \nabla^2 f(x_k + \theta_k p_k)] p_k| \quad \text{mit } \theta_k \in (0, 1) \\ &\leq \frac{1}{\epsilon} \|\nabla^2 f(x_k) - \nabla^2 f(x_k + \theta_k p_k)\|_2. \end{aligned}$$

Hieraus folgt $\lim_{k \rightarrow \infty} r_k = 1$, damit $r_k \geq \rho_2$ und $\Delta_{k+1} \geq \Delta_k$ für alle hinreichend großen k . Das ist ein Widerspruch zu $\lim_{k \rightarrow \infty} \Delta_k = 0$. Der zweite Teil des Satzes ist schließlich bewiesen.

Sei für den Rest des Beweises x^* ein Häufungspunkt der Folge $\{x_k\}$ mit der Eigenschaft, daß $\nabla^2 f(x^*)$ positiv definit ist, so daß in x^* die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt sind.

Zunächst wird die Konvergenz der Folge $\{x_k\}$ gegen x^* nachgewiesen. Da $\nabla^2 f(x^*)$ positiv definit ist, gibt es positive Konstanten c und δ mit $c \leq \lambda_{\min}[\nabla^2 f(x)]$ für alle $x \in B[x^*; \delta]$, der (euklidischen) Kugel um x^* mit Radius δ . Ein $\epsilon \in (0, \delta]$ sei beliebig vorgegeben. Da x^* Häufungspunkt von $\{x_k\}$ ist und nach dem ersten Teil des Satzes $\lim_{k \rightarrow \infty} \|g_k\|_2 = 0$ gilt, gibt es ein $l \in \mathbb{N}$ mit

$$\|x_l - x^*\|_2 \leq \frac{\epsilon}{2}, \quad \|g_k\|_2 \leq \frac{c\epsilon}{4} \quad \text{für alle } k \geq l.$$

Wir wollen $\|x_k - x^*\|_2 \leq \frac{1}{2}\epsilon$ für alle $k \geq l$ zeigen, womit die Konvergenz der Folge $\{x_k\}$ gegen x^* bewiesen sein wird. Der Beweisgang hierfür ist:

$$\|x_k - x^*\|_2 \leq \frac{\epsilon}{2}, \quad k \geq l \implies \|p_k\|_2 \leq \frac{\epsilon}{2} \implies \|x_{k+1} - x^*\|_2 \leq \epsilon \implies \|x_{k+1} - x^*\|_2 \leq \frac{\epsilon}{2}.$$

Sei also $k \geq l$ und $\|x_k - x^*\| \leq \frac{1}{2}\epsilon$. Da p_k Lösung des Hilfsproblems (P_k) und $p = 0$ hierfür zulässig ist, erhalten wir

$$g_k^T p_k + \frac{1}{2} c \|p_k\|_2^2 \leq g_k^T p_k + \frac{1}{2} p_k^T B_k p_k \leq 0$$

und hieraus mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\frac{1}{2} c \|p_k\|_2 \leq \|g_k\|_2 \leq \frac{1}{4} c \epsilon,$$

also $\|p_k\|_2 \leq \frac{1}{2}\epsilon$. Daher ist wegen $\|x_{k+1} - x_k\|_2 \leq \|p_k\|_2$ und der Dreiecksungleichung $\|x_{k+1} - x^*\|_2 \leq \epsilon$. Aus

$$\frac{1}{2} c \|x^* - x_{k+1}\|_2^2 + g_{k+1}^T (x^* - x_{k+1}) \leq f(x^*) - f(x_{k+1}) \leq 0$$

erhält man ebenso $\|x_{k+1} - x^*\|_2 \leq \frac{1}{2}\epsilon$. Also konvergiert $\{x_k\}$ gegen x^* .

Nun zeigen wir $\lim_{k \rightarrow \infty} r_k = 1$, womit bewiesen sein wird, daß sogar der verschärzte Test $r_k \geq \rho_2$ für alle hinreichend großen k erfüllt ist. Eben haben wir u. a. erhalten, daß eine Konstante $c > 0$ mit $\frac{1}{2}c \|p_k\|_2 \leq \|g_k\|_2$ für alle hinreichend großen k existiert. Berücksichtigt man noch $\|p_k\|_2 \leq \Delta_k$, die Abschätzung in Lemma 5.1 und die Beschränktheit von $\|B_k\|_2$, so erhält man die Existenz einer Konstanten $c_0 > 0$ mit

$$f(x_k) - f_k(p_k) \geq \frac{1}{2} \|g_k\|_2 \min\left(\Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2}\right) \geq c_0 \|p_k\|_2^2$$

für alle hinreichend großen k . Wie oben folgt hieraus $\lim_{k \rightarrow \infty} r_k = 1$.

Bis auf endlich viele vergebliche Versuche sind alle Iterationsschritte erfolgreich, für alle hinreichend großen k ist sogar $r_k \geq \rho_2$, so daß nach endlich vielen Schritten der Trust-Region-Radius nicht mehr verkleinert wird und daher $\inf_{k=0,1,\dots} \Delta_k > 0$ gilt. Andererseits ist $x_{k+1} = x_k + p_k$ für alle hinreichend großen k . Die Konvergenz der Folge $\{x_k\}$ liefert $\lim_{k \rightarrow \infty} \|p_k\|_2 = 0$, so daß $\|p_k\|_2 < \Delta_k$ für alle hinreichend großen k . Wegen Lemma 5.1 ist $B_k p_k = -g_k$ für alle hinreichend großen k . Nach endlich vielen Schritten geht die Trust-Region-Modifikation des Newton-Verfahrens in das ungedämpfte Newton-Verfahren über und erbt deswegen dessen Konvergenzeigenschaften.

Damit ist der Satz endlich bewiesen. \square

Bemerkungen: Etliche Varianten zu Satz 5.2 sind denkbar, siehe insbesondere G. A. SHULTZ, R. B. SCHNABEL, R. H. BYRD (1985). Verzichtet man z. B. darauf, daß $B_k = \nabla^2 f(x_k)$, in jedem erfolgreichen Iterationsschritt also die Hessesche der Zielfunktion neu zu berechnen ist, und setzt statt dessen voraus, daß $\{B_k\} \subset \mathbb{R}^{n \times n}$ eine beschränkte Folge symmetrischer Matrizen ist, so gilt immer noch $\lim_{k \rightarrow \infty} \|g_k\|_2 = 0$.

Auch in diesem Falle ist jeder Häufungspunkt der durch das Verfahren erzeugten Folge $\{x_k\}$ eine stationäre Lösung, und $f(x_k) = f_k(p_k)$ impliziert noch, daß x_k stationär ist. Dies erkennt man einfach durch Inspektion der ersten beiden Beweisteile des obigen Satzes.

Ferner kann man sich überlegen, wie exakt die Hilfsprobleme (P_k) zu lösen sind, um immer noch Aussagen der gewünschten Form zu erhalten. Schließlich kann es sinnvoll sein, statt der Trust-Region-Kugeln $\|p\|_2 \leq \Delta_k$ bezüglich der festen euklidischen Norm $\|\cdot\|_2$ solche bezüglich einer transformierten Norm zu betrachten, mit gewissen nichtsingulären Matrizen D_k also von den Trust-Region-Kugeln $\|D_k p\|_2 \leq \Delta_k$ auszugehen. Wir verweisen hierzu lediglich auf J. J. MORÉ (1983).

Auch bei der Update-Strategie für die Radien Δ_k sind einige Modifikationen und Spezifikationen möglich oder auch notwendig. Für den obigen Konvergenzbeweis wichtig war lediglich, daß $\Delta_{k+1} \leq \sigma_1 \Delta_k$ (mit $\sigma_1 \in (0, 1)$) für einen nicht erfolgreichen Schritt (also $r_k < \rho_1$), daß $\Delta_{k+1} \leq \sigma_2 \Delta_k$ für einen erfolgreichen Schritt und Δ_{k+1} nicht verkleinert wird, wenn $r_k \geq \rho_2$. In der Literatur (siehe z. B. M. J. D. POWELL (1975) und, für halbglatte Aufgaben, K. MADSEN (1975a) und Y. YUAN (1985)) wird häufig $\Delta_{k+1} \in [\|p_k\|_2, \sigma_2 \|p_k\|_2]$ mit $\sigma_2 > 1$ gewählt, z. B. $\Delta_{k+1} := \sigma_2 \|p_k\|_2$, wenn der Test $r_k \geq \rho_2$ erfüllt ist. Diese Wahl ist durch den obigen Konvergenzbeweis nicht gedeckt, wir gehen darauf später noch ein. \square

Auch wenn die Konvergenzaussagen in Satz 5.2 sehr befriedigend sind, so ist das dort angegebene Trust-Region-Verfahren doch nur von theoretischem Interesse, wenn das Hilfsproblem

$$(P_\Delta) \quad \text{Minimiere } \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

wobei $\Delta > 0$, $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ gegeben sind, nicht effizient gelöst werden kann. Einige Bemerkungen müssen daher hierzu noch gemacht werden.

Lemma 5.1 sagt uns, daß ein $p^* \in \mathbb{R}^n$ genau dann eine globale Lösung von (P_Δ) ist, wenn ein $\lambda^* \geq 0$ existiert mit

- (a) $(B + \lambda^* I)p^* = -g$,
- (b) $\lambda^*(\|p^*\|_2 - \Delta) = 0$,
- (c) $B + \lambda^* I$ ist positiv semidefinit.

Wir wollen lediglich den Fall betrachten, daß es ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ und ein $\lambda^* \geq 0$ gibt, für die (a) und (b) erfüllt sind und $B + \lambda^* I$ sogar positiv definit ist. Dieser Fall liegt insbesondere dann vor, wenn B positiv definit ist. Ferner sei $g \neq 0$. Wegen Lemma 5.1 ist p^* in diesem Falle sogar eindeutige globale Lösung von (P) . Mit λ_{\min} bezeichnen wir den kleinsten Eigenwert von B . Die sogenannte *Levenberg-Marquardt-Trajektorie* $p: (-\lambda_{\min}, +\infty) \rightarrow \mathbb{R}^n$ sei durch

$$p(\lambda) := -(B + \lambda I)^{-1} g$$

definiert, sei $w(\lambda) := \|p(\lambda)\|_2$. Zur Berechnung von λ^* bietet sich die folgende Vorgehensweise an:

- Berechne die Cholesky-Zerlegung von B , um zu erkennen, ob B positiv definit ist. Wenn das der Fall ist, so berechne $p(0)$ durch Vorwärts- und Rückwärtseinsetzen. Wenn $\|p(0)\|_2 \leq \Delta$, so ist $p^* := p(0)$ (mit $\lambda^* = 0$) globale Lösung von (P_Δ) , STOP.
- Bestimme iterativ eine positive Lösung $\lambda^* \in (-\lambda_{\min}, +\infty)$ von $w(\lambda) = \Delta$. Mit $p^* := p(\lambda^*)$ hat man dann die globale Lösung von (P_Δ) gefunden.

Zum zweiten Punkt muß etwas mehr gesagt werden. Da $w(\cdot)$ in $-\lambda_{\min}$ singulär ist, wird nicht auf die Gleichung $w(\lambda) - \Delta = 0$, sondern auf

$$\psi(\lambda) := \frac{1}{\Delta} - \frac{1}{w(\lambda)} = 0$$

das Newton-Verfahren angewandt. Nun ist

$$w'(\lambda) = \frac{p(\lambda)^T p'(\lambda)}{\|p(\lambda)\|_2} = -\frac{p(\lambda)^T (B + \lambda I)^{-1} p(\lambda)}{\|p(\lambda)\|_2},$$

woran man erkennt, daß $w(\cdot)$ auf $(-\lambda_{\min}, +\infty)$ monoton fallend ist. Hiermit erhält man die Iterationsvorschrift

$$\lambda_+ := \lambda - \frac{\psi(\lambda)}{\psi'(\lambda)} = \lambda - \left[\frac{1}{\Delta} - \frac{1}{w(\lambda)} \right] \frac{w(\lambda)^2}{w'(\lambda)} = \lambda + \frac{w(\lambda)}{w'(\lambda)} \left[1 - \frac{w(\lambda)}{\Delta} \right].$$

Die Durchführung dieses Iterationsschrittes kann folgendermaßen verlaufen:

- Sei $\lambda \geq 0$ und $B + \lambda I$ positiv definit. Man berechne die Cholesky-Zerlegung $B + \lambda I = LL^T$.
- Berechne $p(\lambda)$ aus $LL^T p(\lambda) = -g$ durch Vorwärts- und Rückwärtseinsetzen. Anschließend berechne $w(\lambda) := \|p(\lambda)\|_2$.
- Berechne $q(\lambda)$ aus $Lq(\lambda) = p(\lambda)$ durch Vorwärtseinsetzen und anschließend

$$w'(\lambda) := -\frac{\|q(\lambda)\|_2^2}{w(\lambda)}.$$

- Berechne λ_+ aus

$$\lambda_+ := \lambda + \frac{w(\lambda)}{w'(\lambda)} \left[1 - \frac{w(\lambda)}{\Delta} \right].$$

Auf nähere Einzelheiten (Wahl eines Startwertes für obiges Verfahren zur Berechnung von λ^* und Untersuchung des Falles, in dem $B + \lambda^* I$ lediglich positiv semidefinit ist) soll nicht mehr eingegangen werden. Hierzu verweisen wir lediglich auf J. E. DENNIS, R. B. SCHNABEL (1983, S. 134 ff.) und D. C. SORENSEN (1982).

