Studienbücherei

H. Kaiser

Numerische Mathematik und Rechentechnik II

VEB Deutscher Verlag der Wissenschaften

Mathematik für Lehrer
Band 10

Herausgegeben von:
W. Engel, S. Brehmer, M. Schneider, H. Wussing

Unter Mitarbeit von:
G. Asser, J. Böhm, J. Flachsmeyer, G. Geise, T. Glocke,
K. Härtig, G. Kasdorf, O. Krötenheerdt, H. Lugowski,
P. H. Müller, G. Porath

Studienbücherei

Numerische Mathematik und Rechentechnik II

H. Kaiser

Mit 25 Abbildungen
und 19 Tabellen

VEB Deutscher Verlag
der Wissenschaften
Berlin 1980

Verlagslektor: Dipl.-Math. B. Mai
Verlagshersteller: B. Burkhardt
Umschlaggestaltung: R. Wendt

Printed in the German Democratic Republic
Lizenz-Nr. 206 · 435/86/80
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", 74 Altenburg
LSV 1084
Bestellnummer: 570 861 9
DDR 16,80 M

Vorwort

Aufbauend auf den im Grundkurs Mathematik erworbenen Kenntnissen werden die im Lehrprogramm für die Disziplin Numerische Mathematik ausgewiesenen Gebiete

Approximationstheorie (Kapitel 5),
Lösung von Gleichungen (Kapitel 6) und
Lineare Optimierung (Kapitel 7)

behandelt. Während Kapitel 7 im Umfang den verfügbaren Vorlesungsstunden entspricht, enthalten die beiden anderen Kapitel ergänzenden Stoff und weiterführende Betrachtungen. Diese Teile des Buches können den Anschluß für einen mit dem Numerikstudium gekoppelten Kurs der wahlobligatorischen Ausbildung vermitteln und eignen sich auch für die Gestaltung des mathematischen Fachpraktikums. Letzteres gilt zum Beispiel für die Abschnitte 5.2.3. bis 5.2.5., 5.3.2. und die Betrachtungen zur Lokalisierung von Polynomnullstellen in 6.2.2. Der Abschnitt 5.4. zu Theorie und Anwendung von Splinefunktionen geht thematisch über das Lehrprogramm hinaus. Splinefunktionen haben aber in den letzten Jahren eine solche Bedeutung für die Numerische Mathematik gewonnen, daß eine Einführung in deren Theorie in dem hier dargebotenen Umfang vertretbar erscheint.

Kapitel 8 enthält Ausführungen zum linguistischen Aspekt der Informationsverarbeitung. Sie ergänzen die allgemeinen Betrachtungen der Kapitel 1 und 2 aus MfL Bd. 9 und präzisieren die Vorstellungen zur Syntax formaler Sprachen, die in Kapitel 3 entwickelt wurden. Der Verfasser hofft, daß dieser Text trotz seines fragmentarischen Charakters zum tieferen Verständnis der mit den Begriffen Signal, Information, Zeichen zu beschreibenden informationellen Prozesse beiträgt, deren methodische Gestaltung auch für die Schule wachsende Bedeutung erhält, z. B. im Biologieunterricht der Abiturstufe. Vor allem soll Kapitel 8 aber eine Verbindung zum Lehrgebiet Grundlagen der Mathematik herstellen.

Die letzten Bemerkungen weisen darauf hin, daß die nunmehr vorliegenden Teile I und II einer Einführung in die Numerische Mathematik und Rechentechnik als eine Einheit zu betrachten sind. Dessen ungeachtet war der erste Band mit Rücksicht auf die Studierenden des Nebenfachs als selbständiges, den einschlägigen Stoff des

mathematischen Grundkurses umfassendes Lehrbuch zu konzipieren. Diese Nebenbedingung mußte sich notwendigerweise auf die Systematik der Darstellung auswirken, die aber so dem tatsächlichen Vorgehen im Studienbetrieb entspricht. Der Einheitlichkeit wegen wurden die in Teil I benutzten PAP-Symbole der TGL 22451/1967 beibehalten.

Das zu beiden Bänden in Zusammenarbeit mit Herrn Dr. A. FÜHRICH entwickelte Übungsmaterial wird in MfL Bd. 15 erscheinen und liegt bereits als Sonderdruck vor. Es enthält Anwendungsaufgaben, deren mathematische Modellierung mit den in der EOS vermittelten naturwissenschaftlichen Kenntnissen möglich ist. Die Gliederung der Übungen entspricht der von MfL Bd. 9 und 10. Jedem Abschnitt sind einige durchgerechnete Beispielaufgaben vorangestellt, die zum Teil den Charakter einer komplexen Belegübung haben und als Muster für Praktikumsaufgaben dienen können.

Bei der Arbeit am Manuskript bin ich wieder durch mehrere Hinweise und Ratschläge der Herren des Herausgeberkollegiums, namentlich von Professor Dr. ENGEL und Professor Dr. SCHNEIDER, unterstützt worden, denen ich dafür meinen herzlichen Dank ausspreche. Dieser gilt in besonderer Weise meiner Frau, die das Manuskript geschrieben und alle Organisationsarbeiten im Rechenzentrum der Akademie übernommen hat, sowie meinem Sohn HANS-CHRISTOPH, der die Programmierung und Durchführung numerischer Experimente besorgte. Sehr zu danken habe ich auch Frau Dipl.-Math. B. MAI vom DVW für Hinweise auf Errata und die angenehme Zusammenarbeit während der Drucklegung sowie allen, die an der Herstellung des schwierigen Satzes beteiligt waren.

Berlin, Sommer 1979 H. KAISER

Inhalt

5. Approximation von Funktionen

5.1. Begriffsbildungen und allgemeine Sachverhalte

5.1.1. Formulierung des Approximationsproblems

Gegeben sei eine reelle Funktion f einer reellen Veränderlichen x,

$$f: X \to \mathbf{R}, \tag{1}$$

wobei $X \subseteq \mathbf{R}$ im folgenden meist ein Intervall ist. Daneben wird eine von n Parametern

$$a_1, a_2, \ldots, a_n, \qquad \boldsymbol{a} := (a_1, a_2, \ldots, a_n)^{\mathsf{T}}, \qquad \boldsymbol{a} \in A \subseteq \mathbf{R}^n \tag{2}$$

abhängige Schar solcher Funktionen

$$F_{\boldsymbol{a}}: X \to \mathbf{R} \tag{3}$$

betrachtet, deren Werte wir an Stelle von $F_{\boldsymbol{a}}(x)$ auch mit $F(\boldsymbol{a}, x)$ oder $F(a_1, a_2, \ldots, a_n, x)$ bezeichnen. Wir stellen uns die Aufgabe, f durch eine Funktion der Schar (3) über X oder einer Teilmenge $M \subseteq X$ in einem gewissen Sinn möglichst gut anzunähern. Nach Einschränkung aller betrachteten Funktionen auf M kann stets $X = M$ angenommen werden.

Solche approximativen Darstellungen von Funktionen ergeben sich bei der *konstruktiven* Lösung von Aufgaben der Analysis. Im besonderen ist darauf hinzuweisen, daß in einem Computer Funktionswerte im allgemeinen aktuell erzeugt und nicht — dem Arbeiten mit einem Tafelwerk vergleichbar — aus einem Speicher entnommen werden. Da der Aufruf von Funktionsprozeduren in einem Programm häufig erfolgen kann, kommt es im Hinblick auf die Rechenzeit sehr darauf an, einfache Verfahren zu entwickeln, die Näherungswerte hoher Genauigkeit liefern. Das geschieht meist durch Konstruktion einer rationalen Näherungsfunktion, die dann der Berechnung zugrunde liegt.

Neben der Auswahl der Schar (3) für die Approximation zugelassener Funktionen ist ein Maß für deren Güte festzulegen. Dies geschieht mit Hilfe einer *Abstands-* oder *Distanzfunktion* ϱ, die dem Funktionspaar f, $F_{\boldsymbol{a}}$ eine nichtnegative Zahl zuordnet, wobei kleine Werte von ϱ „gute" Näherungen charakterisieren. Man beachte, daß dann $\varrho(f, F_{\boldsymbol{a}})$ eine auf A definierte von $\boldsymbol{a}$ abhängende Funktion ist, die wir weiterhin

mit Z bezeichnen:

$$Z(\boldsymbol{a}) := \varrho(f, F_{\boldsymbol{a}}). \tag{4}$$

Meist hat ϱ die Eigenschaft einer Norm $\|\cdot\|$ oder wird mit Hilfe von Normen gebildet. Das setzt voraus, daß die Funktionen (1) und (3) Elemente eines reellen linearen Vektorraumes E[1]) im Sinne von MfL Bd. 3, Kap. 9., sind. Ein Beispiel dafür ist der in MfL Bd. 4, 2.6.3., behandelte Raum C_D der auf einer abgeschlossenen beschränkten Menge D stetigen Funktionen mit der durch die *Tschebyscheff-Metrik* gegebenen Abstandsbestimmung. Wählen wir D als das abgeschlossene Intervall $[\![a, b]\!]$, so gilt hier für $f, F_{\boldsymbol{a}} \in C_{[a,b]}$

$$Z(\boldsymbol{a}) = \|f - F_{\boldsymbol{a}}\| := \max_{x \in [a,b]} |f(x) - F(\boldsymbol{a}, x)|. \tag{5}$$

Im folgenden werden wir die Funktionen (1) und (3) gewöhnlich als Elemente von $C_{[a,b]}$ annehmen, an Stelle von (5) aber auch andere Normen betrachten. Wir erinnern daran, daß diese durch folgende Eigenschaften charakterisiert sind:

1.	$\|f\| \geqq 0$	für alle $f \in E$ und $\|f\| = 0$ nur für das Nullelement von E;	
2.	$\|\alpha f\| = \|\alpha\| \cdot \|f\|$	für alle $\alpha \in \mathbf{R}$ und $f \in E$;	(6)
3.	$\|f + g\| \leqq \|f\| + \|g\|$	für alle $f, g \in E$ (Dreiecksungleichung).	

Das uns interessierende *Approximationsproblem* kann endgültig so formuliert werden:

> *Für eine gegebene Funktion* (1) *wird bezüglich einer ausgezeichneten Funktionenschar* (3) *und einer Distanzfunktion* ϱ *ein solcher Parametervektor* $\boldsymbol{a}^* \in A \subseteq \mathbf{R}^n$ *gesucht, daß*
>
> $$Z(\boldsymbol{a}) \geqq Z(\boldsymbol{a}^*) \tag{7}$$
>
> *für alle* $\boldsymbol{a} \in A$ *gilt.*

Die Aufgabe besteht also darin, für die (Ziel-) Funktion (4) ein absolutes Minimum auf A zu bestimmen. Ist $\boldsymbol{a}^*$ ein solches, so heißt $F_{\boldsymbol{a}^*}$ *Bestapproximation* von f oder wegen (7) *Minimallösung*. Die Theorie befaßt sich vor allem mit zwei Fragenkomplexen:

1. Entscheidung, ob ein (7) genügender Parametervektor $\boldsymbol{a}^*$ existiert, in Verbindung mit dessen Charakterisierung durch Kriterien (gegebenenfalls interessiert die Einzigkeit der Minimallösung).
2. Entwicklung von Algorithmen zur effektiven Bestimmung von $\boldsymbol{a}^*$.