7.5.3 Nichtlineare Ausgleichsprobleme

In diesem Unterabschnitt betrachten wir das nichtlineare Ausgleichsproblem

$$(P) \quad \text{Minimiere } f(x) := \frac{1}{2} \|F(x)\|_2^2, \quad x \in \mathbb{R}^n,$$

wobei $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine mindestens einmal stetig differenzierbare Abbildung und $m \geq n$ ist. Als zugehörige Modell-Funktion nehmen wir naheliegenderweise

$$f_x(p) := \frac{1}{2} \|F(x) + F'(x)p\|_2^2,$$

so daß das beim Trust-Region-Verfahren auftretende Hilfsproblem die Form

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := f(x) + [F'(x)^T F(x)]^T p + \frac{1}{2} p^T F'(x)^T F'(x)p, \quad \|p\|_2 \leq \Delta$$

hat. Der Unterschied zu der Situation im letzten Abschnitt besteht u. a. darin, daß hier die Matrix $B := F'(x)^T F'(x)$ nicht nur symmetrisch, sondern auch positiv semidefinit ist. Daher ist eine lokale Lösung p^* von $(P_{x,\Delta})$ auch eine globale Lösung, eine solche ist nach Lemma 5.1 charakterisiert durch die Existenz einer Konstanten $\lambda^* \geq 0$ mit

- (a) $[F'(x)^T F'(x) + \lambda^* I]p^* = -F'(x)^T F(x),$
- (b) $\lambda^*(\|p^*\|_2 - \Delta) = 0.$

Die Bedingung (c) in Lemma 5.1 ist hier natürlich überflüssig. Ähnlich wie am Schluß von 7.5.2 definieren wir die *Levenberg-Marquardt-Trajektorie* $p: [0, +\infty) \rightarrow \mathbb{R}^n$ durch

$$p(\lambda) := \begin{cases} -F'(x)^+ F(x) & \text{falls } \lambda = 0 \text{ und Rang } F'(x) < n, \\ -[F'(x)^T F'(x) + \lambda I]^{-1} F'(x)^T F(x) & \text{sonst.} \end{cases}$$

Hierbei bedeutet $F'(x)^+ \in \mathbb{R}^{n \times m}$ die Pseudoinverse von $F'(x) \in \mathbb{R}^{m \times n}$, so daß $p(0) := -F'(x)^+ F(x)$ die eindeutige Lösung *minimaler euklidischer Norm* des linearen Ausgleichsproblems

$$\text{Minimiere } \|F(x) + F'(x)p\|_2, \quad p \in \mathbb{R}^n$$

ist. Ist $\lambda > 0$ oder besitzt $F'(x)$ vollen Rang, so kann man $p(\lambda)$ natürlich mit Hilfe einer Cholesky-Zerlegung von $F'(x)^T F'(x) + \lambda I$ als Lösung von

$$[F'(x)^T F'(x) + \lambda I]p = -F'(x)^T F(x)$$

bestimmen. Erkennt man aber, daß dieses Gleichungssystem genau mit den Normalgleichungen zum linearen Ausgleichsproblem

$$\text{Minimiere } \frac{1}{2} \left\| \begin{pmatrix} F(x) \\ 0 \end{pmatrix} + \begin{pmatrix} F'(x) \\ \sqrt{\lambda} I \end{pmatrix} p \right\|_2^2, \quad p \in \mathbb{R}^n$$

übereinstimmt, so liegt es näher, $p(\lambda)$ mit Hilfe von QR-Zerlegungen zu berechnen. Hierbei sollte natürlich die hier vorliegende spezielle Form der Koeffizientenmatrix berücksichtigt werden. Der Einfachheit halber wollen wir nur den unkritischen Fall betrachten, daß nämlich Rang $F'(x) = n$ gilt. Die Berechnung von $p(\lambda)$ kann dann folgendermaßen vorgenommen werden:

- Berechne eine orthogonale Matrix $Q_1 \in \mathbb{R}^{m \times m}$ und eine (bei vollem Rang von $F'(x)$) nichtsinguläre, obere Dreiecksmatrix $R_1 \in \mathbb{R}^{n \times n}$ mit

$$Q_1^T F'(x) = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}, \quad Q_1^T F(x) =: \begin{pmatrix} c_1 \\ d_1 \end{pmatrix},$$

d.h. man bestimme eine QR -Zerlegung von $F'(x)$ und wende Q_1^T gleich auf $F(x)$ an, damit man sich Q_1 nicht zu merken braucht. Dieser Schritt ist von λ unabhängig.

- Ist $\lambda = 0$, so kann man $p(0)$ als Lösung von $c_1 + R_1 p = 0$ erhalten.
- Sei $\lambda > 0$. Wegen

$$\begin{pmatrix} Q_1^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} F'(x) \\ \sqrt{\lambda} I \end{pmatrix} = \begin{pmatrix} R_1 \\ 0 \\ \sqrt{\lambda} I \end{pmatrix}$$

hat man schon fast die gewünschte QR -Zerlegung erhalten. Lediglich die Diagonale im unteren Block stört die obere Dreiecksgestalt der Matrix auf der rechten Seite. Deren QR -Zerlegung kann aber offenbar in naheliegender Weise mit Hilfe von $n(n+1)/2$ Givens-Rotationen berechnet werden. Faßt man das Produkt der Givens-Rotationen in der orthogonalen Matrix Q_2^T zusammen, so berechnet man also

$$Q_2^T \begin{pmatrix} R_1 \\ 0 \\ \sqrt{\lambda} I \end{pmatrix} = \begin{pmatrix} R_2 \\ 0 \\ 0 \end{pmatrix}, \quad Q_2^T \begin{pmatrix} c_1 \\ d_1 \\ 0 \end{pmatrix} = \begin{pmatrix} c_2 \\ d_2 \\ e_2 \end{pmatrix}$$

und anschließend $p = p(\lambda)$ aus $c_2 + R_2 p = 0$.

Nachdem nun klar ist (jedenfalls für den Fall, daß $F'(x)$ vollen Rang besitzt), wie man bei gegebenem $\lambda \geq 0$ das zugehörige $p(\lambda)$ berechnet, ist es auch naheliegend, daß man zur Berechnung von λ^* im Prinzip wie in dem allgemeineren Fall zum Schluß von 7.5.2 vorgehen kann. Wir wollen hierauf nicht näher eingehen, sondern verweisen hierzu nur auf J. J. MORÉ (1978). Dort wird auch auf weitere Feinheiten einer effizienten Implementation des zugehörigen *Levenberg-Marquardt-Verfahrens* eingegangen (z. B. Skalierung, Update-Regeln für die Trust-Region-Radien) und ein Konvergenzsatz formuliert.

7.5.4 Diskrete, nichtlineare Approximationsaufgaben

Als Beispiel einer halbglatten, unrestringierten Optimierungsaufgabe betrachten wir in diesem Unterabschnitt die diskrete, nichtlineare Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei ist $F = (F_i) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ eine hinreichend glatte Abbildung und $\|\cdot\|$ eine Norm auf dem \mathbb{R}^m . Die Zielfunktion f läßt sich auch als $f = g \circ F$ mit $g(y) := \|y\|$

schreiben. Diese Darstellung und die Konvexität von g werden oft ausgenutzt. Ziel wird es sein, ein von K. MADSEN (1975a) für diskrete Tschebyscheffsche Approximationsaufgaben (hier ist also $\|\cdot\| := \|\cdot\|_\infty$) entwickeltes Trust-Region-Verfahren anzugeben und seine Konvergenz zu beweisen. Wir werden versuchen, weitgehend so zu argumentieren, daß der allgemeine Hintergrund sichtbar wird und es einfach ist, die Vorgehensweise auf andere Aufgaben zu übertragen.

Als Modell-Funktion $f_x(\cdot)$ für $f(x + \cdot)$ nehmen wir

$$f_x(p) := \|F(x) + F'(x)p\|,$$

und betrachten als Hilfsproblem bei gegebenem $\Delta > 0$ die Aufgabe

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|, \quad \|p\| \leq \Delta.$$

Hier werden die Trust-Region-Kugeln bezüglich einer Norm $\|\cdot\|$ auf dem \mathbb{R}^n betrachtet.

Beispiel: Ist die gegebene Norm auf dem \mathbb{R}^n die Maximumnorm, handelt es sich also bei (P) um eine diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe, so ist es zweckmäßig, als Norm auf dem \mathbb{R}^n ebenfalls die Maximumnorm zu wählen. Wie schon mehrfach erwähnt, läßt sich dann das Hilfsproblem

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|_\infty, \quad \|p\|_\infty \leq \Delta$$

als äquivalentes lineares Programm

$$\begin{aligned} &\text{Minimiere } \delta \text{ unter den Nebenbedingungen} \\ &\delta \geq 0, \quad -\delta e \leq F(x) + F'(x)p \leq \delta e, \quad -\Delta e \leq p \leq \Delta e \end{aligned}$$

schreiben, wobei e der Vektor im \mathbb{R}^m bzw. \mathbb{R}^n ist, dessen Komponenten alle gleich 1 sind. Das Hilfsproblem $(P_{x,\Delta})$ kann also durch eine geeignete Modifikation des Simplexverfahrens gelöst werden (siehe auch Aufgabe 6). \square

Als Abbruchbedingung wird im folgenden Algorithmus wieder $f(x) = f_x(p^*)$ auftreten, wobei p^* eine Lösung von $(P_{x,\Delta})$ ist. Das folgende Lemma sagt aus, daß dies genau dann der Fall ist, wenn x stationäre Lösung von (P) ist, also $f'(x;p) \geq 0$ für alle $p \in \mathbb{R}^n$ gilt.

Lemma 5.3 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P) . Die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ sei in $x \in \mathbb{R}^n$ stetig differenzierbar. Mit einem vorgegebenem $\Delta > 0$ sei $p^* \in \mathbb{R}^n$ eine Lösung von

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|, \quad \|p\| \leq \Delta.$$

Dann ist x genau dann eine stationäre Lösung von (P) , wenn $f(x) = f_x(p^*)$.

Beweis: Sei x eine stationäre Lösung von (P) . Insbesondere ist (siehe Definition 1.5)

$$0 \leq f'(x; p^*) = g'(F(x); F'(x)p^*) \leq g(F(x) + F'(x)p^*) - g(F(x)) = f_x(p^*) - f(x),$$

wobei für die Abschätzung Lemma 1.6 benutzt wurde und die konvexe, lipschitzstetige Abbildung $g: \mathbb{R}^m \rightarrow \mathbb{R}$ durch $g(y) := \|y\|$ definiert ist. Also ist

$$f(x) \leq f_x(p^*) \leq f_x(0) = f(x)$$

und daher $f(x) = f_x(p^*)$.

Sei umgekehrt $f(x) = f_x(p^*)$. Dann ist auch $p^{**} := 0$ eine Lösung von $(P_{x,\Delta})$. Für alle $p \in \mathbb{R}^n$ ist daher

$$0 \leq f'_x(0; p) = \lim_{t \rightarrow 0+} \frac{g(F(x) + tF'(x)p) - g(F(x))}{t} = g'(F(x); F'(x)p) = f'(x; p).$$

Also ist x eine stationäre Lösung von (P) . \square

Es wird zweckmäßig sein, eine Bezeichnung einzuführen. Bei gegebenen $x \in \mathbb{R}^n$ und $\Delta > 0$ verstehen wir unter $v(x, \Delta)$ den *Optimalwert* von $(P_{x,\Delta})$, d. h. es ist

$$v(x, \Delta) := \min\{\|F(x) + F'(x)p\| : \|p\| \leq \Delta\}.$$

Ist $x^* \in \mathbb{R}^n$ keine stationäre Lösung von (P) , so ist $f(x^*) - v(x^*, \Delta) > 0$ für alle $\Delta > 0$ (siehe Lemma 5.3). Das folgende Lemma impliziert insbesondere, daß alle Punkte x aus einer hinreichend kleinen Umgebung eines nichtstationären Punktes x^* ebenfalls nicht stationär sind.

Lemma 5.4 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P) . Ein $\Delta^* > 0$ sei vorgegeben. Dann gilt:

1. Ist F in x differenzierbar, so ist

$$f(x) - v(x, \Delta) \geq \frac{\Delta}{\Delta^*} [f(x) - v(x, \Delta^*)] \quad \text{für alle } \Delta \in (0, \Delta^*].$$

2. Ist F in x^* stetig differenzierbar und x^* keine stationäre Lösung von (P) , so gibt es ein $\delta > 0$ mit

$$f(x) - v(x, \Delta^*) \geq \frac{1}{2} [f(x^*) - v(x^*, \Delta^*)] \quad \text{für alle } x \text{ mit } \|x - x^*\| \leq \delta.$$

Beweis: Sei $\Delta \in (0, \Delta^*]$ beliebig gewählt und p^* eine Lösung von (P_{x,Δ^*}) , also $\|p^*\| \leq \Delta^*$ und $\|F(x) + F'(x)p^*\| = v(x, \Delta^*)$. Dann ist $p := \lambda p^*$ mit $\lambda := \Delta/\Delta^*$ zulässig für $(P_{x,\Delta})$ und daher

$$\begin{aligned} f(x) - v(x, \Delta) &\geq \|F(x)\| - \|F(x) + \lambda F'(x)p^*\| \\ &\geq \|F(x)\| - [(1 - \lambda)\|F(x)\| + \lambda\|F(x) + F'(x)p^*\|] \\ &\quad (\text{wegen der Konvexität von } \|\cdot\| \text{ bzw. von } g(\cdot)) \\ &= \lambda [\|F(x)\| - \|F(x) + F'(x)p^*\|] \\ &= \frac{\Delta}{\Delta^*} [f(x) - v(x, \Delta^*)], \end{aligned}$$

womit der erste Teil des Lemmas bewiesen ist.

Nun sei $p^* \in \mathbb{R}^n$ eine feste Lösung von (P_{x^*, Δ^*}) . Die Abbildung

$$x \mapsto \|F(x)\| - \|F(x) + F'(x)p^*\|$$

ist in x^* stetig, dort hat sie den Wert $f(x^*) - v(x^*, \Delta^*) > 0$ (nach Voraussetzung ist x^* keine stationäre Lösung von (P)). Daher gibt es ein $\delta > 0$ derart, daß

$$f(x) - v(x, \Delta^*) \geq \|F(x)\| - \|F(x) + F'(x)p^*\| \geq \frac{f(x^*) - v(x^*, \Delta^*)}{2}$$

für alle x mit $\|x - x^*\| \leq \delta$. Damit ist der zweite Teil des Lemmas bewiesen. \square

In dem folgenden Satz geben wir das Madsen-Verfahren (in fast genau derselben Form wie bei K. MADSEN (1975a), dort ist allerdings $\|\cdot\| := \|\cdot\|_\infty$) an und beweisen (unter unnötig starken Voraussetzungen an die Zielfunktion) eine Konvergenzaussage.

Satz 5.5 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Mit dem Startwert x_0 des gleich anzugebenden Verfahrens sei die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt. Die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ sei auf einer Umgebung von L_0 stetig differenzierbar, die Funktionalmatrix $F'(\cdot)$ sei dort lipschitzstetig. Man betrachte das folgende Verfahren.

- Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$ und $0 < \sigma_1 < 1 < \sigma_2$.
(Bei K. MADSEN (1975a) wird $\rho_1 := 0.01$, $\rho_2 := 0.25$, $\sigma_1 := 0.25$ und $\sigma_2 := 2$ empfohlen.)
- Gegeben seien $x_0 \in \mathbb{R}^n$ und $\Delta_0 > 0$. Berechne $F(x_0)$ und $F'(x_0)$.
- Für $k = 0, 1, \dots$:

Berechne eine Lösung p_k der Aufgabe

$$(P_k) \quad \text{Minimiere } f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad \|p_k\| \leq \Delta_k.$$

Falls $f(x_k) = f_k(p_k)$, dann: STOP, x_k ist eine stationäre Lösung von (P) .

Berechne

$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$

Falls $r_k \geq \rho_1$, dann setze $x_{k+1} := x_k + p_k$ und berechne $F(x_{k+1})$ und $F'(x_{k+1})$.

Andernfalls setze $x_{k+1} := x_k$.

- Falls $r_k \leq \rho_2$, dann setze $\Delta_{k+1} := \sigma_1 \|p_k\|$.
- Falls $r_k > \rho_2$, dann wähle $\Delta_{k+1} \in [\|p_k\|, \sigma_2 \|p_k\|]$.
(Bei K. MADSEN (1975a) wird $\Delta_{k+1} := \sigma_2 \|p_k\|$ gesetzt, wenn ein weiterer Test

$$\|F(x_k + p_k) - F(x_k) - F'(x_k)p_k\| \leq \rho_3 [f(x_k) - f(x_k + p_k)]$$

mit einem $\rho_3 \in (0, 1)$, etwa $\rho_3 := 0.25$, erfüllt ist. Wenn das nicht der Fall ist, wird $\Delta_{k+1} := \|p_k\|$ gesetzt.)

Das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Dann liefert es eine Folge $\{x_k\}$ mit der Eigenschaft, daß jeder Häufungspunkt von $\{x_k\}$ eine stationäre Lösung von (P) ist.

Beweis: Zunächst beachten wir, daß die Folge $\{\Delta_k\} \subset \mathbb{R}_+$ beschränkt ist, da es die Folge $\{x_k\}$ ist, und Δ_{k+1} im k -ten Iterationsschritt nur dann vergrößert werden kann, wenn für diesen $r_k > \rho_2$ und damit $\Delta_{k+1} \leq \sigma_2 \|p_k\| = \sigma_2 \|x_{k+1} - x_k\|$ gilt. Daher existiert ein $\Delta^* > 0$ mit $0 < \Delta_k \leq \Delta^*$ für $k = 0, 1, \dots$. Der Beweis zerfällt in zwei Teile.

- (1) Konvergiert die durch das Verfahren erzeugte Folge $\{x_k\}$ gegen einen Punkt x^* , so ist x^* eine stationäre Lösung von (P).

Wir machen einen Widerspruchsbeweis und nehmen an, x^* sei keine stationäre Lösung. Wegen $\lim_{k \rightarrow \infty} x_k = x^*$ und Lemma 5.4 existiert eine positive Konstante c_0 mit

$$(*) \quad f(x_k) - f_k(p_k) \geq c_0 \Delta_k \quad \text{für alle hinreichend großen } k.$$

Wie im Beweis zu Satz 5.2 folgt hieraus $\sum_{k=0}^{\infty} \Delta_k < \infty$. Insbesondere ist $\{\Delta_k\}$ und damit auch $\{\|p_k\|\}$ eine Nullfolge. Zum Beweis von (1) machen wir eine Fallunterscheidung.

- (i) Es gibt eine unendliche Teilmenge $K \subset \mathbb{N}$ mit $\|p_k\| < \Delta_k$ für alle $k \in K$.

Sei $k \in K$ fest. Wir überlegen uns, daß $f_k(p_k) \leq f_k(p)$ für alle $p \in \mathbb{R}^n$. Denn nach Definition von p_k ist trivialerweise $f_k(p_k) \leq f_k(p)$ für alle p mit $\|p\| \leq \Delta_k$. Ist dagegen $\|p\| > \Delta_k$, so bestimme man ein $\lambda \in (0, 1)$ mit $\|(1 - \lambda)p_k + \lambda p\| = \Delta_k$, nutze die Konvexität von f_k aus:

$$f_k(p_k) \leq f_k((1 - \lambda)p_k + \lambda p) \leq (1 - \lambda)f_k(p_k) + \lambda f_k(p)$$

und schließe hieraus auf $f_k(p_k) \leq f_k(p)$. Läßt man $k \in K$ gegen unendlich streben, so folgt wegen $x_k \rightarrow x^*$ und $p_k \rightarrow 0$, daß $f(x^*) \leq f_{x^*}(p)$ für alle $p \in \mathbb{R}^n$. Wegen Lemma 5.3 ist x^* eine stationäre Lösung von (P), ein Widerspruch zu der Annahme, daß das gerade nicht der Fall ist.

- (ii) Für alle hinreichend großen k ist $\|p_k\| = \Delta_k$.

Wir zeigen, daß $\lim_{k \rightarrow \infty} r_k = 1$. Dann ist $r_k > \rho_2$ für alle hinreichend großen k und daher $\|p_{k+1}\| = \Delta_{k+1} \geq \|p_k\|$ für alle hinreichend großen k , ein Widerspruch zu $p_k \rightarrow 0$.

Wegen (*) und (ii) ist für alle hinreichend großen k

$$\begin{aligned} |r_k - 1| &= \left| \frac{\|F(x_k + p_k)\| - \|F(x_k) + F'(x_k)p_k\|}{f(x_k) - f_k(p_k)} \right| \\ &\leq \frac{\|F(x_k + p_k) - F(x_k) - F'(x_k)p_k\|}{c_0 \|p_k\|} \\ &\leq \frac{1}{c_0} \int_0^1 \|F'(x_k + tp_k) - F'(x_k)\| dt. \end{aligned}$$

Wegen $p_k \rightarrow 0$ und $x_k \rightarrow x^*$ folgt $\lim_{k \rightarrow \infty} r_k = 1$. Damit ist (1) bewiesen.

(2) Jeder Häufungspunkt x^* der Folge $\{x_k\}$ ist eine stationäre Lösung von (P).

Angenommen, x^* sei ein Häufungspunkt der Folge $\{x_k\}$, der keine stationäre Lösung von (P) ist. Eine Anwendung von Lemma 5.4 zeigt die Existenz positiver Konstanten δ und c_0 mit

$$\|x_k - x^*\| \leq \delta \implies f(x_k) - f_k(p_k) \geq c_0 \|p_k\|.$$

Wegen Beweisteil (1) wissen wir, daß nicht die gesamte Folge $\{x_k\}$ gegen x^* konvergiert, so daß angenommen werden kann, daß es unendlich viele k mit $\|x_k - x^*\| > \delta$ gibt (notfalls verkleinere man δ). Da ferner x^* ein Häufungspunkt von $\{x_k\}$ ist, gibt es in jeder Umgebung von x^* , insbesondere in der $\delta/2$ -Kugel um x^* , unendlich viele Elemente von $\{x_k\}$. Daher existieren Folgen $\{k(i)\}_{i \in \mathbb{N}}$ und $\{l(i)\}_{i \in \mathbb{N}}$ in \mathbb{N} mit $k(i) < l(i) < k(i+1) < l(i+1)$ und

$$\|x_{k(i)} - x^*\| \leq \frac{\delta}{2}, \quad \|x_k - x^*\| \leq \delta \quad \text{für } k(i) < k < l(i), \quad \|x_{l(i)} - x^*\| > \delta.$$

Hieraus folgt: Ist $k(i) \leq k < l(i)$, so ist

$$f(x_k) - f(x_{k+1}) \geq \rho_1 c_0 \|x_{k+1} - x_k\|, \quad k(i) \leq k < l(i),$$

und damit

$$\begin{aligned} \rho_1 c_0 \|x_{k(i)} - x_{l(i)}\| &\leq \rho_1 c_0 \sum_{k=k(i)}^{l(i)-1} \|x_{k+1} - x_k\| \\ &\leq \sum_{k=k(i)}^{l(i)-1} [f(x_k) - f(x_{k+1})] \\ &= f(x_{k(i)}) - f(x_{l(i)}). \end{aligned}$$

Da die Folge $\{f(x_k)\}$ konvergiert, insbesondere eine Cauchy-Folge ist, konvergiert die rechte Seite und damit auch $\{\|x_{k(i)} - x_{l(i)}\|\}$ gegen Null. Andererseits ist nach Konstruktion

$$\|x_{k(i)} - x_{l(i)}\| \geq \|x_{l(i)} - x^*\|_2 - \|x_{k(i)} - x^*\|_2 > \delta - \frac{\delta}{2} = \frac{\delta}{2},$$

ein Widerspruch. Der Satz ist damit bewiesen. \square

Nun soll noch etwas zur Konvergenzgeschwindigkeit des Madsen-Verfahrens ausgesagt werden. Entscheidend hierfür ist wieder der Begriff der *lokalen starken Eindeutigkeit* (siehe Definition 2.12, die Bemerkungen im Anschluß an den Beweis zu Satz 2.13 sowie Satz 2.14). Unser Ziel ist es, den folgenden Satz zu beweisen (siehe auch K. MADSEN (1975b) und C. GEIGER (1977)).

Satz 5.6 Das in Satz 5.5 angegebene Verfahren zur Lösung der diskreten, nichtlinearen Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n$$

liefere eine Folge $\{x_k\}$, die gegen ein $x^* \in \mathbb{R}^n$ konvergiert. Sei x^* lokal stark eindeutige Lösung von (P) und $F'(\cdot)$ auf einer Umgebung von x^* lipschitzstetig. Dann sind fast alle Iterationsschritte erfolgreich und die Folge $\{x_k\}$ konvergiert sogar quadratisch gegen x^* , d. h. es existiert eine Konstante $C > 0$ mit

$$\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2 \quad \text{für alle hinreichend großen } k.$$

Beweis: Es kann angenommen werden, daß die wegen der starken Eindeutigkeit von x^* existierende Konstante $\delta > 0$ so klein ist, daß $F'(\cdot)$ auf der Kugel $B[x^*; \delta]$ lipschitzstetig mit einer Lipschitzkonstanten $L > 0$ ist. Mit positiven Konstanten σ und δ ist also

$$f(x) \geq f(x^*) + \sigma \|x - x^*\|, \quad \|F'(x) - F'(y)\| \leq L \|x - y\|$$

für alle $x, y \in B[x^*; \delta]$. O. B. d. A. bricht das Verfahren nicht vorzeitig ab, so daß $x_k \neq x^*$ für alle k angenommen werden kann. Wir erinnern an das beim Madsen-Verfahren auftretende Hilfsproblem

$$(P_k) \quad \text{Minimiere } f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad \|p\| \leq \Delta_k$$

mit der Lösung p_k . Der Beweis des Satzes zerfällt in mehrere Teile.

(1) Sind $x_k, x_k + p \in B[x^*; \delta]$, so ist

$$|f(x_k + p) - f_k(p)| \leq \|F(x_k + p) - F(x_k) - F'(x_k)p\| \leq \frac{L}{2} \|p\|^2.$$

(2) Es ist $\|p_k\| \leq 2 \|x_k - x^*\|$ für alle hinreichend großen k .