Wir betrachten ein *Beispiel:* Es seien $X \subseteq \mathbf{R}$ und f gegeben. $F_{\boldsymbol{a}}$ bedeutet ein beliebiges Polynom höchstens $(m-1)$-ten Grades ($m \geqq 1$, ganz), wobei der Parametervektor $\boldsymbol{a}$ durch das System der Koeffizienten bestimmt wird:

$$F_{\boldsymbol{a}}(x) = \sum_{\mu=0}^{m-1} a_\mu x^\mu. \tag{8}$$

[1]) Im folgenden wird man erkennen, daß die Natur der Elemente von E im allgemeinen keine Rolle spielt.

In X wählen wir m paarweise verschiedene Punkte x_i und bilden damit den Vektor

$$\boldsymbol{d} = (d_1, d_2, \ldots, d_m)^{\mathrm{T}}, \qquad d_i := |f(x_i) - F_a(x_i)|. \tag{9}$$

Eine Distanzfunktion wird etwa mit Hilfe der euklidischen Norm $\|\cdot\|_2$ (MfL Bd. 4, 1.5.) von $\boldsymbol{d}$ definiert:

$$Z(\boldsymbol{a}) = \varrho(f, F_a) = \|\boldsymbol{d}\|_2 = \left(\sum_{i=1}^{m} |d_i|^2\right)^{1/2}. \tag{10}$$

Wegen $Z(\boldsymbol{a}) \geqq 0$ wird Bestapproximation erreicht, wenn $Z(\boldsymbol{a}) = 0$, also $d_i = 0$ für $i = 1, 2, \ldots, m$ ist, d. h., wenn F_a die mit den Wertepaaren

$$\big(x_i, f(x_i)\big), \qquad i = 1, 2, \ldots, m,$$

formulierte Interpolationsaufgabe löst. Auf Grund des Satzes 1 in MfL Bd. 9, 4.2.1., gibt es genau ein Polynom mit dieser Eigenschaft in der Schar (8); es ist u. a. in Form des Lagrangeschen Interpolationspolynoms bestimmbar.

An Stelle von $\|\cdot\|_2$ kann man für beliebiges $p \geqq 1$ Distanzfunktionen mit der Vektornorm

$$\|\boldsymbol{d}\|_p := \left(\sum_{i=1}^{m} |d_i|^p\right)^{1/p} \tag{11}$$

bilden:

$$\varrho(f, F_a) := \|\boldsymbol{d}\|_p = \left(\sum_{i=1}^{m} |f(x_i) - F(a, x_i)|^p\right)^{1/p}. \tag{12}$$

Für $\|\cdot\|_p$ sind die ersten beiden Normeigenschaften (6) offensichtlich. Der Beweis der Dreiecksungleichung kann im Falle $p = 1$ dem Leser überlassen bleiben. Wenn $p > 1$ ist, benötigt man dazu eine Verallgemeinerung der Schwarzschen Ungleichung. Ohne darauf näher einzugehen, sei nur erwähnt, daß man im Falle $p = 2$ mit dieser selbst auskommt. Die Durchführung der Betrachtungen findet man in MfL Bd. 4, 1.5.2.

Wir wollen noch $\|\boldsymbol{d}\|_p$ für $p \to \infty$ untersuchen. Dabei wird diese Größe bei festem $\boldsymbol{d} \neq 0$ als Funktion von p betrachtet. Es sei

$$h(p) := \ln \|\boldsymbol{d}\|_p = \frac{1}{p} \ln \left(\sum_{i=1}^{m} |d_i|^p\right)$$

und

$$\mu := \max_{i=1,2,\ldots,m} |d_i|.$$

Die Indexmenge $\{1, 2, \ldots, m\}$ denke man sich in zwei disjunkte Teilmengen $\{i_1, i_2, \ldots, i_k\}$, $\{i'_1, i'_2, \ldots, i'_l\}$ $(k + l = m)$ derart zerlegt, daß

$$|d_{i_j}| = \mu \quad \text{für} \quad j = 1, 2, \ldots, k$$

und

$$|d_{i'_j}| < \mu \quad \text{für} \quad j = 1, 2, \ldots, l$$

gilt. Dann ist — wenn $|d_{i'_j}| = \alpha_{i'_j}\mu$, $0 \leqq \alpha_{i'_j} < 1$ —

$$h(p) = \frac{1}{p} \ln \left(k + \sum_{j=1}^{l} \alpha_{i'_j}^p \right) + \ln \mu .$$

Wegen $\lim\limits_{p\to\infty} \left(k + \sum\limits_{j=1}^{l} \alpha_{i'_j}^p \right) = k$ folgt $\lim\limits_{p\to\infty} h(p) = \ln \mu$ und

$$\lim_{p\to\infty} \|\boldsymbol{d}\|_p = \mu. \tag{13}$$

Auf Grund von (13) führen wir die Größe

$$\|\boldsymbol{d}\|_\infty := \max_{i=1,2,\ldots,m} |d_i| \tag{14}$$

ein, die als Grenzwert der p-Norm selbst die Normeigenschaften (6) besitzt. Letzteres kann man auch leicht direkt zeigen.

Die Funktion

$$\varrho(f, F_{\boldsymbol{a}}) := \|\boldsymbol{d}\|_\infty = \max_{i=1,2,\ldots,m} |d_i|, \tag{15}$$

wobei $\boldsymbol{d}$ den Vektor (9) bedeutet, ist eine weitere häufig benutzte Distanzfunktion.

Die Norm (14) ist das diskrete Analogon zu der in (5) auftretenden Norm, die der Tschebyscheff-Metrik zugrunde liegt. Für $1 \leqq p < \infty$ kann gezeigt werden, daß die $\|\cdot\|_p$ entsprechende Größe

$$\|f\|_p := \left(\int_a^b |f(x)|^p \, dx \right)^{1/p} \tag{16}$$

auf $C_{(a,b)}$ die Eigenschaften (6) besitzt. Schwierigkeiten bereitet dabei nur der Beweis der Dreiecksungleichung. Dazu benötigt man die im Zusammenhang mit $\|\cdot\|_p$ erwähnten Ungleichungen für Integrale. Mit der Norm (16) wird die Distanzfunktion

$$\varrho(f, F_{\boldsymbol{a}}) = \|f - F_{\boldsymbol{a}}\| = \left(\int_a^b |f(x) - F(\boldsymbol{a}, x)|^p \, dx \right)^{1/p} \tag{17}$$

gebildet.

In der Praxis werden über X an die Güte der Approximation gelegentlich unterschiedliche Anforderungen gestellt. Das legt nahe, die in den Distanzfunktionen auftretenden Abweichungen $f(x) - F(\boldsymbol{a}, x)$ mit Gewichten zu versehen und etwa an Stelle von (12) und (17)

$$\varrho(f, F_{\boldsymbol{a}}) = \left(\sum_{i=1}^{m} w_i \, |f(x_i) - F(\boldsymbol{a}, x_i)|^p \right)^{1/p}, \tag{18}$$

$$w_i > 0, \qquad i = 1, 2, \ldots, m,$$

bzw.

$$\varrho(f, F_{\boldsymbol{a}}) = \left(\int_a^b w(x) \, |f(x) - F(\boldsymbol{a}, x)|^p \, dx \right)^{1/p}, \tag{19}$$

$$w(x) \geqq 0 \quad \text{für} \quad a \leqq x \leqq b,$$

zu betrachten.

Es trifft den Sachverhalt gut, wenn man die bei der Formulierung eines Approximationsproblems vorzunehmende Bestimmung der Funktionenschar (3) und einer Distanzfunktion, J. R. Rice [43] folgend, als die Wahl von „Form" und „Norm" bezeichnet. Erstere ist durch den Anwendungsbezug häufig vorgegeben, während hinsichtlich der Distanzfunktion eine gewisse Willkür herrscht. Dabei ist zu beachten, daß sich für dieselbe Funktionenschar (3), aber verschiedene Distanzfunktionen, im allgemeinen voneinander abweichende Bestapproximationen ergeben. Das sei an einem einfachen Beispiel erläutert. Eine auf $[\![a, b]\!]$ definierte Funktion f mit positiver zweiter Ableitung soll durch ein lineares Polynom

$$P(x) = a_0 + a_1 x \tag{20}$$

approximiert werden. Wählt man (10) als Distanzfunktion mit $m = 2$ und $x_1 = a$, $x_2 = b$, so ist das durch die Verbindungsgerade der Punkte $(a, f(a))$ und $(b, f(b))$ repräsentierte Polynom (20) beste Approximation. Bezüglich der mit der Tschebyscheff-Norm gebildeten Abstandsfunktion (5) gewinnt man dafür ein Polynom, dessen Graph die nach der Parallelenkonstruktion der Abb. 5.1 bestimmte Gerade g ist; g schneidet die Sehne AT im Mittelpunkt M. Wir werden diesen Sachverhalt in 5.3. mit Hilfe des Tschebyscheffschen Alternantensatzes begründen.

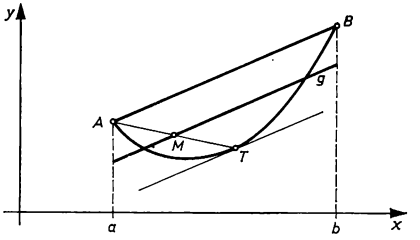


Abb. 5.1

5.1.2. Lineare Approximationsprobleme

Die in 5.1.1. erörterte Polynomapproximation weist die Besonderheit auf, daß die als Koeffizienten von (8) bestimmten Komponenten des Parametervektors in $F(a, x)$ linear auftreten. Allgemein bezeichnet man ein Approximationsproblem als *linear*, wenn n Funktionen

$$\varphi_i \colon X \to \mathbf{R}, \qquad i = 1, 2, \ldots, n, \tag{21}$$

derart existieren, daß die Funktionen (3) in der Form

$$F_a = \sum_{i=1}^{n} a_i \varphi_i \tag{22}$$

darstellbar sind. Dem Beispiel vergleichbar lassen auch andere lineare Probleme eine in gewissem Sinne „explizite" Bestimmung der (oder einer) Bestapproximation zu.

Offenbar ist es vernünftig, die φ_i auf X von vornherein als *linear unabhängig* anzunehmen (vgl. MfL Bd. 4, 1.3.3.), da sonst jede der Funktionen (21) durch eine Linearkombination linear unabhängiger φ_i ausgedrückt, d. h. eine Reduktion von (22), also der Zahl der zu bestimmenden Parameter, vorgenommen werden könnte. Dieser Gesichtspunkt ist in dem betrachteten Beispiel berücksichtigt, da gemäß MfL Bd. 4, 1.3.4., Satz 2, das System der Potenzfunktionen i-ten Grades ($i = 0, 1, \ldots m-1$) linear unabhängig ist.

Endgültig wird ein lineares Approximationsproblem so charakterisiert:

> *E sei ein linearer Raum von Funktionen*[1]) *$f: X \to \mathbb{R}$ (z. B. $E = C_{(a,b)}$) und (φ_i) ($i = 1, 2, \ldots, n$) ein System linear unabhängiger Elemente aus E. Für eine gegebene Funktion $f \in E$ wird ein Vektor $\boldsymbol{a}^* \in \mathbb{R}^n$ derart gesucht, daß bezüglich einer Abstandsfunktion* (4)
>
> $$F_{\boldsymbol{a}^*} = \sum_{i=1}^{n} a_i^* \varphi_i \tag{23}$$
>
> *Bestapproximation von f ist im Vergleich mit allen anderen Linearkombinationen*
>
> $$F_{\boldsymbol{a}} = \sum_{i=1}^{n} a_i \varphi_i .$$

Für lineare Probleme gilt:

Satz 1. *Die mit einer Norm $\|\cdot\|$ über E gebildete Distanzfunktion Z $\left(Z(\boldsymbol{a}) := \|f - F_{\boldsymbol{a}}\|\right)$ ist für beliebiges $f \in E$ stetig.*

Beweis. Wir haben die Werte von Z an benachbarten Stellen $\boldsymbol{a}$ und $\boldsymbol{a} + \boldsymbol{h}$ zu betrachten, wobei $\boldsymbol{h} = (h_1, h_2, \ldots, h_n)^{\mathsf{T}}$ sei. Dann ist

$$|Z(\boldsymbol{a} + \boldsymbol{h}) - Z(\boldsymbol{a})| = \left| \|f - F_{\boldsymbol{a}+\boldsymbol{h}}\| - \|f - F_{\boldsymbol{a}}\| \right| \leqq \|F_{\boldsymbol{a}+\boldsymbol{h}} - F_{\boldsymbol{a}}\| = \left\| \sum_{i=1}^{n} h_i \varphi_i \right\|$$

$$\leqq \sum_{i=1}^{n} |h_i| \cdot \|\varphi_i\| .$$

Diese Abschätzung läßt erkennen, daß $|Z(\boldsymbol{a} + \boldsymbol{h}) - Z(\boldsymbol{a})|$ mit $\boldsymbol{h} \to \boldsymbol{0}$ gegen Null strebt.

Bemerkung. Zum Beweis der Ungleichung

$$| \|f\| - \|g\| | \leqq \|f - g\| \quad \text{für} \quad f, g \in E \tag{24}$$

benötigt man nur die zweite und dritte der Normeigenschaften (6). Eine Abbildung $h: E \to \mathbb{R}_+$ für welche

$$h(\alpha f) = |\alpha| \, h(f) \tag{25}$$

und

$$h(f + g) \leqq h(f) + h(g), \qquad f, g \in E, \qquad \alpha \in \mathbb{R},$$

[1]) Vgl. dazu die Fußnote auf S. 10.

gilt, heißt eine *Halbnorm* über E. Betrachten wir als Beispiel

$$E = C_{[a,b]}.$$

x_j, $j = 1, 2, \ldots, m$, seien paarweise verschiedene Punkte des Intervalls $[\![a, b]\!]$, und es sei

$$f = (f_1, f_2, \ldots, f_m)^T, \qquad f_j := |f(x_j)|.$$

Dann ist die mit einer beliebigen Vektornorm des $\mathbf{R}^m$ definierte Abbildung

$$h(f) = \|f\|$$

Halbnorm über $C_{[a,b]}$. Offenbar gilt Satz 1 auch noch, wenn man $|\cdot|$ durch eine Halbnorm ersetzt.

Im folgenden wird die Existenz einer Bestapproximation nachgewiesen. Dem Leser sei die Analyse der Beweise von Satz 2 und des vorangestellten Hilfssatzes empfohlen, da in diesen für die Approximationstheorie typische Schlußweisen benutzt werden, die auf dem Satz von Bolzano-Weierstrass, d. h. Kompaktheitseigenschaften des $\mathbf{R}^n$ (vgl. MfL Bd. 4, 2.1.6. und 2.4.2.) beruhen.

Hilfssatz 1. *$|\cdot|$ sei eine Norm über E und $\|\cdot\|$ eine Norm über $\mathbf{R}^n$, z. B. die euklidische. Dann gilt*

$$\bigvee_{\mu>0} \mu \bigwedge_{a\in\mathbf{R}^n} a \quad (\|a\| = 1 \Rightarrow |F_a| \geqq \mu).$$

Beweis. Im Sinne eines indirekten Beweises nehmen wir an, daß die Negation der Aussage wahr ist:

$$\bigwedge_{\mu>0} \mu \bigvee_{a\in\mathbf{R}^n} a \quad (\|a\| = 1 \wedge |F_a| < \mu).$$

Für $\mu_k = \dfrac{1}{k}$, $k \in \mathbf{N}^*$, sei $a^{(k)}$ ein erfüllendes Element, d. h., es gilt $\|a^{(k)}\| = 1$ und $|F_a(k)| < \dfrac{1}{k}$. Die beschränkte Folge der $a^{(k)}$ enthält nach dem Satz von Bolzano-Weierstrass eine konvergente Teilfolge $(a^{(k_j)})$, deren Limes a' sei. Nach Satz 1 (für f wähle man das Nullelement von E) ist dann

$$\lim_{j\to\infty} |F_{a^{(k_j)}}| = |F_{a'}|$$

und wegen der Stetigkeit der Norm

$$\|a'\| = \lim_{j\to\infty} \|a^{(k_j)}\| = 1. \tag{26}$$

Andererseits gilt aber auf Grund der Bestimmung der $a^{(k)}$

$$\lim_{j\to\infty} |F_{a^{(k_j)}}| = 0.$$

Wir haben also

$$|F_{a'}| = 0$$

und auf Grund der ersten Normeigenschaft (6)

$$F_{a'} = \sum_{i=1}^{n} a_i' \varphi_i = 0.$$

Wegen der linearen Unabhängigkeit der φ_i folgt daraus $a' = 0$, im Widerspruch zu (26).

Satz 2. *$\|\cdot\|$ sei Norm über E, und Z sei die durch $Z(a) = \|f - F_a\|$ definierte Distanzfunktion. Dann besitzt das diesbezüglich formulierte lineare Approximationsproblem (23) eine Lösung.*

Beweis. Da $\|f - F_a\| \geqq 0$ ist, existiert

$$I := \inf_{a \in \mathbb{R}^n} \|f - F_a\| = \inf_{a \in \mathbb{R}^n} Z(a),$$

und es gibt eine Folge $(a^{(k)})$, $a^{(k)} \in \mathbb{R}^n$, mit der Eigenschaft

$$\lim_{k \to \infty} Z(a^{(k)}) = I. \tag{27}$$

Für $k > K$ sei $Z(a^{(k)}) = \|f - F_{a^{(k)}}\| < I + 1$, d. h., wenn $M := \|f\|$,

$$\|F_{a^{(k)}}\| < I + M + 1. \tag{28}$$

Die Menge der $a^{(k)}$ ist beschränkt. Zum Beweis betrachten wir eine Norm $\|\cdot\|$ des $\mathbb{R}^n$ und ein Element a dieses Raumes, für welches $\|a\| > \dfrac{I + M + 1}{\mu}$ ist, wobei μ eine die Aussage des Hilfssatzes 1 verifizierende positive Zahl bedeutet. Dann gilt mit $\tilde{a} := \dfrac{a}{\|a\|}$ auf Grund dieses Hilfssatzes

$$\|F_a\| = \left\| \|a\| \frac{1}{\|a\|} F_a \right\| = \|a\| \cdot \|F_{\tilde{a}}\| \geqq \|a\| \mu > I + M + 1.$$

Wegen (28) ist also für $k > K$

$$\|a^{(k)}\| \leqq \frac{I + M + 1}{\mu},$$

d. h., die Menge der $a^{(k)}$ ist beschränkt.

Nach dem Satz von Bolzano-Weierstrass enthält $(a^{(k)})$ eine konvergente Teilfolge $(a^{(k_j)})$, deren Limes a^* sei. Für diese gilt auf Grund von Satz 1

$$\lim_{j \to \infty} Z(a^{(k_j)}) = Z(a^*),$$

und wegen (27) ist $\lim\limits_{j \to \infty} Z(a^{(k_j)}) = I$, also

$$Z(a^*) = I.$$

Das heißt, das Infimum der Werte der Distanzfunktion wird als Funktionswert bei a^* angenommen. a^* ist absolutes Minimum von Z und F_{a^*} folglich Minimallösung.

Wir beschließen diesen Abschnitt mit *Unitätsbetrachtungen*, welche die eindeutige Bestimmtheit der durch Satz 2 gesicherten Bestapproximation zum Gegenstand haben.

Mit der in E gegebenen Norm $\|\cdot\|$ definieren wir für $r > 0$ und $f \in E$ den Begriff der (abgeschlossenen) r-Umgebung von f als die Menge der Elemente $g \in E$, für die

$$g \in \overline{U}_r(f) \Leftrightarrow \|g - f\| \leqq r \tag{29}$$

gilt. Maßgebend für die Einzigkeit der Lösung von (26) sind gewisse Eigenschaften der *Konvexität* von $\overline{U}_r(f)$. Die Definition der Konvexität einer Menge $M \subseteq \mathbb{R}^p$ in MfL Bd. 4, 1.5.2., läßt sich zwanglos auf beliebige normierte Räume E übertragen und in einer für unser Problem wesentlichen Hinsicht verschärfen:

$$M\ (\subseteq E)\ \textit{konvex} :\Leftrightarrow \bigwedge_{x,y \in M} \bigwedge_{\lambda,\mu \in \mathbb{R}_+} \{\lambda + \mu = 1 \Rightarrow \lambda x + \mu y \in M\}; \tag{30}$$

$$M\ (\subseteq E)\ \textit{streng konvex} :\Leftrightarrow \bigwedge_{\substack{x,y \in M \\ x \neq y}} \bigwedge_{\lambda,\mu \in \mathbb{R}_+^0} \{\lambda + \mu = 1 \Rightarrow \lambda x + \mu y\ \textit{innerer Punkt von}\ M\}. \tag{31}$$

Ein innerer Punkt von M ist dadurch charakterisiert, daß noch eine gewisse r-Umgebung desselben zu M gehört.