Denn: Wegen der vorausgesetzten Konvergenz der Folge $\{x_k\}$ gegen x^* existiert ein $k_0 \in \mathbb{N}$ mit

$$\|x_k - x^*\| \leq \frac{\delta}{3}, \quad \sigma - \frac{5L}{2} \|x_k - x^*\| > 0 \quad \text{für alle } k \geq k_0.$$

Wir wollen zeigen:

$$(*) \quad k \geq k_0, \quad \|p\| \geq 2 \|x_k - x^*\| \implies f_k(p) > f_k(x^* - x_k).$$

Ist (*) bewiesen, so wissen wir: Ist $k \geq k_0$ und $x^* - x_k$ zulässig für (P_k) bzw. $\|x^* - x_k\| \leq \Delta_k$, so ist $\|p_k\| \leq 2 \|x_k - x^*\|$. Ist dagegen $\|x^* - x_k\| > \Delta_k$, so gilt $\|p_k\| \leq 2 \|x_k - x^*\|$ trivialerweise. Daher bleibt (*) zu zeigen.

Hierzu nehmen wir zunächst an, es sei $\|p\| = 2 \|x_k - x^*\|$. Nach (zweimaliger) Anwendung von (1) und wegen der lokal starken Eindeutigkeit von x^* ist

$$\begin{aligned} f_k(p) - f_k(x^* - x_k) &\stackrel{(1)}{\geq} -2L \|x_k - x^*\|^2 + f(x_k + p) - f_k(x^* - x_k) \\ &\geq -2L \|x_k - x^*\|^2 + \sigma \|x_k + p - x^*\| + f(x^*) - f_k(x^* - x_k) \\ &\stackrel{(1)}{\geq} -\frac{5L}{2} \|x_k - x^*\| + \sigma \|x_k + p - x^*\| \\ &\geq \left(\sigma - \frac{5L}{2} \|x_k - x^*\| \right) \|x_k - x^*\| \\ &> 0 \end{aligned}$$

für alle $k \geq k_0$. Nun sei $\|p\| > 2\|x_k - x^*\|$. Man bestimme ein $\lambda \in (0, 1)$ mit

$$\|(1-\lambda)p + \lambda(x^* - x_k)\| = 2\|x_k - x^*\|$$

und setze anschließend $q := (1-\lambda)p + \lambda(x^* - x_k)$. Für alle $k \geq k_0$ ist dann

$$f_k(x^* - x_k) < f_k(q) \leq (1-\lambda)f_k(p) + \lambda f_k(x^* - x_k)$$

(wegen der Konvexität von f_k) und daher $f_k(x^* - x_k) < f_k(p)$. Damit ist (2) bewiesen.

(3) Es existiert eine Konstante $c_0 > 0$ derart, daß $f(x_k) - f_k(p_k) \geq c_0 \|p_k\|$ für alle hinreichend großen k .

Denn: Sei $k_0 \in \mathbb{N}$ wie im Beweis von (2) bestimmt. Ist $\|x_k - x^*\| \leq \Delta_k$, also $x_k - x^*$ zulässig für (P_k) , und $k \geq k_0$, so ist wegen (1) und (2)

$$f(x_k) - f_k(p_k) \geq f(x_k) - f_k(x^* - x_k) \geq \frac{4\sigma}{5} \|x_k - x^*\| \geq \frac{2\sigma}{5} \|p_k\|.$$

Ist dagegen $\|x_k - x^*\| > \Delta_k$ und $k \geq k_0$, so ist wegen der Konvexität von f_k

$$f(x_k) - f_k(p_k) \geq \frac{\Delta_k}{\|x_k - x^*\|} [f(x_k) - f_k(x^* - x_k)] \geq \frac{4\sigma}{5} \Delta_k \geq \frac{4\sigma}{5} \|p_k\|.$$

Damit ist auch (3) bewiesen.

(4) Mit

$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}$$

ist $\lim_{k \rightarrow \infty} r_k = 1$. Insbesondere ist $r_k \geq \rho_1$ für alle hinreichend großen k , folglich sind fast alle Iterationsschritte erfolgreich und damit $x_{k+1} = x_k + p_k$ für alle hinreichend großen k .

Denn: Wegen (1)–(3) ist

$$|r_k - 1| = \left| \frac{|f(x_k + p_k) - f_k(p_k)|}{f(x_k) - f_k(p_k)} \right| \leq \frac{L}{2c_0} \|p_k\| \leq \frac{L}{c_0} \|x_k - x^*\|$$

für alle hinreichend großen k . Wegen $\lim_{k \rightarrow \infty} x_k = x^*$ folgt (4).

(5) Es ist

$$(*) \quad f_k(p_k) \leq f_k(x^* - x_k)$$

und

$$(**) \quad \|x_{k+1} - x^*\| \leq \frac{5L}{2\sigma} \|x_k - x^*\|^2$$

für alle hinreichend großen k . Insbesondere konvergiert die Folge $\{x_k\}$ von mindestens zweiter Ordnung gegen x^* .

Denn: Sei $k_0 \in \mathbb{N}$ so groß gewählt, daß

$$\|x_k - x^*\| \leq \min\left(\delta, \frac{\sigma}{5L}\right), \quad \|p_k\| \leq 2\|x_k - x^*\|, \quad r_k \geq \rho_2 \quad \text{für alle } k \geq k_0,$$

was wegen $x_k \rightarrow x^*$ sowie (2) und (4) möglich ist. Insbesondere ist $x_{k+1} = x_k + p_k$ und $\Delta_{k+1} \geq \|p_k\|$ für alle $k \geq k_0$. Letzteres liefert zusammen mit $p_k \rightarrow 0$ die Existenz eines $l \geq k_0$ mit $\|p_l\| < \Delta_l$. Wegen der Konvexität von f_l ist p_l nicht nur Minimum von f_l auf der Kugel $B[0; \Delta_l]$, sondern auf dem gesamten \mathbb{R}^n . Daher ist speziell (*) für $k = l$ richtig. Ferner ist

$$\begin{aligned} \|x_{l+1} - x^*\| &\leq \frac{1}{\sigma} [f(x_l + p_l) - f(x^*)] \\ &\stackrel{(1),(2)}{\leq} \frac{1}{\sigma} [2L\|x_l - x^*\|^2 + f_l(p_l) - f(x^*)] \\ &\leq \frac{1}{\sigma} [2L\|x_l - x^*\|^2 + f_l(x^* - x_l) - f(x^*)] \\ &\stackrel{(1)}{\leq} \frac{5L}{2\sigma} \|x_l - x^*\|^2. \end{aligned}$$

Damit ist auch (**) für $k = l$ bewiesen. Nach Wahl von k_0 folgt

$$\|x_{l+1} - x^*\| \leq \frac{1}{2} \|x_l - x^*\| \leq \frac{1}{2} [\|p_l\| + \|x_{l+1} - x^*\|],$$

damit

$$\|x_{l+1} - x^*\| \leq \|p_l\| \leq \Delta_{l+1}$$

und hieraus schließlich (*) für $k = l + 1$. In dieser Weise kann man fortfahren und erhält, daß (*) und (**) für alle $k \geq l$ gelten. Der Satz ist bewiesen⁶. \square

Bemerkungen: Die Beweise zu den Sätzen 5.5 und 5.6 wurden so geführt, daß ihre Übertragung auf andere halbglatte Aufgaben

$$\text{Minimiere } f(x) := g \circ F(x), \quad x \in \mathbb{R}^n$$

mit einer konvexen, lipschitzstetigen Abbildung $g: \mathbb{R}^m \rightarrow \mathbb{R}$ und einer glatten Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ fast offensichtlich ist.

Es kann sinnvoll sein, als Modell-Funktion zur Approximation von $f(x + \cdot)$ Funktionen $f_x(\cdot)$ der Form

$$f_x(p) := g(F(x) + F'(x)p) + \frac{1}{2} p^T B p$$

mit einer symmetrischen Matrix $B \in \mathbb{R}^{n \times n}$ zu betrachten, um auch noch Informationen zweiter Ordnung zu berücksichtigen (siehe z. B. Y. YUAN (1985)). \square

⁶Den angegebenen Beweis verdanke ich Thomas Bannert.

Aufgaben

1. Sei $p^* \in \mathbb{R}^n$ eine lokale Lösung der Aufgabe

$$(P_\Delta) \quad \text{Minimiere } \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

wobei $\Delta > 0$, $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ gegeben sind. Dann existiert ein $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$ und $\lambda^*(\|p^*\|_2 - \Delta) = 0$. Ist ferner $\|p^*\|_2 < \Delta$, so ist B positiv semidefinit. Ist dagegen $\|p^*\|_2 = \Delta$, so ist $h^T(B + \lambda^* I)h \geq 0$ für alle $h \in \mathbb{R}^n$ mit $(p^*)^T h = 0$.

Hinweis: Die Behauptung folgt aus wesentlich allgemeineren Aussagen. Wir wollen einen ad hoc Beweis andeuten.

Ist $\|p^*\|_2 < \Delta$, so spielt die Restriktion $\|p\|_2 \leq \Delta$ lokal keine Rolle. Für jedes $h \in \mathbb{R}^n$ ist $\|p^* + th\|_2 \leq \Delta$ für alle hinreichend kleinen $|t|$, die Behauptung mit $\lambda^* := 0$ ist einfach zu beweisen.

Interessant ist daher nur der zweite Fall, daß $\|p^*\|_2 = \Delta$. Angenommen, es gibt kein $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$. Etwas anders geschrieben bedeutet das, daß

$$p^* \lambda^* = -(g + B p^*), \quad \lambda^* \geq 0$$

nicht lösbar ist. Das Farkas-Lemma liefert die Existenz eines $h \in \mathbb{R}^n$ mit

$$(p^*)^T h \leq 0, \quad -(g + B p^*)^T h > 0.$$

Ist sogar $(p^*)^T h < 0$, so ist $\|p^* + th\|_2 < \Delta$ für alle hinreichend kleine $t > 0$ und daher

$$0 \leq \lim_{t \rightarrow 0+} \frac{\phi(p^* + th) - \phi(p^*)}{t} = \nabla \phi(p^*)^T h = (g + B p^*)^T h < 0,$$

ein Widerspruch. Ist dagegen $(p^*)^T h = 0$, so sei $p(t) \in \mathbb{R}^n$ durch

$$p(t) := \|p^*\|_2 \frac{p^* + th}{\|p^* + th\|_2}$$

für alle hinreichend kleinen $|t|$ definiert. Die auf einem offenen Intervall um den Nullpunkt durch $\psi(t) := \phi(p(t))$ definierte reellwertige Funktion ψ besitzt bei $t^* = 0$ ein lokales Minimum. Daher ist (hier wird $(p^*)^T h = 0$ benutzt)

$$0 = \psi'(0) = \nabla \phi(p(0))^T h = (g + B p^*)^T h < 0,$$

ein Widerspruch. Insgesamt ist die Existenz eines $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$ nachgewiesen. Gegeben sei nun ein $h \in \mathbb{R}^n$ mit $(p^*)^T h = 0$. Definiert man ψ wie oben, so ist unter Benutzung von $g + B p^* = -\lambda^* p^*$ nach einfacher Rechnung

$$0 \leq \psi''(0) = h^T B h - \frac{\|h\|_2^2}{\|p^*\|_2^2} (g + B p^*)^T p^* = h^T (B + \lambda^* I)h,$$

womit der Beweis abgeschlossen ist.

2. Sei $B \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix mit kleinstem Eigenwert λ_{\min} . Mit einem $g \in \mathbb{R}^n \setminus \{0\}$ und $\Delta > 0$ seien $\phi: (-\lambda_{\min}, +\infty) \rightarrow \mathbb{R}$ und $\psi: (-\lambda_{\min}, +\infty) \rightarrow \mathbb{R}$ durch

$$\phi(\lambda) := \|(B + \lambda I)^{-1}g\|_2, \quad \psi(\lambda) := \frac{1}{\Delta} - \frac{1}{\|(B + \lambda I)^{-1}g\|_2}$$

definiert. Man zeige, daß ϕ und ψ auf $(-\lambda_{\min}, +\infty)$ monoton fallend und konvex sind.

3. Seien $\Delta > 0$, $g \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B \in \mathbb{R}^{n \times n}$ gegeben. Man programmiere das am Schluß von 7.5.2 angegebene Verfahren, eine (globale) Lösung von

$$(P_\Delta) \quad \text{Minimiere } \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta$$

zu berechnen, das auf der Anwendung des Newton-Verfahrens auf

$$\psi(\lambda) := \frac{1}{\Delta} - \frac{1}{\|(B + \lambda I)^{-1}g\|_2}$$

basiert. Anschließend teste man das Programm an $\Delta := 0.5$ sowie

$$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad B := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Hinweis: Startet man das Verfahren mit $\lambda_0 := 3.0$, so erhält man die in Tabelle 7.8 angegebenen Werte. Hierbei ist $p(\lambda) := -(B + \lambda I)^{-1}g$. Bessere Ergebnisse erhält man,

k	λ_k	$p(\lambda_k)$	$\ p(\lambda_k)\ _2$
0	3.0000000000000	-0.157894736842	0.210526315789
1	0.554347826087	-0.186627421344	0.523288652002
2	0.736875374131	-0.196311042319	0.462721142605
3	0.747508390383	-0.196645773070	0.459714088557
4	0.747535241295	-0.196646592909	0.459706555901
5	0.747535241461	-0.196646592914	0.459706555854

Tabelle 7.8: Ergebnisse zu Aufgabe 3

wenn man sich vorher Gedanken über einen geeigneten Startwert λ_0 macht. Z. B. ist es nicht schwer, eine obere Schranke u für die gesuchte Nullstelle λ^* zu erhalten. Denn es ist

$$\Delta = \|(B + \lambda^* I)^{-1}g\|_2 \leq \frac{\|g\|_2}{\lambda^*}$$

und daher $\lambda^* \leq \|g\|_2/\Delta =: u$. Da man am Anfang sowieso testet, ob nicht $\lambda^* = 0$ bzw. für $p(0) = -B^{-1}g$ die Restriktion $\|p(0)\|_2 \leq \Delta$ erfüllt ist, liegt am Anfang der Iteration die Cholesky-Zerlegung von B vor. Im wesentlichen ohne Mehraufwand kann man daher einen Newton-Schritt, mit 0 startend, durchführen. Da 0 links von λ^* liegt und ψ nach Aufgabe 2 monoton fallend und konvex ist, ist auch die Iterierte eine untere Schranke für die gesuchte Nullstelle λ^* . Definiert man wieder $w(\lambda) := \|(B + \lambda I)^{-1}g\|_2$, so setzt man also

$$l := \frac{w(0)}{w'(0)} \left[1 - \frac{w(0)}{\Delta} \right].$$

Gestartet wird das Verfahren dann etwa mit $\lambda_0 := (l u)^{1/2}$. Im obigen Beispiel erhält man $l = 0.50$, $u = 2.83$, $\lambda_0 = 1.19$.