Hilfssatz 2. *In einem normierten Raum E ist jede Umgebung $\overline{U}_r(f)$ $(r > 0, f \in E)$ konvex.*

Beweis. Für $x, y \in \overline{U}_r(f)$, $\lambda, \mu \geqq 0$ und $\lambda + \mu = 1$ ist zu zeigen, daß $\lambda x + \mu y$ zu $\overline{U}_r(f)$ gehört:

$$\|f - (\lambda x + \mu y)\| = \|(\lambda + \mu) f - (\lambda x + \mu y)\| \leqq \lambda \|f - x\| + \mu \|f - y\| \leqq \lambda r + \mu r = r.$$

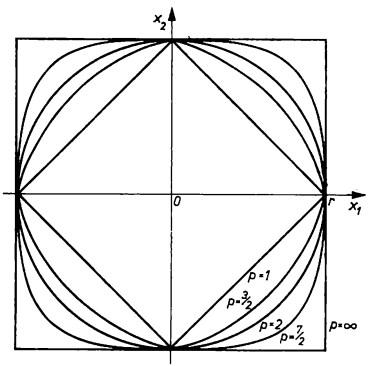


Abb. 5.2. $\overline{U}_r(0)$ für $p = 1, 2, \infty$ und $1 < p < 2$, $2 < p < \infty$

Zur Veranschaulichung betrachten wir Nullpunktumgebungen, die mit verschiedenen p-Normen des $\mathbb{R}^2$ (vgl. (12), (14)) gebildet sind (vgl. Abb. 5.2):

$$x = (x_1, x_2)^T \in \mathbb{R}^2, \qquad \|x\|_p = \sqrt[p]{|x_1|^p + |x_2|^p}, \qquad \|x\|_\infty = \max(|x_1|, |x_2|).$$

Man vermutet zu Recht, daß die $\overline{U}_r(f)$-Umgebungen des betrachteten Beispiels für $1 \leqq p \leqq \infty$ mit Ausnahme von $p = 1$ und $p = \infty$ streng konvex sind.

Definition 1. Ein normierter Raum E heißt *streng normiert*, wenn jede Umgebung $\overline{U}_r(f)$, $(r > 0, f \in E)$ im Sinne von (31) streng konvex ist.

Satz 3. *E sei ein bezüglich $\|\cdot\|$ streng normierter Raum, und es sei $Z(a) = \|f - F_a\|$. Dann hat das Approximationsproblem (23) genau eine Lösung.*

Beweis. Im Sinne eines indirekten Beweises sei angenommen, daß F_{a^*} und $F_{a^{**}}$ $(a^*, a^{**} \in \mathbf{R}^n$, $a^* \neq a^{**})$ Bestapproximationen von f sind. Wir setzen

$$r := \|f - F_{a^*}\| = \|f - F_{a^{**}}\|.$$

Dann gilt

$$r \leqq \left\|f - F_{\frac{a^* + a^{**}}{2}}\right\| \leqq \frac{1}{2}\|f - F_{a^*}\| + \frac{1}{2}\|f - F_{a^{**}}\| = r, \tag{32}$$

d. h., in dieser Abschätzung muß überall das Gleichheitszeichen stehen. Für $r > 0$ ist $F_{\frac{a^* + a^{**}}{2}}$ wegen F_{a^*}, $F_{a^{**}} \in \overline{U}_r(f)$ und $\frac{1}{2}F_{a^*} + \frac{1}{2}F_{a^{**}} = F_{\frac{a^* + a^{**}}{2}}$ auf Grund der strengen Normiertheit von E innerer Punkt von $\overline{U}_r(f)$, also

$$\left\|f - F_{\frac{a^* + a^{**}}{2}}\right\| < r;$$

dem widerspricht (32). Ist aber $r = 0$, so gilt $f = F_{a^*} = \sum_{i=1}^{n} a_i^* \varphi_i = F_{a^{**}} = \sum_{i=1}^{n} a_i^{**} \varphi_i$, also

$$\sum_{i=1}^{n} (a_i^* - a_i^{**})\, \varphi_i = 0.$$

Wegen der linearen Unabhängigkeit der φ_i folgt daraus

$$a_i^* = a_i^{**} \quad \text{für} \quad i = 1, 2, \ldots, n,$$

im Widerspruch zu $a^* \neq a^{**}$.

Ohne Beweis sei erwähnt:

Satz 4. *Der Funktionenraum $C_{\langle a,b\rangle}$ ist bezüglich der Normen $\|\cdot\|_p$, $1 < p < \infty$, streng normiert; für $\|\cdot\|_1$ und $\|\cdot\|_\infty$ gilt das nicht.*

In diesem Abschnitt wurden gelegentlich verschiedene Normen zugleich betrachtet; sie waren daher auch in der Bezeichnung zu unterscheiden. Weiterhin werden wir — wenn nur eine Norm im Spiele ist — dafür wie üblich $\|\cdot\|$ schreiben.

5.2. Quadratmittelapproximation

5.2.1. Problemformulierung

Die in 5.1.1. betrachteten Normen

$$\|\boldsymbol{d}\|_2 := \left(\sum_{i=1}^{m} d_i^2\right)^{1/2}, \qquad \boldsymbol{d} \in \mathbf{R}^m,$$

und

$$\|f\|_2 := \left(\int_a^b |f(x)|^2 \, dx \right)^{1/2}, \qquad f \in C_{\langle a,b \rangle},$$

haben die besondere Eigenschaft, daß ihre Quadrate als ein Skalarprodukt der Elemente $\boldsymbol{d}$ bzw. f mit sich selbst dargestellt werden können. Wie üblich identifizieren wir dabei den Begriff des Skalarproduktes mit dem der positiv definiten symmetrischen Bilinearform im Sinne von MfL Bd. 3, 7.1., beziehen uns weiterhin also auf folgende

Definition 1. E sei ein reeller linearer Raum gemäß MfL Bd. 3, Kap. 9. Eine Abbildung

$$(\cdot, \cdot): E \times E \to \mathbf{R}$$

heißt *Skalarprodukt* auf E, wenn für beliebige Elemente f, g, h aus E folgendes gilt:

1. $(f, f) > 0$, sofern f vom Nullelement des Raumes verschieden ist;
2. $(f, g) = (g, f)$;
3. $(f + g, h) = (f, h) + (g, h)$;
4. $(\alpha f, g) = \alpha (f, g)$ für jedes $\alpha \in \mathbf{R}$.

Definition 2. Ein linearer Raum E, auf dem ein Skalarprodukt erklärt ist, heißt *unitär* (bezüglich dieses Skalarproduktes).

Für die zu Anfang betrachtete Norm gilt

$$\|\boldsymbol{d}\|_2^2 = \langle \boldsymbol{d}, \boldsymbol{d} \rangle,$$

wenn $\langle \cdot, \cdot \rangle$ das durch

$$\langle \boldsymbol{c}, \boldsymbol{d} \rangle := \sum_{i=1}^{m} c_i d_i, \qquad \boldsymbol{c} = (c_1, c_2, \ldots, c_m)^{\mathsf{T}}, \qquad \boldsymbol{d} = (d_1, d_2, \ldots, d_m)^{\mathsf{T}} \tag{1}$$

definierte Skalarprodukt bedeutet (vgl. MfL Bd. 3, 7.2.).

Für den Raum $E = C_{\langle a,b \rangle}$ ist die Abbildung

$$\begin{aligned} &\langle \cdot, \cdot \rangle: C_{\langle a,b \rangle} \times C_{\langle a,b \rangle} \to \mathbf{R} \\ &\langle f, g \rangle = \int_a^b f(x)\, g(x)\, dx, \qquad f, g \in C_{\langle a,b \rangle}, \end{aligned} \tag{2}$$

ein Skalarprodukt. Die Eigenschaften 2 bis 4 der Definition 1 folgen unmittelbar aus den Rechenregeln für reelle Zahlen in Verbindung mit dem Integralbegriff. Ferner ist

$$\langle f, f \rangle = \int_a^b f^2 \, dx \geqq 0.$$

Das Gleichheitszeichen kann nur gelten, wenn $f(x) \equiv 0$ auf $[\![a, b]\!]$ ist. Wäre nämlich für ein ξ dieses Intervalls

$$\gamma := |f(\xi)| > 0,$$

so folgt aus Stetigkeitsgründen

$$|f(x)| \geqq \frac{\gamma}{2}$$

für alle x eines gewissen ξ enthaltenden Teilintervalls von $[\![a, b]\!]$ und daraus, wenn l dessen Länge bedeutet,

$$\int_a^b f(x)^2 \, dx \geqq \frac{\gamma^2}{4} l > 0.$$

Damit erhalten wir

$$\|f\|_2^2 = \langle f, f \rangle. \tag{3}$$

Allgemein hat über einem unitären Raum bezüglich des Skalarproduktes $(\cdot, \cdot)$ die durch

$$\begin{aligned} &\|\cdot\| : E \to \mathbf{R}, \\ &\|f\| := \sqrt{(f, f)}, \qquad f \in E, \end{aligned} \tag{4}$$

definierte Abbildung die Eigenschaften einer Norm. Das ist die Aussage von MfL Bd. 3, 7.2., Satz 1, wenn man dort $B(\boldsymbol{x}, \boldsymbol{y})$ durch $(\boldsymbol{x}, \boldsymbol{y})$ und $\mathbf{R}^n$ durch E ersetzt. Auch im folgenden wird diese Substitution bei Verweisen auf MfL Bd. 3 vorzunehmen sein, ohne daß dies noch besonders erwähnt wird. Es ist in jedem Fall leicht zu erkennen, daß die Verallgemeinerung der Sachverhalte auf beliebige lineare bzw. unitäre Räume möglich ist.

Ein unitärer Raum ist bezüglich (4) *streng normiert:* Für $x, y \in \overline{U}_r(f)$, $x \neq y$ und $\lambda, \mu > 0$, $\lambda + \mu = 1$ ergibt sich mit Beachtung der Schwarzschen Ungleichung

$$\begin{aligned} \|\lambda x + \mu y - f\|^2 &= \|\lambda(x - f) + \mu(y - f)\|^2 \\ &= \lambda^2 \|x - f\|^2 + \mu^2 \|y - f\|^2 + 2\lambda\mu(x - f, y - f) \\ &\leqq \lambda^2 \|x - f\|^2 + \mu^2 \|y - f\|^2 + 2\lambda\mu \|x - f\| \, \|y - f\|. \end{aligned}$$

Daraus folgt, wenn wenigstens eines der Elemente x, y innerer Punkt von $\overline{U}_r(f)$ ist,

$$\|\lambda x + \mu y - f\|^2 < r^2(\lambda^2 + \mu^2 + 2\lambda\mu) = r^2,$$

d. h., $\lambda x + \mu y$ ist innerer Punkt von $\overline{U}_r(f)$. Gilt aber $\|x - f\| = r$ und $\|y - f\| = r$, so gewinnt man dieses Resultat jedenfalls dann, wenn in der Schwarzschen Ungleichung nicht das Gleichheitszeichen steht. Dieses kann wegen $x \neq y$ nur auftreten, wenn $x - f = f - y$ ist. Dann gilt aber

$$\|\lambda x + \mu y - f\|^2 = r^2(\lambda^2 + \mu^2 - 2\lambda\mu) = r^2(\lambda - \mu)^2 < r^2.$$

Nach Satz 3 besitzt daher das mit (4) formulierte Approximationsproblem (23) genau eine Lösung. Wir werden im folgenden Abschnitt noch einen anderen Beweis für die Einzigkeit kennenlernen.

Die eingangs erwähnte Besonderheit hat bemerkenswerte Konsequenzen für die Lösung des mit einer solchen Norm formulierten Approximationsproblems. Wir werden diese im weiteren genauer untersuchen und präzisieren zunächst die zu betrachtende Spezialisierung von 5.1. (23), die man als das *Problem der Quadratmittelapproximation* bezeichnet:

E sei ein unitärer Raum von Funktionen[1] $f: X \to \mathbf{R}$ *bezüglich des Skalarproduktes* $(\cdot,\cdot)$ *und* $\|\cdot\|: E \to \mathbf{R}$ *die durch* (4) *definierte Norm;* $\{\varphi_i\}$, $i = 1, 2, \ldots, n$, *bedeutet ein System linear unabhängiger Elemente aus E. Für eine gegebene Funktion* $f \in E$ *wird ein Vektor* $\boldsymbol{a}^* \in \mathbf{R}^n$ *derart gesucht, daß*

$$\|f - F_{\boldsymbol{a}^*}\|^2 \leqq \|f - F_{\boldsymbol{a}}\|^2 \tag{5}$$

für alle $\boldsymbol{a} \in \mathbf{R}^n$, $F_{\boldsymbol{a}} = \sum_{i=1}^{n} a_i \varphi_i$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$ *ist.*

Bemerkung 1. Offensichtlich sind die Minimumprobleme für die durch $\|f - F_{\boldsymbol{a}}\|$ und $\|f - F_{\boldsymbol{a}}\|^2$ bestimmten Distanzfunktionen äquivalent. Auf Grund von (4) ist es zweckmäßig, mit der Zielfunktion

$$Z: \boldsymbol{a} \mapsto \|f - F_{\boldsymbol{a}}\|^2$$

zu arbeiten.

Bemerkung 2. Die Bezeichnung für das Problem (5) nimmt auf den Spezialfall

$$E = C_{\langle a,b \rangle}, \qquad (\cdot,\cdot) = \langle \cdot, \cdot \rangle$$

Bezug, da hier die bis auf einen konstanten Faktor mit $\|f - F_{\boldsymbol{a}}\|$ übereinstimmende Größe

$$\sqrt{\frac{1}{b-a} \int_a^b (f - F_{\boldsymbol{a}})^2 \, dx}$$

die sogenannte mittlere quadratische Abweichung der Funktionen f und $F_{\boldsymbol{a}}$ ist.

5.2.2. Die Normalgleichungen

Das Problem (5) besitzt nach 5.1.2., Satz 2, eine Lösung $\boldsymbol{a}^*$, für welche die Distanzfunktion

$$\begin{aligned} Z = Z(\boldsymbol{a}) &= (f - F_{\boldsymbol{a}}, f - F_{\boldsymbol{a}}) \\ &= (f, f) - 2(f, F_{\boldsymbol{a}}) + (F_{\boldsymbol{a}}, F_{\boldsymbol{a}}) \\ &= (f, f) - 2 \sum_{i=1}^{n} a_i (f, \varphi_i) + \sum_{i=1}^{n} \sum_{k=1}^{n} a_i a_k (\varphi_i, \varphi_k) \end{aligned} \tag{6}$$

[1] Vgl. dazu die Fußnote auf S. 10.

ein absolutes Minimum annimmt. Da Z offensichtlich differenzierbar ist, muß

$$\left.\frac{\partial Z}{\partial a_j}\right|_{a=a^*} = 0, \qquad j = 1, 2, \ldots, n, \tag{7}$$

gelten. Man findet mit Benutzung des Kronecker-Symbols

$$\delta_{ij} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases}$$

die Beziehungen

$$\begin{aligned} \frac{\partial Z}{\partial a_j} &= -2(f, \varphi_j) + \sum_{i=1}^{n} \delta_{ij} \sum_{k=1}^{n} a_k(\varphi_i, \varphi_k) + \sum_{i=1}^{n} a_i \sum_{k=1}^{n} \delta_{kj}(\varphi_i, \varphi_k) \\ &= -2(f, \varphi_j) + \sum_{k=1}^{n} a_k(\varphi_j, \varphi_k) + \sum_{i=1}^{n} a_i(\varphi_i, \varphi_j) \\ &= -2(f, \varphi_j) + 2\sum_{i=1}^{n} a_i(\varphi_i, \varphi_j) \end{aligned}$$

und an Stelle von (7) die sogenannten *Normalgleichungen*

$$\sum_{i=1}^{n} a_i(\varphi_i, \varphi_j) = (f, \varphi_j), \qquad j = 1, 2, \ldots, n, \tag{8}$$

die in den a_i linear sind.

Satz 1. *Für das Problem der Quadratmittelapproximation besitzt das System* (8) *der Normalgleichungen — und damit* (5) *— genau eine Lösung.*

Das folgt aus dem Nichtverschwinden der Koeffizientendeterminante, welches mit dem nächsten Satz bewiesen wird.

Satz 2. *E sei ein unitärer Raum bezüglich des Skalarproduktes* $(\cdot, \cdot)$ *und* $\{\varphi_i\}$, $i = 1, 2, \ldots, n$, *ein System von Elementen aus E. Dann ist die sogenannte Gramsche Determinante*

$$\begin{vmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) & \ldots & (\varphi_1, \varphi_n) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) & \ldots & (\varphi_2, \varphi_n) \\ \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots\ldots \\ (\varphi_n, \varphi_1) & (\varphi_n, \varphi_2) & \ldots & (\varphi_n, \varphi_n) \end{vmatrix} \tag{9}$$

genau dann von Null verschieden, wenn $\{\varphi_i\}$ *ein System linear unabhängiger Elemente ist.*

Beweis. Wir nehmen zunächst an, daß die φ_i linear abhängig sind und etwa

$$\varphi_j = \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \varphi_i \tag{10}$$

gilt. Auf Grund der Definition 1, Bedingungen 3 und 4, erweist sich dann in (9) die j-te Zeile nach Ersetzung des ersten Operanden φ_j gemäß (10) als eine Linearkombination der übrigen Zeilen, d. h., die Gramsche Determinante verschwindet. Wäre dies andererseits bei linearer Unabhängigkeit der φ_i der Fall, so existierten Konstanten $\beta_i \in \mathbf{R}$ $(i = 1, 2, \ldots, n)$, für die

$$\sum_{i=1}^{n} |\beta_i| > 0 \tag{11}$$

und

$$\sum_{i=1}^{n} \beta_i(\varphi_j, \varphi_i) = 0, \qquad j = 1, 2, \ldots, n, \tag{12}$$

ist. Wir betrachten die Linearkombination

$$F_\beta = \sum_{i=1}^{n} \beta_i \varphi_i \tag{13}$$

und erhalten wegen (12) $(\varphi_j, F_\beta) = 0$ für $j = 1, 2, \ldots, n$. Damit ist auch

$$(F_\beta, F_\beta) = \|F_\beta\|^2 = 0,$$

d. h., (13) ist das Nullelement von E. Auf Grund der linearen Unabhängigkeit der φ_i folgt daraus $\beta_i = 0$ für $i = 1, 2, \ldots, n$. Dem widerspricht aber (11).

Die Bestimmung der besten Quadratmittelapproximation kann somit durch Lösen des Gleichungssystems (8), wesentlich also durch Invertierung seiner Koeffizientenmatrix erfolgen. Letzteres ist von dem zu approximierenden f unabhängig und läßt sich für ein gegebenes System $\{\varphi_i\}$ ein für alle Mal durchführen. Auf die numerischen Probleme, die bei der Lösung der Normalgleichungen auftreten, werden wir in 6.1. eingehen.

5.2.3. Orthogonalsysteme

Besonders einfach ist die Lösung der Normalgleichungen, wenn das System (8) Diagonalgestalt hat, d. h., wenn (φ_i, φ_j) für $i \neq j$ verschwindet und für $i = j$ positiv ist. Derartige Elementsysteme $\{\varphi_i\}$ nennt man *orthogonal*, speziell *orthonormiert*, wenn

$$(\varphi_i, \varphi_j) = \delta_{ij}$$

(vgl. MfL Bd. 3, 7.3.). Ein Orthogonalsystem $\{\varphi_i\}$ ist linear unabhängig. Das folgt unmittelbar aus MfL Bd. 3, 7.3., Satz 1, oder aus Satz 2.

Im Prinzip kann man bei der Quadratmittelapproximation immer von der Annahme ausgehen, daß das Elementsystem $\{\varphi_i\}$ orthonormiert ist. Nach MfL Bd. 3, 7.3., Satz 2, ist nämlich jedes linear unabhängige Elementsystem $\varphi_1, \varphi_2, \ldots, \varphi_n$ durch ein orthonormiertes System $\psi_1, \psi_2, \ldots, \psi_n$ ersetzbar, dessen lineare Hülle mit

der von $\{\varphi_i\}$ übereinstimmt. Der Beweis dieses Satzes enthält ein Verfahren zur Konstruktion der ψ_i. Wir wiederholen — ohne auf Einzelheiten einzugehen — die wesentlichen Schritte der von E. SCHMIDT stammenden Methode:

Zunächst wird mit Bezug auf (4)

$$\psi_1 = \frac{\varphi_1}{\|\varphi_1\|} \tag{14}$$

gesetzt. Sodann bestimmt man ein Element χ_2 der Form

$$\chi_2 = \varphi_2 - \lambda_1^{(2)}\psi_1,$$

für welches $\chi_2 \perp \psi_1$ gilt. Das führt auf

$$\lambda_1^{(2)} = (\varphi_2, \psi_1).$$

ψ_2 ergibt sich aus χ_2 durch Normierung:

$$\psi_2 = \frac{\chi_2}{\|\chi_2\|}.$$

Sind $\psi_1, \psi_2, \ldots, \psi_k$ $(k < n)$ bereits konstruiert, so bildet man

$$\chi_{k+1} = \varphi_{k+1} - \sum_{i=1}^{k} \lambda_i^{(k+1)}\psi_i \tag{15}$$

und fordert $\chi_{k+1} \perp \psi_1, \psi_2, \ldots, \psi_k$. Daraus folgt

$$\lambda_i^{(k+1)} = (\varphi_{k+1}, \psi_i), \qquad i = 1, 2, \ldots, k. \tag{16}$$

ψ_{k+1} ergibt sich aus χ_{k+1} durch Normierung:

$$\psi_{k+1} = \frac{\chi_{k+1}}{\|\chi_{k+1}\|}. \tag{17}$$

Man beachte, daß das Verfahren in dem Sinne endgültig ist, daß man bei Hinzunahme weiterer Elemente $\varphi_{n+1}, \varphi_{n+2}, \ldots$ (unter Wahrung der linearen Unabhängigkeit) die Konstruktion der ψ fortsetzen kann, ohne die schon bestimmten Elemente zu ändern.