4. Sei $f(x) := \frac{1}{2} \|F(x)\|_2^2$ mit einer stetig differenzierbaren Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$, durch $f_x(p) := \frac{1}{2} \|F(x) + F'(x)p\|_2^2$ sei die zugehörige Modell-Funktion gegeben. Bei einer vorgegebenen Richtung p sei

$$r_x(p) := \frac{f(x) - f(x + p)}{f(x) - f_x(p)}$$

der Quotient aus tatsächlicher und vorhergesagter Verminderung der Zielfunktion f . Sei $p = p(\lambda)$ mit einem $\lambda \geq 0$ Lösung des linearen Ausgleichsproblems

$$\text{Minimiere } \frac{1}{2} \left\| \begin{pmatrix} F(x) \\ 0 \end{pmatrix} + \begin{pmatrix} F'(x) \\ \sqrt{\lambda} I \end{pmatrix} p \right\|_2^2, \quad p \in \mathbb{R}^n.$$

(a) Man zeige:

$$r_x(p) = \left[1 - \left(\frac{\|F(x + p)\|_2}{\|F(x)\|_2} \right)^2 \right] / \left[\left(\frac{\|F'(x)p\|_2}{\|F(x)\|_2} \right)^2 + 2 \left(\frac{\sqrt{\lambda} \|p\|_2}{\|F(x)\|_2} \right)^2 \right].$$

- (b) Sei $\phi(t) := \frac{1}{2} \|F(x + tp)\|_2^2$. Man bestimme das quadratische Polynom $\psi \in \Pi_2$, für welches $\psi(0) = \phi(0)$, $\psi'(0) = \phi'(0)$ und $\psi(1) = \phi(1)$ und anschließend das Minimum μ von $\psi(\cdot)$.

Hinweis: Man benutze die Normalgleichungen. Siehe auch J. J. MORÉ (1978, S. 108 ff.).

5. Gegeben sei die halbglatte, unrestringierte Optimierungsaufgabe

$$(P) \quad \text{Minimiere } f(x) := g \circ F(x), \quad x \in \mathbb{R}^n.$$

Hierbei sei $g: \mathbb{R}^m \rightarrow \mathbb{R}$ konvex und lipschitzstetig, die Abbildung $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ in $x \in \mathbb{R}^n$ stetig differenzierbar. Mit gegebenem $\Delta > 0$, einer symmetrischen Matrix $B \in \mathbb{R}^{n \times n}$ und einer Norm $\|\cdot\|$ auf dem \mathbb{R}^n sei p^* eine Lösung von

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := g(F(x) + F'(x)p) + \frac{1}{2} p^T B p, \quad \|p\| \leq \Delta.$$

Dann gilt die folgende Verallgemeinerung von Lemma 5.3:

- (a) Ist $f(x) = f_x(p^*)$, so ist x eine stationäre Lösung von (P).
- (b) Ist zusätzlich B positiv semidefinit und x eine stationäre Lösung von (P), so ist $f(x) = f_x(p^*)$.

Als Verallgemeinerung des ersten Teils von Lemma 5.4 zeige man:

- (c) Ist B positiv semidefinit und bezeichnet man den Optimalwert von $(P_{x,\Delta})$ mit $v(x, \Delta)$, so ist

$$f(x) - v(x, \Delta) \geq \frac{\Delta}{\Delta^*} [f(x) - v(x, \Delta^*)] \quad \text{für } 0 < \Delta \leq \Delta^*.$$

6. Man betrachte bei gegebenen $\Delta > 0$, $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ die restringierte, diskrete, lineare Tschebyscheffsche Approximationsaufgabe

$$(P) \quad \text{Minimiere } \|Ax - b\|_\infty, \quad \|x\|_\infty \leq \Delta.$$

Zu (P) stelle man ein äquivalentes lineares Programm auf, bilde hierzu das duale Programm und überführe dieses in Normalform. Anschließend erläutere man, wie man (P) mit Hilfe des (revidierten) Simplexverfahrens lösen kann.

Hinweis: Bezeichnet man mit e den Vektor im \mathbb{R}^m bzw. \mathbb{R}^n , dessen Komponenten sämtlich gleich 1 sind, so ist (P) äquivalent zum linearen Programm

$$\begin{aligned} & \text{Minimiere } \delta \text{ unter den Nebenbedingungen} \\ & \delta \geq 0, \quad -\delta e \leq Ax - b \leq \delta e, \quad -\Delta e \leq x \leq \Delta e. \end{aligned}$$

In Normalform geschrieben ist dieses lineare Programm durch die Daten

0^T	0^T	1	0^T	0^T	0^T	0^T	
A	$-A$	e	$-I$	0	0	0	b
$-A$	A	e	0	$-I$	0	0	$-b$
$-I$	I	0	0	0	$-I$	0	$-\Delta e$
I	$-I$	0	0	0	0	$-I$	$-\Delta e$

gegeben. Das hierzu duale Programm besitzt, wiederum in Normalform geschrieben, die Daten

$-b^T$	b^T	Δe^T	Δe^T	0	
A^T	$-A^T$	$-I$	I	0	0
e^T	e^T	0^T	0^T	1	1

Man erkennt, daß die Phase II des Simplexverfahrens sofort gestartet werden kann. Aus einem optimalen Tableau liest man eine Lösung des hierzu dualen linearen Programms ab und erhält damit, bis auf das Vorzeichen, eine Lösung des Ausgangsproblems.

7. Man programmiere das in Satz 5.5 angegebene Madsen-Verfahren zur numerischen Lösung der diskreten, nichtlinearen Tschebyscheffschen Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^n.$$

Eine Lösung der in jedem Iterationsschritt anfallenden linearen Aufgabe

$$(P_{x,\Delta}) \quad \text{Minimiere } f_x(p) := \|F(x) + F'(x)p\|_\infty, \quad \|p\|_\infty \leq \Delta$$

bestimme man mit der im Hinweis zu Aufgabe 6 angedeuteten Modifikation des Simplexverfahrens. Anschließend teste man das Programm an den folgenden Beispielen (siehe auch Aufgabe 15 in Abschnitt 7.2):

- (a) Man setze $t_i := (i - 11)/10$, $i = 1, \dots, 21$. Die Abbildung $F: \mathbb{R}^5 \rightarrow \mathbb{R}^{21}$ sei durch

$$F_i(x_1, x_2, x_3, x_4, x_5) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2 + x_5 t_i^3} - \exp(t_i), \quad i = 1, \dots, 21,$$

gegeben.

k	x_k					$\ p_k\ _\infty$	Δ_k
0	0.500000	0.000000	0.000000	0.000000	0.000000	0.125000	0.125000
1	0.625000	0.125000	-0.125000	-0.125000	-0.125000	0.125000	0.125000
2	0.750000	0.250000	-0.250000	-0.250000	-0.250000	0.125000	0.125000
3	0.875000	0.375000	-0.374079	-0.125000	-0.125000	0.125000	0.125000
4	1.000000	0.500000	-0.438373	0.000000	-0.037940	0.125000	0.125000
5	1.000315	0.425272	-0.563373	0.055038	0.033347	0.125000	0.125000
6	0.999897	0.312307	-0.683809	0.180038	-0.013358	0.063643	0.125000
7	0.999911	0.255105	-0.744992	0.243681	-0.037016	0.001614	0.063643
8	0.999878	0.253591	-0.746605	0.245199	-0.037490	0.000002	0.001614
9	0.999878	0.253588	-0.746608	0.245202	-0.037490	0.000000	0.000002

Tabelle 7.9: Ergebnisse zu Aufgabe 7 (a)

(b) Sei $F: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ durch

$$F(x_1, x_2) := \begin{pmatrix} x_1^2 + x_2^2 + x_1 x_2 \\ \sin x_1 \\ \cos x_2 \end{pmatrix}$$

gegeben.

Hinweis: Nimmt man in (a) den Startwert $x_0 := (0.5, 0.0, 0.0, 0.0, 0.0)^T$, setzt man $\Delta_0 := 0.125$ und benutzt man die von Madsen empfohlenen Werte

$$(\rho_1, \rho_2, \rho_3, \sigma_1, \sigma_2) := (0.01, 0.25, 0.25, 0.25, 2),$$

so erhält man die in Tabelle 7.9 angegebenen Werte.

In (b) haben wir (wie K. MADSEN (1975a)) den Startwert $x_0 := (3.0, 1.0)^T$ genommen und $\Delta_0 := 1.2$ gesetzt. Die übrigen Parameter sind wie in (a) gewählt. Man erhält die in Tabelle 7.10 angegebenen Werte. Man erkennt, daß hier, im Gegensatz zu (a), die Restriktion $\|p\|_\infty \leq \Delta_k$ im Hilfsproblem aktiv bleibt.

8. Gegeben sei die diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe

$$(P) \quad \text{Minimiere } f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^2,$$

wobei die Abbildung $F: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ durch

$$F(x_1, x_2) := \begin{pmatrix} x_1^2 + x_2^2 + x_1 x_2 \\ \sin x_1 \\ \cos x_2 \end{pmatrix}$$

gegeben ist (siehe auch Beispiel (b) in Aufgabe 7). Mit Hilfe der notwendigen Optimierungsbedingungen erster Ordnung (siehe Satz 1.9) berechne man (zwei) stationäre

k	x_k		$\ F(x_k)\ _\infty$	$\ p_k\ _\infty$	Δ_k
0	3.000000	1.000000	13.000000	1.097914	1.200000
1	1.902086	0.182688	3.998797	0.545962	1.097914
2	1.481128	-0.363273	1.787653	0.545962	0.545962
3	0.935166	-0.909235	0.850958	0.545962	0.545962
4	0.935166	-0.909235	0.850958	0.136490	0.136490
5	0.798676	-0.772745	0.716433	0.136490	0.136490
6	0.662185	-0.909235	0.663116	0.136490	0.136490
7	0.525695	-0.913294	0.630347	0.136490	0.136490
8	0.389205	-0.913124	0.629883	0.034123	0.034123
9	0.423327	-0.906837	0.617670	0.034123	0.034123
10	0.457450	-0.907123	0.617169	0.034123	0.034123
15	0.453184	-0.906595	0.616435	0.000533	0.000533
20	0.453284	-0.906592	0.616432	0.000033	0.000033

Tabelle 7.10: Ergebnisse zu Aufgabe 7 (b)

Lösungen von (P). Durch eine Anwendung der hinreichenden Bedingungen zweiter Ordnung (siehe Satz 1.12) zeige man, daß diese stationären Lösungen zumindestens isolierte lokale Lösungen von (P) sind. Ferner überzeuge man sich davon, daß in diesen Lösungen die Haarsche Bedingung (siehe Satz 2.14) *nicht* erfüllt ist.

Hinweis: Man weist leicht nach, daß man stationäre Lösungen (x_1^*, x_2^*) erhält, indem man x_1^* als Lösung von $3x^2 - \cos 2x = 0$ berechnet und $x_2^* := -2x_1^*$ setzt. Hiermit ist $(x_1^*, x_2^*) \approx (\pm 0.4532962370, \mp 0.9065924741)$.

9. Gegeben sei die nichtlineare Min-Max-Aufgabe

$$(P) \quad \text{Minimiere } f(x) := \max_{i=1,\dots,m} F_i(x), \quad x \in \mathbb{R}^n.$$

Die Abbildung $F = (F_i): \mathbb{R}^n \rightarrow \mathbb{R}^m$ sei in $x^* \in \mathbb{R}^n$ stetig differenzierbar. Ist x^* eine stationäre Lösung von (P), in der die Haarsche Bedingung (d. h. jede $n \times n$ -Untermatrix von $F'(x^*)$ ist nichtsingulär) erfüllt ist, so ist x^* eine lokal stark eindeutige Lösung von (P), d. h. es existieren positive Konstanten σ und δ mit

$$\sigma \|x - x^*\|_\infty + f(x^*) \leq f(x) \quad \text{für alle } x \in B[x^*; \delta] := \{x \in \mathbb{R}^n : \|x - x^*\|_\infty \leq \delta\}.$$

Hinweis: Man gehe ganz ähnlich vor wie beim Beweis von Satz 2.14 und berücksichtige hierbei, daß stationäre Lösungen von (P) in der Aufgabe 4 im Abschnitt 7.1 charakterisiert wurden.

10. Bei gegebenen Vektoren $a_i \in \mathbb{R}^n$, reellen Zahlen $b_i \in \mathbb{R}$, $i = 1, \dots, m$, und $\Delta > 0$ sei die lineare, restriktive Min-Max-Aufgabe

$$(P) \quad \text{Minimiere } \max_{i=1,\dots,m} (a_i^T x - b_i), \quad \|x\|_\infty \leq \Delta$$

gegeben. Man überlege sich, wie man (P) mit Hilfe des Simplexverfahrens lösen kann.

11. Gegeben sei die nichtlineare Min-Max-Aufgabe

$$(P) \quad \text{Minimiere } f(x) := \max_{i=1,\dots,m} F_i(x), \quad x \in \mathbb{R}^n.$$

Man formuliere und beweise hierzu Aussagen, die denen in Satz 5.5 und Satz 5.6 entsprechen.

Literaturverzeichnis

- [1] AL-BAALI, M. (1985) "Descent property and global convergence of the Fletcher–Reeves method with inexact line search." *IMA J. Numer. Anal.* 5, 121–124.
- [2] ANDERSON, D. H. AND M. R. OSBORNE (1977) "Discrete, nonlinear approximation problems in polyhedral norms." *Numer. Math.* 28, 143–156.
- [3] ANSTREICHER, K. M. (1986) "A monotonic projective algorithm for fractional linear programming." *Algorithmica* 1, 483–498.
- [4] ANSTREICHER, K. M. (1989) "A combined phase I-phase II projective algorithm for linear programming." *Mathematical Programming* 43, 209–223.
- [5] ATKINSON, K. E. (1978) *An Introduction to Numerical Analysis*. John Wiley & Sons, New York.
- [6] AXELSSON, O. (1985) "A survey of preconditioned iterative methods for linear systems of algebraic equations." *BIT* 25, 166–187.
- [7] BARNES, E. R. (1986) "A variation on Karmarkar's algorithm for solving linear programming problems." *Mathematical Programming* 36, 174–182.
- [8] BARTELS, R. H. (1971) "A stabilization of the simplex method." *Numer. Math.* 16, 414–434.
- [9] BARTELS, B. H. AND G. H. GOLUB (1969) "The simplex method for linear programming using LU decomposition." *Communications of the ACM* 12, 266–268.
- [10] BAUER, F. L. AND C. T. FIKE (1960) "Norms and exclusion theorems." *Numer. Math.* 2, 137–144.
- [11] BAUER, F. L., J. STOER AND C. WITZGALL (1961) "Absolute and monotonic norms." *Numer. Math.* 3, 257–264.
- [12] BEALE, E. M. L. (1955) "Cycling in the dual simplex algorithm." *Naval Research Logistics Quarterly* 2, 269–275.
- [13] BEALE, E. M. L. (1972) "A derivation of conjugate gradients." In: F. A. Lootsma, ed., *Numerical Methods for Non-linear Optimization*, 49–53, Academic Press, London-New York.
- [14] BLAND, R. G. (1977) "New finite pivoting rules for the simplex method." *Mathematics of Operations Research* 2, 103–107.

- [15] BORGWARDT, K. H. (1987) *The Simplex Method. A Probabilistic Analysis*. Springer-Verlag, Berlin-Heidelberg-New York.
- [16] BUNSE, W. UND A. BUNSE-GERSTNER (1985) *Numerische lineare Algebra*. Teubner, Stuttgart.
- [17] BYRD, R. H., J. NOCEDAL AND Y. YUAN (1987) "Global convergence of a class of quasi-Newton methods on convex problems." *SIAM J. Numer. Anal.* 24, 1171–1190.
- [18] BYRD, R. H. AND J. NOCEDAL (1989) "A tool for the analysis of quasi-Newton methods with application to unconstrained minimization." *SIAM J. Numer. Anal.* 26, 727–739.
- [19] CHVÁTAL, V. (1983) *Linear Programming*. W. H. Freeman and Company, New York.
- [20] COLLATZ, L. UND W. WETTERLING (1971) *Optimierungsaufgaben*. Springer-Verlag, Berlin-Heidelberg-New York.
- [21] CROMME, L. (1976) "Eine Klasse von Verfahren zur Ermittlung bester nichtlinearer Tschebyscheff-Approximationen." *Numer. Math.* 25, 447–459.
- [22] CROMME, L. (1978) "Strong uniqueness. A far-reaching criterion for the convergence analysis of iterative processes." *Numer. Math.* 29, 179–193.
- [23] DANTZIG, G. B. (1966) *Lineare Programmierung und Erweiterungen*. Springer-Verlag, Berlin-Heidelberg-New York.
- [24] DENNIS, J. E. AND J. J. MORÉ (1974) "A characterizaton of superlinear convergence and its application to quasi-Newton methods." *Math. Comp.* 28, 549–560.
- [25] DENNIS, J. E. AND J. J. MORÉ (1977) "Quasi-Newton methods, motivation and theory." *SIAM Review* 19, 46–89.
- [26] DENNIS, J. E. AND R. B. SCHNABEL (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs.
- [27] DENNIS, J. E. AND R. B. SCHNABEL (1989) "A view of unconstrained optimization." In: *Handbooks in Operations Research and Management Science. Volume 1 Optimization*. (G. L. Nemhauser, A. H. G. Rinnooy Kan and M. J. Todd, eds.), 1–72. North-Holland, Amsterdam-New York-Oxford-Tokyo.
- [28] DENNIS, J. E. AND T. STEIHAUG (1986) "On the successive projections approach to least-squares problems." *SIAM J. Numer. Anal.* 23, 717–733.
- [29] DIXON, L. C. W. (1972) "Variable metric algorithms: Necessary and sufficient conditions for identical behavior on nonquadratic functions." *J. Opt. Theory Appl.* 10, 34–40.
- [30] EL-ATTAR, R. A. ET AL. (1979) "An algorithm for L_1 -norm minimization with application to nonlinear L_1 -approximation. *SIAM J. Numer. Anal.* 16, 70–86.
- [31] FLETCHER, R. (1970) "A new approach to variable metric algorithms." *Computer Journal* 13, 317–322.

- [32] FLETCHER, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
- [33] FLETCHER, R. AND M. J. D. POWELL (1963) "A rapidly convergent descent method for minimization." *Computer Journal* 6, 163–168.
- [34] FLETCHER, T. AND C. M. REEVES (1964) "Function minimization by conjugate gradients." *Computer Journal* 7, 149–154.
- [35] FORSTER, O. (1983) *Analysis 1.* 4., durchgesehene Auflage. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.
- [36] FORSTER, O. (1984) *Analysis 2.* 5., durchgesehene Auflage. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.
- [37] FORSYTHE, G. E. AND C. B. MOLER (1967) *Computer Solution of Linear Algebraic Systems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.
- [38] FORSYTHE, G. E., M. A. MALCOLM AND C. B. MOLER (1977) *Computer Methods for Mathematical Computations.* Prentice-Hall, Inc., Englewood Cliffs, N. J.
- [39] FRANCIS, J. G. F. (1961/62) "The *QR* transformation: a unitary analogue to the *LR* transformation. Part I and II." *Comput. J.* 4, 265–271, 332–345.
- [40] FRANKLIN, J. (1980) *Mathematical Economics.* Springer-Verlag, New York-Heidelberg-Berlin.
- [41] FRANKLIN, J. (1987) "Convergence in Karmarkar's algorithm for linear programming." *SIAM J. Numer. Anal.* 24, 928–945.
- [42] GALE, D. (1960) *The Theory of Linear Economic Models.* MacGraw-Hill, New York.
- [43] GARCIA-PALOMARES, U. M. (1975) "Superlinearly convergent algorithms for linearly constrained optimization problems." In: *Nonlinear Programming 2* (O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds.), 101–119.
- [44] GEIGER, C. (1977) "Zur Konvergenz eines Abstiegsverfahrens für nichtlineare gleichmäßige Approximationsaufgaben." Preprint 77/14. Universität Hamburg, Institut für Angewandte Mathematik.
- [45] GILL, P. E., W. MURRAY AND M. H. WRIGHT (1981) *Practical Optimization.* Academic Press, London-New York-Toronto-Sydney-San Francisco.
- [46] GOLDFARB, D. (1970) "A family of variable metric methods derived by variational means." *Math. Comp.* 24, 23–26.
- [47] GOLDFARB, D. (1977) "On the Bartels-Golub decomposition for linear programming bases." *Mathematical Programming* 13, 272–279.
- [48] GOLDFARB, D. AND S. MEHROTRA (1988a) "Relaxed variants of Karmarkar's algorithm for linear programs with unknown optimal objective value." *Mathematical Programming* 40, 183–195.
- [49] GOLDFARB, D. AND S. MEHROTRA (1988b) "A relaxed version of Karmarkar's method." *Mathematical Programming* 40, 289–315.

- [50] GOLDFARB, D. AND S. MEHROTRA (1989) "A self-correcting version of Karmarkar's algorithm." *SIAM J. Numer. Anal.* 26, 1006–1015.
- [51] GOLDFARB, D. AND M. J. TODD (1989) "Linear programming." In: *Handbooks in Operations Research and Management Science. Volume 1 Optimization.* (G. L. Nemhauser, A. H. G. Rinnooy Kan and M. J. Todd, eds.), 73–170. North-Holland, Amsterdam-New York-Oxford-Tokyo.
- [52] GOLUB, G. H. AND W. KAHAN (1965) "Calculating the singular values and pseudoinverse of a matrix." *SIAM J. Num. Anal.* 2, 205–224.
- [53] GOLUB, G. H. AND C. F. VAN LOAN (1989) *Matrix Computations. Second Edition.* The Johns Hopkins University Press, Baltimore.
- [54] GONIN, R. AND A. H. MONEY (1989) *Nonlinear L_p -Norm Estimation.* Marcel Dekker, Inc., New York-Basel.
- [55] GOURLAY, A. R. AND G. A. WATSON (1973) *Computational Methods for Matrix Eigenproblems.* John Wiley & Sons, London-New York-Sydney-Toronto.
- [56] GREGORY, R. T. AND D. L. KARNEY (1969) *A collection of matrices for testing computational algorithms.* J. Wiley, New York-London-Sydney-Toronto.
- [57] HADLEY, G. (1962) *Linear Programming.* Addison-Wesley, Reading-Menlo Park-London-Sydney-Manila.
- [58] HÄMMERLIN, G. UND K.-H. HOFFMANN (1989) *Numerische Mathematik.* Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo.
- [59] HAN, S. P. (1981) "Variable metric methods for minimizing a class of nondifferentiable functions." *Mathematical Programming* 20, 1–13.
- [60] HESTENES, M. R. AND E. STIEFEL (1952) "Methods of conjugate gradients for solving linear systems." *J. Res. Nat. Bur. Standards* 48, 409–436.
- [61] KARMARKAR, N. (1984) "A new polynomial-time algorithm for linear programming." *Combinatorica* 4, 373–395.
- [62] KHACHIAN, L. G. (1979) "A polynomial algorithm in linear programming (in Russian)." *Doklady Akademii Nauk SSSR* 244, 1093–1096. English translation: *Soviet Mathematics Doklady* 20, 191–194.
- [63] KIELBASIŃSKI, A. UND H. SCHWETLICK (1988) *Numerische lineare Algebra.* Verlag Harri Deutsch, Thun-Frankfurt am Main.
- [64] KLEE, V. AND G. MINTY (1972) "How good is the simplex algorithm?" In: *Inequalities III* (O. Shisha, ed.), 159–175. Academic Press, New York.
- [65] KÖCKLER, N. (1990) *Numerische Algorithmen in Softwaresystemen.* B. G. Teubner, Stuttgart.
- [66] KOSMOL, P. (1989) *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben.* B. G. Teubner, Stuttgart.

- [67] LANCASTER, P. AND M. TISMONETSKY (1985) *The Theory of Matrices. Second Edition with Applications*. Academic Press, New York.
- [68] LAWLER, E. L. (1980) "The great mathematical sputnik of 1979." *The Mathematical Intelligencer* 2, 191–198.
- [69] LAWSON, C. L. AND R. J. HANSON (1974) *Solving Least Squares Problems*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- [70] LUENBERGER, D. G. (1973) *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Menlo Park, California.
- [71] MADSEN, K. (1975a) "An algorithm for minimax solution of overdetermined systems of non-linear equations." *J. Inst. Maths Applics* 16, 321–328.
- [72] MADSEN, K. (1975b) "Minimax solution of non-linear equations without calculating derivatives." *Mathematical Programming Study* 3, 110–126.
- [73] MAESS, G. (1984) *Vorlesungen über numerische Mathematik I. Lineare Algebra*. Birkhäuser Verlag, Basel-Boston-Stuttgart.
- [74] MAESS, G. (1988) *Vorlesungen über numerische Mathematik II. Analysis*. Birkhäuser Verlag, Basel-Boston-Stuttgart.
- [75] MCCORMICK, G. P. (1983) *Nonlinear Programming. Theory, Algorithms, and Applications*. John Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore.
- [76] MCCORMICK, G. P. AND K. RITTER (1974) "Alternative proofs of the convergence properties of the conjugate-gradient method." *J. Optim. Theor. Appl.* 13, 497–518.
- [77] MEINARDUS, G. UND G. MERZ (1979) *Praktische Mathematik I*. Bibliographisches Institut, Mannheim-Wien-Zürich.
- [78] MEINARDUS, G. UND G. MERZ (1981) *Praktische Mathematik II*. Bibliographisches Institut, Mannheim-Wien-Zürich.
- [79] MORÉ, J. J. (1978) "The Levenberg-Marquardt algorithm: implementation and theory." In: *Lecture Notes in Mathematics* 630, G. A. Watson, ed., Springer-Verlag, Berlin-Heidelberg-New York, 105–116.
- [80] MORÉ, J. J. (1983) "Recent developments in algorithms and software for trust region methods." In: *Mathematical Programming. The State of the Art Bonn 1982*. A. Bachem, M. Grötschel, B. Korte, eds., Springer-Verlag, Berlin-Heidelberg-New York-Tokyo.
- [81] MURTY, K. G. (1983) *Linear Programming*. J. Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore.
- [82] NASH, J. C. (1979) *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Adam Hilger Ltd, Bristol.
- [83] OREN, S. S. AND D. G. LUENBERGER (1974) "Self-scaling variable metric (SSVM) algorithms. Part I: Criteria and sufficient conditions for scaling a class of algorithms." *Management Science* 20, 845–862.