Für ein Orthogonalsystem $\{\varphi_i\}$, $\varphi_i \in E$, folgt aus (8)

$$a_i^* = \frac{(f, \varphi_i)}{\|\varphi_i\|^2}, \qquad i = 1, 2, \ldots, n. \tag{18}$$

Diese Größen werden die *Fourierkoeffizienten* von f bezüglich des Systems $\{\varphi_i\}$ genannt. Nach Satz 1 ist die mit den Koeffizienten (18) gebildete Linearkombination die beste Quadratmittelapproximation von f. Man nennt das gelegentlich die *Minimaleigenschaft der Fourierkoeffizienten*. Dafür werden wir einen weiteren direkten

Beweis finden im Zusammenhang mit der folgenden Herleitung einer wichtigen Abschätzung für die a_j^*.

Im Falle eines Orthogonalsystems ergibt sich aus (6) mit (18)

$$Z(\boldsymbol{a}) = \|f - F_{\boldsymbol{a}}\|^2 = \|f\|^2 - 2\sum_{i=1}^{n} a_i(f, \varphi_i) + \sum_{i=1}^{n} a_i^2 \|\varphi_i\|^2$$

$$= \|f\|^2 - 2\sum_{i=1}^{n} a_i a_i^* \|\varphi_i\|^2 + \sum_{i=1}^{n} a_i^2 \|\varphi_i\|^2 \tag{19}$$

und für $a_i = a_i^*$

$$Z(\boldsymbol{a}^*) = \|f - F_{\boldsymbol{a}^*}\|^2 = \|f\|^2 - \sum_{i=1}^{n} a_i^{*2} \|\varphi_i\|^2. \tag{20}$$

Damit folgt

$$Z(\boldsymbol{a}) - Z(\boldsymbol{a}^*) = \sum_{i=1}^{n} (a_i^* - 2a_i a_i^* + a_i^2) \|\varphi_i\|^2 = \sum_{i=1}^{n} (a_i^* - a_i)^2 \|\varphi_i\|^2 \geqq 0. \tag{21}$$

Das Gleichheitszeichen gilt in (21) genau dann, wenn $\boldsymbol{a} = \boldsymbol{a}^*$ ist, d. h., die mit den Fourierkoeffizienten von f gebildete Linearkombination und nur diese ist beste Quadratmittelapproximation von f. Aus (20) erhalten wir

$$\|f\|^2 - \sum_{i=1}^{n} a_i^{*2} \|\varphi_i\|^2 \geqq 0,$$

also

$$\sum_{i=1}^{n} a_i^{*2} \|\varphi_i\|^2 \leqq \|f\|^2, \tag{22}$$

und speziell für ein Orthonormalsystem

$$\sum_{i=1}^{n} a_i^{*} \leqq \|f\|^2. \tag{22a}$$

(22) wird als *Besselsche Ungleichung* bezeichnet.

Wir wollen annehmen, daß

$$\varphi_1, \varphi_2, \ldots, \varphi_k, \ldots \tag{23}$$

eine unendliche Folge zueinander orthogonaler und linear unabhängiger Elemente aus E ist. Damit ist gemeint, daß jedes endliche Teilsystem $\{\varphi_i\}$, $i = 1, 2, \ldots, n$, von (23) orthogonal und linear unabhängig ist. Zunächst sei darauf hingewiesen, daß die für ein System $\{\varphi_i\}$, $i = 1, 2, \ldots, n$, bestimmten a_i^* auf Grund von (18) keine Änderung erfahren, wenn man zu einem erweiterten System der φ_i übergeht (*Endgültigkeit der Fourierkoeffizienten*). Die unendliche Reihe

$$\sum_{i=1}^{\infty} a_i^* \|\varphi_i\|^2 \tag{24}$$

konvergiert, da ihre Partialsummen auf Grund von (22) beschränkt sind. Notwendigerweise ist dann

$$\lim_{i\to\infty} a_i^{*2}\,\|\varphi_i\|^2 = 0. \tag{25}$$

Wir betrachten Beispiele.

1. Es sei $E = C_{(0,p)}$ und $(\cdot,\cdot)$ das gemäß (2) entsprechend gebildete Skalarprodukt.

Satz 3. *Für die Folge der durch*

$$\begin{aligned}
&\varphi_0(x) = \frac{1}{2}, \qquad \varphi_1(x) = \sin\frac{2\pi}{p}x, \qquad \varphi_2(x) = \cos\frac{2\pi}{p}x,\\
&\qquad\qquad \varphi_3(x) = \sin\frac{4\pi}{p}x, \qquad \varphi_4(x) = \cos\frac{4\pi}{p}x,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\qquad\qquad \varphi_{2n-1}(x) = \sin\frac{2\pi}{p}nx, \quad \varphi_{2n}(x) = \cos\frac{2\pi}{p}nx
\end{aligned} \tag{26}$$

definierten Funktionen gilt

$$(\varphi_i, \varphi_j) = 0 \qquad \text{für} \quad i \neq j,$$

$$(\varphi_i, \varphi_i) = \begin{cases} \dfrac{p}{4} & \text{für} \quad i = 0, \\ \dfrac{p}{2} & \text{für} \quad i \neq 0. \end{cases} \tag{27}$$

Die Funktionen (26) *bilden also ein Orthogonalsystem. Wir bezeichnen die Fourierkoeffizienten einer Funktion* $f \in C_{(0,p)}$ *zu* φ_{2k}, φ_{2k-1} *mit* a_k^* *bzw.* b_k^*. *Dann ist*

$$\begin{aligned}
a_k^* &:= \frac{(f, \varphi_{2k})}{\|\varphi_{2k}\|^2} = \frac{2}{p}\int_0^p f(x)\cos\frac{2\pi}{p}kx\,dx, \qquad k = 0, 1, 2, \ldots,\\
b_k^* &:= \frac{(f, \varphi_{2k-1})}{\|\varphi_{2k-1}\|^2} = \frac{2}{p}\int_0^p f(x)\sin\frac{2\pi}{p}kx\,dx, \qquad k = 1, 2, \ldots .
\end{aligned} \tag{28}$$

Unter allen Linearkombinationen der Funktionen $\varphi_0, \varphi_1, \ldots, \varphi_{2n}$ *ist*

$$F = \frac{a_0^*}{2} + \sum_{\nu=1}^{n}\left(a_\nu^*\cos\frac{2\pi}{p}\nu x + b_\nu^*\sin\frac{2\pi}{p}\nu x\right) \tag{29}$$

die beste Quadratmittelapproximation von f.

Beweis. Mit Bezugnahme auf die in 2.2. und 2.3. entwickelte Theorie reduziert sich der Beweis auf die Verifikation der folgenden Integralformeln, in denen $k, l \in \mathbf{N}$ ist:

$$\int_0^p \cos\frac{2\pi}{p}kx \sin\frac{2\pi}{p}lx = 0 \qquad \text{für beliebige } k, l,$$

$$\int_0^p \cos\frac{2\pi}{p}kx \cos\frac{2\pi}{p}lx = \begin{cases} 0 & \text{für } k \neq l, \\ \frac{p}{2} & \text{für } k = l \neq 0, \\ p & \text{für } k = l = 0, \end{cases} \tag{30}$$

$$\int_0^p \sin\frac{2\pi}{p}kx \sin\frac{2\pi}{p}lx = \begin{cases} 0 & \text{für } k \neq l, \\ \frac{p}{2} & \text{für } k = l \neq 0, \\ 0 & \text{für } k = l = 0. \end{cases}$$

Diese ergeben sich leicht mit Hilfe der trigonometrischen Beziehungen

$$2\cos u \sin v = \sin(u+v) - \sin(u-v),$$
$$2\cos u \cos v = \cos(u+v) + \cos(u-v),$$
$$2\sin u \sin v = \cos(u-v) - \cos(u+v).$$

Bemerkung 3. Satz 3 bleibt gültig, wenn man an Stelle von $[\![0, p]\!]$ ein beliebiges Intervall $[\![a, b]\!]$ der Länge p betrachtet, die Definition des Skalarproduktes gemäß (2) entsprechend ändert und die Integrale (28) mit den Grenzen a, b schreibt. Das folgt aus der Tatsache, daß die Integralformeln (30) auch gelten, wenn man an Stelle von $0, p$ die Grenzen a, b einsetzt. Zur Begründung sei an folgendes erinnert: Eine Funktion $f: \mathbf{R} \to \mathbf{R}$ heißt periodisch mit der Periode p ($p > 0$), wenn für alle $x \in \mathbf{R}$

$$f(x+p) = f(x) \tag{31}$$

gilt. Ist f bezüglich eines bestimmten Intervalls I der Länge p integrabel, so gilt das für jedes derartige Intervall, und die entsprechenden Integralwerte sind gleich (vgl. Abb. 5.3).

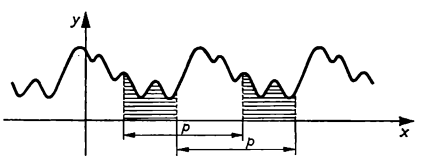


Abb. 5.3

Die Berechnung der Fourierkoeffizienten (28) wird als *harmonische Analyse* bezeichnet.

2. Wir betrachten die über einem Intervall $[\![0, p]\!]$ stückweise stetigen Funktionen. Zwei derartige Funktionen f, g seien *äquivalent* genannt ($f \sim g$) genau dann, wenn die Menge

$$M = \{x: x \in [\![0, p]\!] \wedge f(x) \neq g(x)\}$$

endlich ist. Die Relation $\sim$ ist offenbar reflexiv, symmetrisch und transitiv. Wir bezeichnen die davon erzeugten Äquivalenzklassen mit $\hat{f}, \hat{g}, \ldots$, wenn diese f bzw. g bzw. ... als Repräsentanten enthalten. In der Gesamtheit E dieser Klassen definieren wir eine Addition und Vervielfältigung mit Skalaren durch

$$\hat{f} + \hat{g} = \widehat{f + g}, \tag{32a}$$

$$\lambda \hat{f} = \widehat{\lambda f}, \qquad \lambda \in \mathbf{R}. \tag{32b}$$

Offenbar sind diese Festlegungen repräsentantenunabhängig und genügen den Axiomen des linearen Raumes. Nullvektor ist das durch die auf $[\![0, p]\!]$ identisch verschwindende Funktion repräsentierte Element aus E. Schließlich verifiziert man leicht, daß durch

$$(\hat{f}, \hat{g}) := \int_0^p f(x)\, g(x)\, dx \tag{33}$$

in repräsentantenunabhängiger Weise ein Skalarprodukt in E definiert wird. Zu den Funktionen (26) betrachten wir die Vektoren

$$\phi_0, \phi_1, \ldots, \phi_{2n}, \tag{34}$$

die auf Grund von (33) und (27) ein Orthogonalsystem bilden. Die beste Quadratmittelapproximation eines Elementes $\hat{f} \in E$ durch Linearkombinationen der ϕ_i wird mit den Fourierkoeffizienten (18) gewonnen. Wir wählen dafür entsprechend die gleichen Bezeichnungen wie im vorigen Beispiel und erhalten

$$\begin{aligned} a_k^* &:= \frac{(\hat{f}, \phi_{2k})}{\|\phi_{2k}\|^2} = \frac{2}{p} \int_0^p f(x) \cos \frac{2\pi}{p} kx\, dx, \qquad k = 0, 1, 2, \ldots, \\ b_k^* &:= \frac{(\hat{f}, \phi_{2k-1})}{\|\phi_{2k-1}\|^2} = \frac{2}{p} \int_0^p f(x) \sin \frac{2\pi}{p} kx\, dx, \qquad k = 1, 2, \ldots. \end{aligned} \tag{35}$$

Daraus folgt:

Satz 4. *Ist f ein stetiger Repräsentant von $\hat{f} \in E$, so erweist sich die Funktion* (29) *als ein Repräsentant der aus den Elementen ϕ_i, $i = 0, 1, \ldots, 2n$, gebildeten Bestapproximation von $\hat{f}$.*

Bemerkung 4. Im Sinne der Bemerkung 3 lassen sich die Betrachtungen dieses Beispiels auf beliebige Intervalle $[\![a, b]\!]$ der Länge p übertragen. Eine auf $[\![a, b]\!]$ stückweise stetige Funktion f kann gemäß (31) periodisch über die ganze Zahlengerade fortgesetzt werden. Die Definition der Funktionswerte in den Punkten $a + kp$ (k beliebig ganz) ist dabei unbestimmt, wenn $f(a) \neq f(b)$ ist. Wir wollen daher $f(a + kp)$ durch einen beliebigen von k unabhängigen Wert c erklären und die so für alle reellen x definierte Funktion mit f_c bezeichnen und eine *periodische Fortsetzung* von f nennen. Dann gilt offenbar

Satz 5. *Es sei E der unitäre Raum der Klassen äquivalenter auf $[\![a, b]\!]$ stückweise stetiger Funktionen, $p := b - a$ und ϕ_i das Orthogonalsystem* (34). *Dann können die Fourierkoeffizienten eines Elementes $\hat{f} \in E$ mit einer beliebigen periodischen Fortsetzung f_c eines Repräsentanten von $\hat{f}$*

gemäß

$$a_k^* = \frac{2}{p} \int_{(I)} f_c(x) \cos \frac{2\pi}{p} kx \, dx, \qquad k = 0, 1, \ldots,$$

$$b_k^* = \frac{2}{p} \int_{(I)} f_c(x) \sin \frac{2\pi}{p} kx \, dx, \qquad k = 1, 2, \ldots,$$

berechnet werden, wobei I ein beliebiges Intervall der Länge p bedeutet.

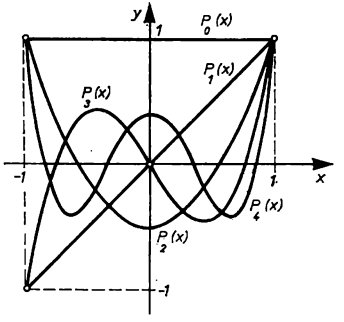


Abb. 5.4

3. Durch

$$\varphi_n(x) = P_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} [(x^2 - 1)^n],^{1}) \qquad n \in \mathbf{N}, \tag{36}$$

sind die *Legendreschen Polynome* definiert. Es ist zum Beispiel (vgl. Abb. 5.4)

$$\begin{aligned}
\varphi_0(x) &= 1, \\
\varphi_1(x) &= x, \\
\varphi_2(x) &= \frac{1}{2}(3x^2 - 1), \\
\varphi_3(x) &= \frac{1}{2}(5x^3 - 3x), \\
\varphi_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \\
\varphi_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x).
\end{aligned} \tag{37}$$

[1]) Diese Darstellung der Legendreschen Polynome wird als *Formel von Rodrigues* bezeichnet.

Betrachtet man die φ_n als Funktionen des Raumes $C_{[-1,1]}$, so sind diese bezüglich des entsprechenden Skalarproduktes (2) orthogonal.

Durch partielle Integration ergibt sich, wenn $m \leqq n$ ist,

$$
\begin{aligned}
(\varphi_n, \varphi_m) &= \int_{-1}^{1} \varphi_n(x)\, \varphi_m(x)\, dx \\
&= \frac{1}{2^n n!}\, \frac{1}{2^m m!} \int_{-1}^{1} \frac{d^n[(x^2-1)^n]}{dx^n}\, \frac{d^m[(x^2-1)^m]}{dx^m}\, dx \\
&= \frac{1}{2^n n!}\, \frac{1}{2^m m!} \left\{ \left[\frac{d^{n-1}[(x^2-1)^n]}{dx^{n-1}}\, \frac{d^m[(x^2-1)^m]}{dx^m} \right]_{-1}^{1} \right. \\
&\quad \left. - \int_{-1}^{1} \frac{d^{n-1}[(x^2-1)^m]}{dx^{n-1}}\, \frac{d^{m+1}[(x^2-1)^m]}{dx^{m+1}}\, dx \right\}.
\end{aligned}
$$

Der integralfreie Term verschwindet, da die $(n-1)$-te Ableitung von $(x^2-1)^n$ offensichtlich bei ± 1 eine Nullstelle hat. Das verbleibende Integral wird entsprechend behandelt; so fortfahrend erhält man nach n Schritten

$$
\int_{-1}^{1} \varphi_n(x)\, \varphi_m(x)\, dx = \frac{1}{2^n n!}\, \frac{1}{2^m m!}\, (-1)^n \int_{-1}^{1} (x^2-1)^n\, \frac{d^{n+m}[(x^2-1)^m]}{dx^{n+m}}\, dx. \tag{38}
$$

Wenn $m < n$ und daher $n + m > 2m$ ist, verschwindet das Integral auf der rechten Seite von (38), d. h.

$$
(\varphi_n, \varphi_m) = 0 \quad \text{für} \quad m \neq n.
$$

Im Fall $m = n$ folgt aus (38)

$$
\|\varphi_n\|^2 = \int_{-1}^{1} \varphi_n^2\, dx = \left[\frac{1}{2^n n!} \right]^2 (-1)^n\, (2n)! \int_{-1}^{1} (x^2-1)^n\, dx.
$$

Das letzte Integral geht mit $x = \cos t$ in

$$
\int_{-1}^{1} (x^2-1)^n\, dx = 2 \int_{0}^{1} (x^2-1)^n\, dx = 2(-1)^n \int_{0}^{\pi/2} \sin^{2n+1} t\, dt
$$

über; partielle Integration liefert weiter für $n \geqq 1$

$$
\begin{aligned}
\int_{0}^{\pi/2} \sin^{2n+1} t\, dt &= [-\sin^{2n} t \cos t]_0^{\pi/2} + 2n \int_{0}^{\pi/2} \sin^{2n-1} t \cos^2 t\, dt \\
&= 2n \int_{0}^{\pi/2} \sin^{2n-1} t\, dt - 2n \int_{0}^{\pi/2} \sin^{2n+1} t\, dt,
\end{aligned}
$$

also

$$
\int_{0}^{\pi/2} \sin^{2n+1} t\, dt = \frac{2n}{2n+1} \int_{0}^{\pi/2} \sin^{2n-1} t\, dt. \tag{39}
$$

Mehrmalige Anwendung der Rekursionsformel (39) ergibt

$$\int_0^{\pi/2} \sin^{2n+1} t \, dt = \frac{2n}{2n+1} \cdot \frac{2n-2}{2n-1} \cdots \frac{2}{3} \cdot 1, \tag{40}$$

so daß

$$\|\varphi_n\|^2 = 2 \frac{2^n n! \, (2n)!}{(2n+1)! \, 2^n n!} = \frac{2}{2n+1} \tag{41}$$

ist; (41) gilt offenbar auch im Falle $n = 0$.

Die beste Quadratmittelapproximation einer Funktion $f \in C_{[-1,1]}$ durch eine Linearkombination der Legendreschen Polynome $\varphi_0, \varphi_1, \ldots, \varphi_n$ ergibt sich, wenn diese mit den entsprechenden Fourierkoeffizienten gebildet wird. Gemäß (18), (36) und (41) findet man

$$\begin{aligned} a_j^* &= \frac{(f, \varphi_j)}{\|\varphi_j\|^2} = \frac{2j+1}{2} \int_{-1}^{1} f(x) \, \varphi_j(x) \, dx \\ &= \frac{2j+1}{2^{j+1} j!} \int_{-1}^{1} f(x) \, \frac{d^j}{dx^j} [(x^2 - 1^j)] \, dx, \qquad j = 0, 1, 2, \ldots . \end{aligned} \tag{42}$$

Da die Legendreschen Polynome auch Kugelfunktionen genannt werden, bezeichnet man die Berechnung der Koeffizienten (42) gelegentlich als *Kugelanalyse*.

4. Die Größe $\cos nt$ läßt sich für jedes $t \in \mathbf{R}$ und $n \in \mathbf{N}$ durch ein Polynom n-ten Grades in $x = \cos t$ darstellen. Beispielsweise ist

$$\begin{aligned} \cos 2t &= 2x^2 - 1, \\ \cos 3t &= 4x^3 - 3x, \\ \cos 4t &= 8x^4 - 8x^2 + 1, \\ \cos 5t &= 16x^5 - 20x^3 + 5x. \end{aligned} \tag{43}$$

Allgemein definiert man durch

$$\cos (n \arccos x) = T_n(x),^1) \qquad n = 0, 1, \ldots, \tag{44}$$

die *Tschebyscheff-Polynome* (1. Art) T_n (vgl. Abb. 5.5). Um zu zeigen, daß $\cos (n \arccos x)$ tatsächlich ein Polynom n-ten Grades in x ist, gehen wir von der Moivreschen Formel (MfL Bd. 2, 7.3.) aus. Für $|x| \leqq 1$, $t = \arccos x$ und $n \in \mathbf{N}$ ist

$$\begin{aligned} \cos nt + i \sin nt &= (\cos t + i \sin t)^n, \\ \cos nt - i \sin nt &= (\cos t - i \sin t)^n, \\ \cos nt &= \frac{1}{2} \left[\left(x + i\sqrt{1-x^2}\right)^n + \left(x - i\sqrt{1-x^2}\right)^n \right]. \end{aligned}$$

[1]) Hier kann ein beliebiger Zweig des Arkuskosinus gewählt werden.