- [84] OREN, S. S. (1974) "Self-scaling variable metric (SSVM) algorithms. Part: Implementation and experiments." *Management science* 20, 863–874.
- [85] OREN, S. S. AND E. SPEDICATO (1976) "Optimal conditioning of selfscaling variable metric algorithms." *Mathematical Programming* 10, 70–90.
- [86] ORTEGA, J. M. AND W. C. RHEINBOLDT (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.
- [87] ORTEGA, J. M. (1972) *Numerical Analysis. A Second Course*. Academic Press, New York-London.
- [88] OSBORNE, M. R. (1972) "An algorithm for discrete, nonlinear best approximation problems." *ISNM* 16, 117–126.
- [89] OSBORNE, M. R. (1985) *Finite Algorithms in Optimization and Data Analysis*. J. Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
- [90] OSBORNE, M. R. AND G. A. WATSON (1967) "On the best linear Chebyshev approximation." *The Computer Journal* 10, 172–177.
- [91] OSBORNE, M. R. AND G. A. WATSON (1978) "Nonlinear approximation problems in vector norms." In: *Lecture Notes in Mathematics* 630, G. A. Watson, ed., Springer-Verlag, Berlin-Heidelberg-New York, 117–132.
- [92] PAPADIMITRIOU, C. H. AND K. STEIGLITZ (1982) *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- [93] PARLETT, B. N. (1980) *The Symmetric Eigenvalue Problem*. Prentice-Hall, Inc., Englewood Cliffs, N. J.
- [94] PARLETT, B. N. AND W. G. POOLE (1973) "A geometric theory for the QR , LU and power iterations." *SIAM J. Numer. Anal.* 10, 389–412.
- [95] POLAK, E. AND G. RIBIÈRE (1969) "Note sur la convergence de méthodes de directions conjuguées." *Rev. Fr. Inf. Rech. Oper.* 16, 35–43.
- [96] POWELL, M. J. D. (1971) "On the convergence of the variable metric algorithm." *J. Inst. Math. Appl.* 7, 21–36.
- [97] POWELL, M. J. D. (1975) "Convergence properties of a class of minimization algorithms." In: O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., *Nonlinear Programming 2*. Academic Press, New York-San Francisco-London.
- [98] POWELL, M. J. D. (1976) "Some global convergence properties of a variable metric algorithm for minimization without exact line searches." *SIAM-AMS Proceedings 9: Nonlinear Programming*, 53–72.
- [99] POWELL, M. J. D. (1977) "Restart procedures for the conjugate gradient method." *Mathematical Programming* 12, 241–254.
- [100] REMMERT, R. (1984) *Funktionentheorie I*. Springer-Verlag, Berlin-Heidelberg-New York-Tokyo.

- [101] RUTISHAUSER, H. (1966) "The Jacobi method for real symmetric matrices." *Numer. Math.* 9, 1–10.
- [102] SCHRIJVER, A. (1986) *Theory of Linear and Integer Programming*. John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
- [103] SCHWARZ, H. R. (1988) *Numerische Mathematik*. 2., durchgesehene Auflage. B. G. Teubner, Stuttgart.
- [104] SCHWETLICK, H. (1979) *Numerische Lösung nichtlinearer Gleichungen*. VEB Deutscher Verlag der Wissenschaften, Berlin.
- [105] SHANNO, D. F. (1970) "Conditioning of quasi-Newton methods for function minimization." *Math. Comp.* 24, 647–656.
- [106] SHANNO, D. F. (1978a) "On the convergence of a new conjugate gradient algorithm." *SIAM J. Numer. Anal.* 15, 1247–1257.
- [107] SHANNO, D. F. (1978b) "Conjugate gradient methods with inexact searches." *Mathematics of Operations Research* 3, 244–256.
- [108] SHANNO, D. F. (1980) "On variable-metric methods for sparse hessians." *Math. Comp.* 34, 499–514.
- [109] SHANNO, D. F. (1988) "Computing Karmarkar projections quickly." *Mathematical Programming* 41, 61–71.
- [110] SHULTZ, G. A., R. B. SCHNABEL AND R. H. BYRD (1985) "A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties." *SIAM J. Numer. Anal.* 22, 47–67.
- [111] SMITH, B. T. ET AL. (1974) *Matrix Eigensystem Routines: EISPACK Guide*. Lecture Notes in Computer Science Vol. 6. Springer-Verlag, Berlin-Heidelberg-New York.
- [112] SORENSEN, D. C. (1982) "Newton's method with a model trust region modification." *SIAM J. Numer. Anal.* 19, 409–426.
- [113] STEWART, G. W. (1973) *Introduction to Matrix Computations*. Academic Press, Inc., New York.
- [114] STEWART, G. W. AND JI-GUANG SUN (1990) *Matrix Perturbation Theory*. Academic Press, Inc., Boston-San Diego-New York-London-Sydney-Tokyo-Toronto.
- [115] STOER, J. (1989) *Numerische Mathematik 1*. Fünfte, verbesserte Auflage. Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo-Hong Kong.
- [116] STOER, J. UND R. BULIRSCH (1990) *Numerische Mathematik 2*. Dritte, verbesserte Auflage. Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo-Hong Kong.
- [117] STRANG, G. (1988) *Linear Algebra and its Applications. Third Edition*. Harcourt Brace Jovanovich, Publishers, San Diego.
- [118] TODD, M. J. AND B. P. BURRELL (1986) "An extension of Karmarkar's algorithm for linear programming using dual variables." *Algorithmica* 1, 409–424.

- [119] TOINT, Ph. L. (1977) "On sparse and symmetric matrix updating subject to a linear equation." *Math. Comp.* 31, 954–961.
- [120] WALSH, G. R. (1985) *An Introduction to Linear Programming. Second Edition*. John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
- [121] WALTER, W. (1990) *Analysis I*. 2. Aufl. Springer-Verlag, Berlin-Heidelberg-New York-Tokyo.
- [122] WARTH, W. UND J. WERNER (1977) "Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben." *Computing* 19, 59–72.
- [123] WATKINS, D. S. (1982) "Understanding the QR algorithm." *SIAM Review* 24, 427–440.
- [124] WATSON, G. A. (1980) *Approximation Theory and Numerical Methods*. John Wiley & Sons, Chichester-New York-Brisbane-Toronto.
- [125] WERNER, J. (1978) "Über die globale Konvergenz von Variable-Metrik-Verfahren bei nicht-exakter Schrittweitenbestimmung." *Numer. Math.* 31, 321–334.
- [126] WERNER, J. (1984) *Optimization. Theory and Applications*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.
- [127] WILKINSON, J. H. (1965) *The Algebraic Eigenvalue Problem*. Oxford University Press, London-New York.
- [128] WILKINSON, J. H. (1968) "Global convergence of tridiagonal QR algorithm with origin shifts." *Linear Algebra and Its Applications* 1, 409–420.
- [129] WILKINSON, J. H. AND C. REINSCH (ed.) (1971) *Handbook for Automatic Computation. Vol. II, Linear Algebra*. Springer-Verlag, Berlin-Heidelberg-New York.
- [130] WOLFE, P. (1969) "Convergence conditions for ascent methods." *SIAM Review* 11, 226–235.
- [131] YUAN, Y. (1985) "Conditions for convergence of trust region algorithms for nonsmooth optimization." *Mathematical Programming* 31, 220–228.
- [132] ZIMMERMANN, U. (1988) "On recent developments in linear programming." In: *Trends in Mathematical Optimization* (K.-H. Hoffmann, J.-B. Hiriart-Urruty, C. Lemaréchal, J. Zowe, eds.), 353–390. Birkhäuser Verlag, Basel-Boston.

Index

- absolute Norm 7
- Abstand linearer Teilräume 35
- Abstiegsrichtung 144, 163, 174
 - gradientenähnliche 170
- algebraische Vielfachheit 3
- Approximationsaufgabe
 - diskrete, nichtlineare 143, 173 ff., 249
- Armijo-Schrittweite 166 ff., 175, 187 ff., 193, 195, 204, 207
- Ausgleichsproblem
 - lineares 66, 73, 78, 134, 174, 232
 - mit Rangdefizit 73
 - restringiertes 79, 80
 - nichtlineares 143, 146, 182, 189, 248 ff.
- Basisindizes 92
 - künstliche 102
- Basislösung 92
 - entartete 92, 99
 - nichtentartete 92, 98
 - zulässige 92
- Bauer-Fike, Satz von 6 ff., 15
- BFGS-Update-Formel 196, 212, 215
- BFGS-Verfahren 195, 199, 212, 218, 226
 - gedächtnisloses 226 ff.
 - globale Konvergenz 203 ff.
 - Implementation 201
 - lokale Konvergenz 206
 - superlineare Konvergenz 206 ff.
 - ungedämpftes 212
- Bidiagonalisierungsverfahren 67
- Bidiagonalmatrix, obere 67, 77
- Bisektions-Verfahren 56 ff., 60, 74
- Brouwerscher Fixpunktsatz 17
- Broyden-Klasse 195 ff., 211
 - eingeschränkte 206
- Carathéodory, Satz von 183
- charakteristisches Polynom 2, 56
- Cholesky-Zerlegung 76 ff., 197, 201, 247, 259
- Courantsches Minimum-Maximum Prinzip 10, 16
- Dennis-Moré, Satz von 210
- DFP-Update-Formel 196, 212
- diagonalähnlich, siehe diagonalisierbar
- diagonal dominante Matrix 15
- diagonalisierbare Matrix 8, 29, 36, 38
- Diätproblem 82, 84
- dominanter Eigenwert 29, 49, 50
- duales Programm 110 ff., 139
- duales Simplexverfahren 119 ff., 127
- Dualitätssatz
 - schwacher 111
 - starker 115
- dual zulässige Lösung 110
- Ecke
 - einer konvexen Menge 88
 - eines Polyeders 106
- effiziente Schrittweiten 168
- Eigenwertaufgaben 1 ff.
 - bei symmetrischen Matrizen 52 ff.
- exakte Schrittweite 164, 195, 198, 213, 224
- Farkas-Lemma 116, 125, 151, 160, 258
- Fletcher-Reeves-Verfahren 224 ff., 233
 - mit inexakter Schrittweitenstrategie 224, 234
- freie Variable 83, 84
- Frobenius-Norm 5, 12, 54, 212
 - gewichtete 212
- Funktionalmatrix 146
- Gateaux-Variation 146, 159
 - der Betragssummennorm 158
 - der Maximumnorm 149
 - Kettenregel 150, 159

- konvexer Funktionen 148
- Gauß-Jordan-Matrix 97
- Gauß-Matrix 47
- Gauß-Newton-Verfahren
 - gedämpftes 173, 176 ff., 182, 189 ff.
 - lokaler Konvergenzsatz 187
 - ungedämpftes 179
- Gershgorin-Kreise 3 ff.
- Gershgorin, Satz von 3 ff., 15, 61
- Givens-Rotation 25 ff., 52, 61 ff., 70 ff., 202, 249
- Gleichgewichtsbedingungen 118
- gleichmäßig konvexe Funktion 155 ff., 171 ff., 193
 - Charakterisierung 155 ff.
- goldener Schnitt 185
- Goldstein-Schrittweite 185
- Gradient 146
- gradientenähnlich 170
- Gradientenverfahren 163, 170, 186
- Haarsche Bedingung 183, 263
- halbglatte Optimierungsaufgabe 150, 257, 260
- Halbraum 88, 116
- Hessenberg-Matrix 19, 202
 - obere 19
 - symmetrische 19
 - unreduzierte 19, 40, 43, 49
- Hessesche 152
- hinreichende Optimalitätsbedingung zweiter Ordnung 153, 207
 - bei diskreter L_1 -Approximation 161
 - bei diskreter Tschebyscheff-Approximation 153
 - bei Min-Max-Aufgaben 160
- Householder-Matrix 19 ff., 22 ff., 44 ff., 67
- Hyperebene 88, 116
- $\inf(P)$ 86, 111
- inverse Iteration nach Wielandt 30 ff., 50, 75
- Jacobi-Rotation 52
- Jacobi-Verfahren 12, 52 ff.
 - klassisches 52 ff.
 - Konvergenz 55
 - zyklisches 53, 56, 73
- Kantorowitsch, Ungleichung von 186, 215, 221, 230
- Karmarkar-Normalform 129 ff.
 - Zurückführung auf 138 ff.
- Karmarkar-Verfahren 128 ff.
 - Konvergenz 135 ff.
 - Motivation 129
- Kettenregel 150, 159
- Kondition 8, 170, 186, 215
- konjugierte Richtungen 219, 231
- konvexe Funktion 148
 - Existenz der Gateaux-Variation 148
 - glatte 155 ff.
 - Charakterisierung 155 ff.
 - gleichmäßig 155 ff., 171
 - Charakterisierung 155 ff.
 - konvexe Menge 87
 - Konvexitätskombination 87
 - Kostenfunktion 81
 - Kuhn-Tucker, Satz von 112, 117
 - künstliche Variable 101 ff.
 - L_1 -Approximationsaufgabe, diskrete 149
 - Levenberg-Marquardt-Trajektorie 246, 248
 - Levenberg-Marquardt-Verfahren 249
 - lexikographische Ordnung 99
 - lineare Optimierungsaufgabe 81 ff.
 - allgemeine Form 83
 - duale 110 ff.
 - Dualitätssatz
 - schwacher 111
 - starker 115
 - Existenzsatz 92
 - Normalform 84 ff., 138
 - lineares Ausgleichsproblem
 - 66, 73, 78, 134, 174, 232
 - lineares Programm 81 ff.
 - logarithmische Potentialfunktion 137
 - lokal starke Eindeutigkeit 178 ff., 254 ff.
 - Lösung einer unrestringierten Optimierungsaufgabe
 - globale 143
 - isolierte lokale 153
 - lokale 143
 - stationäre 144
 - LR -Zerlegung 31, 39, 49, 98
 - einer Tridiagonalmatrix 75

- Madsen-Verfahren 252, 261
 Konvergenzgeschwindigkeit 254
- Matrix
 diagonal dominante 15
 diagonalisierbare 8, 29, 36
 nichtnegative 17
 normale 17
 unzerlegbare 17
- Matrixspiele, Hauptsatz 125
- $\max(D)$ 111
- Methode der kleinsten Quadrate 146
- Methode vom goldenen Schnitt 185
- $\min(P)$ 86, 111
- Min-Max-Optimierungsaufgabe 149, 263, 264
 lineare, restringierte 263
- Modellalgorithmus 162
 Konvergenz des 169 ff.
- Modellfunktion 236, 237, 238, 248, 250, 257
- monotone Norm 7
- Moore-Penrose-Gleichungen 72
- Newton-Richtung 166, 187, 194
- Newton-Verfahren 31, 58, 182, 192 ff., 247, 259
 gedämpftes 187, 193, 217
 globaler Konvergenzsatz 193
 Trust-Region-Modifikation 241 ff.
 ungedämpftes 192, 245
 lokaler Konvergenzsatz 192
- nichtlineares Ausgleichsproblem 143, 146, 248 ff.
- Norm
 absolute 7
 monotone 7
 polyhedrale 175
- normale Matrix 17
- Normalform einer linearen Optimierungsaufgabe 84 ff., 138
- notwendige Optimalitätsbedingung
 erster Ordnung 145 ff., 159
 zweiter Ordnung 153
- ökonomische Interpretation der Dualität 118 ff.
- Optimierungsaufgabe
 duale lineare 110 ff.
 geometrische 162
- lineare 81 ff.
 lösbare 86
 unrestringierte 143 ff.
 Wert einer 86
 zulässige 86
- Oren-Luenberger-Klasse 214
 optimal konditionierte 214
- Perron-Frobenius, Satz von 17
- Pivotelement 107
- Polak-Ribiére-Verfahren 225, 227, 233
- Polyeder 88
- Polytop 88, 106
- Potentialfunktion, logarithmische 137
- Powell-Schrittweite 165, 195, 204, 214
- Präkonditionierung 222 ff.
- Produktionsplanungsproblem 82, 84, 118
- Programm, lineares 81 ff.
- projektive Transformation 130
- Pseudoinverse 37, 72, 73, 248
- QR -Doppelschritt 43 ff., 51, 62
- QR -Verfahren 12, 18 ff.
 einfaches 32 ff.
 Konvergenz 38, 64
 für symmetrische Matrizen 61
 Konvergenz 64 ff.
 mit Shifts 40 ff.
- QR -Zerlegung 18 ff., 32, 174, 182, 189, 202, 249
 einer Hessenberg-Matrix 24 ff.
 Existenz, Eindeutigkeit 32
 nach Givens 24
 nach Householder 19
- Quasi-Newton-Gleichung 196, 229
- Quasi-Newton-Verfahren 195 ff.
 BFGS-Verfahren 196, 199
 Broyden-Klasse 196 ff.
 DFP-Verfahren 196, 213
- Rayleigh-Quotient 10, 75
- Rayleigh-Quotienten Iterationsverfahren 76
- Rayleighsches Variationsprinzip 9
- Richtungsableitung 146
- richtungsdifferenzierbar 146
- Richtungsstrategie 145, 162
- Rosenbrock-Funktion 147, 158, 170, 217
 erweiterte 215, 217

- Rouché, Satz von 2
- Satz von
- Bauer-Fike 6 ff., 15
 - Carathéodory 183
 - Dennis-Moré 210
 - Gershgorin 3 ff., 15, 61
 - Kuhn-Tucker 112, 117
 - Perron-Frobenius 17
 - Rouché 2
 - Schur 12 ff.
- Schattenpreise 119
- Schlupfvariable 84
- Schrittweite
- Armijo- 166 ff., 175, 187 ff., 193, 195, 204, 207, 227
 - effiziente 168
 - exakte 164, 195, 198, 204, 213, 224, 227
 - Goldstein- 185
 - Powell- 165, 195, 204, 214, 227, 230
 - semi-effiziente 168
- Schrittweitenstrategie 145, 162, 163 ff.
- Schur-Zerlegung
- komplexe 12
 - reelle 13
- schwacher Dualitätssatz 111
- Schwellenmethode 56, 73
- Shanno-Verfahren 226 ff.
- Sherman-Morrison-Formel 96, 122, 197
- Shift-Parameter 18, 27 ff., 41 ff.
- Simplexverfahren
- duales 119 ff.
 - geometrische Grundlagen 87 ff.
 - geometrische Idee 93
 - Phase I 101 ff.
 - Phase II 93 ff.
 - revidiertes 96
 - vollständige Tableaus 98, 106 ff.
 - Zusatzregel 100
 - Zyklus beim 98
- singuläre Werte 65
- Singulärwertzerlegung 65 ff., 78
- Spektralverschiebung 30
- starker Dualitätssatz 115
- stationäre Lösung 144, 147
- bei nichtlinearer Approximation 174
 - bei diskreter Tschebyscheff-Approximation 151
- Stetigkeit der Eigenwerte 3
- Störung der Eigenwerte 8, 15
- $\sup(D)$ 111
- Trust-Region-Verfahren 236 ff.
- bei diskreten, nichtlinearen Approximationsaufgaben 249 ff.
 - bei glatten, unrestringierten Optimierungsaufgaben 238
 - bei nichtlinearen Ausgleichsproblemen 248
- Motivation 236
- Tschebyscheffsche Approximationsaufgabe
- diskrete, lineare 83, 86, 113 ff., 124
 - restringierte 260 ff.
 - diskrete, nichtlineare 149, 151, 182, 190, 250, 261, 262
- Ungleichung vom geometrisch-arithmetischen Mittel 135
- Ungleichung von Kantorowitsch 186, 215, 221, 230
- unimodale Funktion 185
- unrestringierte Optimierungsaufgabe 143 ff.
- globale Lösung 143
 - lokale Lösung 143
 - Modellalgorithmus 162 ff.
 - stationäre Lösung 144, 147
- Update-Formel 195
- BFGS 196
 - Broyden 196
 - DFP 196
 - Oren-Luenberger 214
- Variable
- freie 83
 - künstliche 101 ff.
 - vorzeichenbeschränkte 83
- Variationsprinzipien
- Courant 10
 - Rayleigh 9
- Vektoriteration nach v. Mises 29 ff., 34, 49, 50
- Verfahren der konjugierten Gradienten 134, 218 ff.
- bei linearem Ausgleichsproblem 232

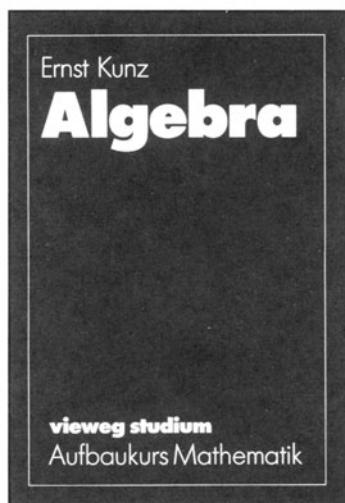
- bei quadratischer Zielfunktion 219
- mit Präkonditionierung 223
- modifiziertes 232
- Verfahren des steilsten Abstiegs 163
- Verfahren von Golub-Reinsch 66, 77
- Vertauschungsmatrix 47
- vorzeichenbeschränkte Variable 83
- Wilkinson-Shift 64, 70
- Wert einer Optimierungsaufgabe 86, 111
- Zielfunktion 81, 143
- zulässige Lösung 81
- zulässige Optimierungsaufgabe 86
- Zusatzregel beim Simplexverfahren
 - Bland-Regel 99
 - lexikographische Ordnung 100
- Zyklus beim Simplexverfahren 99
 - Beispiel von Beale 109

Algebra

von Ernst Kunz

1991. X, 254 S. (vieweg studium, Bd. 43, Aufbaukurs Mathematik;
hrsg. von Gerd Fischer) Paperback.

ISBN 3-528-07243-1



Das Problem, Gleichungen zu lösen, hat die Entwicklung der Algebra über mehr als zwei Jahrtausende begleitet. Geometrische Aufgaben lassen sich in die Algebra übersetzen und in deren präziser Sprache behandeln. Es ist das Leitmotiv des Buches, die Theorie anhand leicht verständlicher Probleme zu entwickeln und durch ihre Lösung zu motivieren. Dabei lernt man kennen, was zu einer Einführung in die Algebra im Grundstudium gehört: Die Körper mit ihren Erweiterungen bis hin zur Galoistheorie, ferner die elementaren Techniken der Gruppen- und

Ringtheorie. Der Text enthält 350 Übungsaufgaben von verschiedenen Schwierigkeitsgraden einschließlich Hinweisen zu ihrer Lösung. Das Buch gründet sich auf die Erfahrungen des Autors mit mehreren Generationen von Studenten und ist besonders zu empfehlen für Lehrer und solche, die es werden wollen.

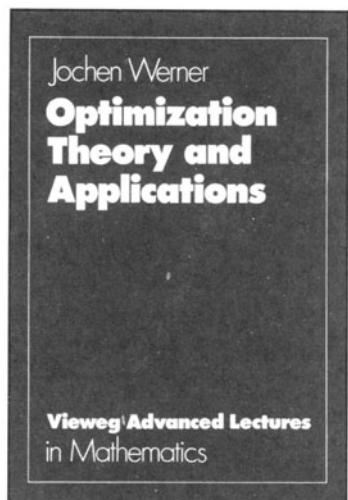
Verlag Vieweg · Postfach 58 29 · D-6200 Wiesbaden



Optimization Theory and Applications

von Jochen Werner

1984. VIII, 233 pp. (*Advanced Lectures in Mathematics; ed. by Gerd Fischer*) Softcover.
ISBN 3-528-08594-0



This book is intended to give the reader an introduction to the foundation and an impression of the applications of optimization theory. It particularly emphasizes the duality theory of convex programming and necessary optimality conditions for nonlinear optimization problems.

Abstract theorems are made more concrete by numerous examples from e. g. approximation theory, calculus of variations and optimal control theory.

Verlag Vieweg · Postfach 58 29 · D-6200 Wiesbaden

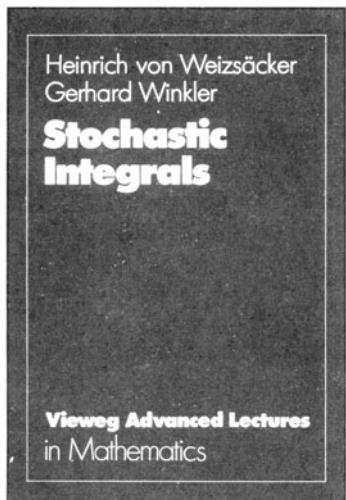


Stochastic Integrals

von Heinrich von Weizsäcker and Gerhard Winkler

An Introduction.

1990. X, 332 S. (*Advanced Lectures in Mathematics*,
ed. by Gerd Fischer and Manfred Knebusch) Softcover.
ISBN 3-528-06310-6



This text introduces in a moderate speed the basic concepts of the theory of stochastic integrals for semimartingales. Having introduced martingales and local martingales the stochastic integral is defined for locally uniform limits of elementary processes. This corresponds to the Riemann integral in one dimensional analysis and it suffices for the study of stochastic differential equations and diffusions including the Feynman-Kac formula and the Stroock-Varadhan martingale problem approach.

Predictability is introduced mainly as a tool for the structure theory of semimartingales which culminates in the Dellacherie-Bichteler characterization theorem. Besides these abstracts parts the material of the text is designed for a one semester course.

Verlag Vieweg · Postfach 58 29 · D-6200 Wiesbaden