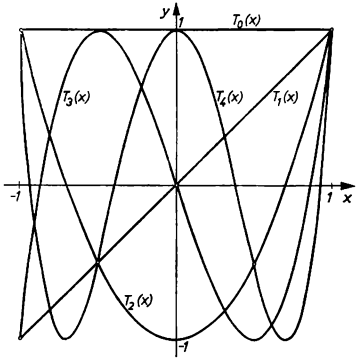


Abb. 5.5

Mit Hilfe des binomischen Satzes erkennt man, daß sich auf der rechten Seite die Terme mit ungeraden Potenzen von $\sqrt{1-x^2}$ aufheben und ein Polynom n-ten Grades in x resultiert:

$$\cos nt = x^n - \binom{n}{2} x^{n-2}(1-x^2) + \binom{n}{4} x^{n-4}(1-x^2)^2 - \binom{n}{6} x^{n-6}(1-x^2)^3 \pm \cdots$$

Dasselbe Polynom ergibt sich, wenn man für $|x| > 1$ den Ausdruck

$$\frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^n + \left(x-\sqrt{x^2-1}\right)^n\right]$$

nach dem binomischen Satz entwickelt, so daß

$$T_n(x) = x^n - \binom{n}{2} x^{n-2}(1-x^2) + \binom{n}{4} x^{n-4}(1-x^2)^2 - \binom{n}{6} x^{n-6}(1-x^2)^3 \pm \cdots$$

$$= \begin{cases} \frac{1}{2}\left[\left(x+i\sqrt{1-x^2}\right)^n + \left(x-i\sqrt{1-x^2}\right)^n\right] & \text{für } |x| \leqq 1, \\ \frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^n + \left(x-\sqrt{x^2-1}\right)^n\right] & \text{für } |x| > 1. \end{cases} \tag{45}$$

Damit folgt

$$\lim_{x\to\infty} \frac{T_n(x)}{x^n} = \lim_{x\to\infty} \frac{1}{2}\left[\left(1+\sqrt{1-\frac{1}{x^2}}\right)^n + \left(1-\sqrt{1-\frac{1}{x^2}}\right)^n\right] = 2^{n-1}. \tag{46}$$

(46) zeigt, daß T_n den Leitkoeffizienten 2^{n-1} besitzt.

Für

$$x_k^{(n)} := \cos\frac{(2k-1)\,\pi}{2n}, \qquad k = 1, 2, \ldots, n,$$

ist nach (44)

$$T_n(x_k^{(n)}) = \cos\left(n \arccos x_k^{(n)}\right) = \cos\left((2k-1)\frac{\pi}{2}\right) = 0$$

(vgl. Abb. 5.6). Damit sind sämtliche Nullstellen von T_n bestimmt.

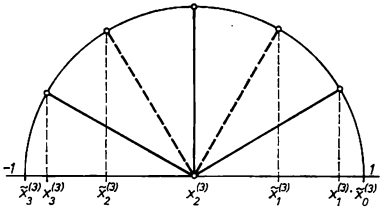


Abb. 5.6 Null- und Extremstellen von T_3

Auf $[\![-1, 1]\!]$ ist gemäß (44) $|T_n(x)| \leqq 1$. Von Interesse sind die Stellen in $[\![-1, 1]\!]$, wo T_n Werte vom Betrag 1 annimmt. Nach (44) sind das die $n + 1$ Argumente

$$\bar{x}_l^{(n)} = \cos\frac{l\pi}{n}, \qquad l = 0, 1, 2, \ldots, n,$$

wofür

$$T_n(\bar{x}_l^{(n)}) = \cos l\pi = (-1)^l$$

gilt (vgl. Abb. 5.6). Wir fassen die Ergebnisse in einem Satz zusammen:

Satz 6. *Die durch* (44) *definierten* T_n *sind Polynome n-ten Grades mit dem Leitkoeffizienten* 2^{n-1}. *Diese besitzen n reelle Nullstellen im Intervall* $[\![-1, 1]\!]$ *bei*

$$x_k^{(n)} = \cos\frac{(2k-1)\,\pi}{2n}, \qquad k = 1, 2, \ldots, n. \tag{47}$$

Auf $|x| \leqq 1$ *ist*

$$|T_n(x)| \leqq 1; \tag{48}$$

das Gleichheitszeichen gilt für

$$\bar{x}_l^{(n)} = \cos\frac{l\pi}{n}, \qquad l = 0, 1, \ldots, n, \tag{49}$$

wobei das Vorzeichen der Funktionswerte gemäß $T_n(\bar{x}_l^{(n)}) = (-1)^l$ *alterniert.*

Im Raum $C_{(-1,1)}$ führen wir durch

$$(f, g) = \int_{-1}^{1} \frac{f(x)\, g(x)}{\sqrt{1 - x^2}}\, dx, \qquad f, g \in C_{(-1,1)}, \tag{50}$$

ein Skalarprodukt ein. Vor Verifikation der Eigenschaften in Definition 1 wird die Existenz des uneigentlichen Integrals gezeigt. Gemäß MfL Bd. 5, 4.1.7., ist

$$\int_{-1}^{1} \frac{f(x)\, g(x)}{\sqrt{1 - x^2}}\, dx = \lim_{a' \downarrow -1} \int_{a'}^{0} \frac{f(x)\, g(x)}{\sqrt{1 - x^2}}\, dx + \lim_{b' \uparrow 1} \int_{0}^{b'} \frac{f(x)\, g(x)}{\sqrt{1 - x^2}}\, dx.$$

Mit der Substitution $x = \cos t$ und

$$F(t) = f(\cos t), \qquad G(t) = g(\cos t) \tag{51}$$

erhält man dafür

$$\int_{-1}^{1} \frac{f(x)\, g(x)}{\sqrt{1 - x^2}}\, dx = \int_{\pi/2}^{\pi} F(t)\, G(t)\, dt + \int_{0}^{\pi/2} F(t)\, G(t)\, dt = \int_{0}^{\pi} F(t)\, G(t)\, dt. \tag{52}$$

Es ist also

$$(f, g) = \langle F, G \rangle, \tag{53}$$

wobei $\langle \cdot, \cdot \rangle$ das in (2) eingeführte Skalarprodukt für das Intervall $[\![0, \pi]\!]$ bedeutet. Aus (53) folgt unmittelbar, daß $(\cdot, \cdot)$ Skalarprodukt in $C_{(-1,1)}$ ist.

Satz 7. *Die Tschebyscheff-Polynome T_n sind bezüglich* (60) *orthogonal; dabei ist*

$$\|T_n\|^2 = \begin{cases} \dfrac{\pi}{2} & \text{für } n \neq 0, \\ \pi & \text{für } n = 0. \end{cases}$$

Beweis. Auf Grund von (44), (53) und (30) ist

$$(T_m, T_n) = \int_{-1}^{1} \frac{T_m(x)\, T_n(x)}{\sqrt{1 - x^2}}\, dx = \int_{0}^{\pi} \cos mt \cos nt\, dt$$

$$= \frac{1}{2} \int_{-\pi}^{\pi} \cos mt \cos nt\, dt = \frac{1}{2} \int_{0}^{2\pi} \cos mt \cos nt\, dt$$

$$= \begin{cases} 0 & \text{für } m \neq n, \\ \dfrac{\pi}{2} & \text{für } m = n \neq 0, \\ \pi & \text{für } m = n = 0. \end{cases}$$

Die bezüglich der Tschebyscheff-Polynome gebildeten Fourierkoeffizienten (18) sind

$$a_i^* = \frac{(f, T_i)}{\|T_i\|^2} = \begin{cases} \dfrac{1}{\pi} \displaystyle\int_0^\pi F(t)\, dt & \text{für } i = 0, \\ \dfrac{2}{\pi} \displaystyle\int_0^\pi F(t) \cos it\, dt & \text{für } i \neq 0, \end{cases} \tag{54}$$

wobei F gemäß (51) zu bilden ist.

Polynomsysteme, die bezüglich eines Skalarproduktes orthogonal sind, besitzen allein auf Grund dieses Umstandes eine Reihe gemeinsamer Eigenschaften. Zum Beispiel gilt der folgende

Satz 8. *$\{P_k\}$, $k = 0, 1, 2, \ldots$, bedeutet ein System von Polynomen k-ten Grades, die (eingeschränkt auf eine Menge $X \subseteq \mathbf{R}$) Elemente eines unitären Raumes E von Funktionen $f: X \to \mathbf{R}$ sind. Bezüglich des in E gegebenen Skalarproduktes sei*

$$(P_j, P_k) \begin{cases} = 0 & \text{für } j \neq k, \\ \neq 0 & \text{für } j = k \end{cases}$$

und

$$(IP_k, P_j) = (P_k, IP_j), \qquad j, k = 0, 1, 2, \ldots, \tag{55}$$

wobei I das Polynom $I(x) = x$ bezeichnet. Dann existieren reelle Konstanten A_k, B_k, C_k ($k = 1, 2, 3, \ldots$) derart, daß für alle $x \in \mathbf{R}$ die Rekursionsformeln

$$P_{k+1}(x) = (A_k x + B_k) P_k(x) + C_k P_{k-1}(x) \tag{56}$$

gelten.

Beweis. E enthält sämtliche Potenzen $1, x, x^2, \ldots, x^k, \ldots$, da diese aus den P_k linear kombinierbar sind.

Es sei p_k der Leitkoeffizient von P_k und

$$A_k := \frac{p_{k+1}}{p_k}.$$

Dann ist $P_{k+1} - A_k IP_k$ ein Polynom maximal k-ten Grades und daher

$$P_{k+1} - A_k IP_k = \sum_{i=0}^{k} a_i^{(k)} P_i \tag{57}$$

mit gewissen $a_i^{(k)} \in \mathbf{R}$. Für $k > 1$ liefert skalare Multiplikation dieser Gleichung mit P_j ($j < k - 1$) unter Beachtung von (55) und der Orthogonalität des Polynomsystems

$$\sum_{i=0}^{k} a_i^{(k)} (P_i, P_j) = a_j^{(k)} \|P_j\|^2 = -A_k (IP_k, P_j) = -A_k (P_k, IP_j) = 0. \tag{58}$$

IP_j ist nämlich ein Polynom maximal $(k-1)$-ten Grades und als solches in der Form

$$IP_j = \sum_{i=0}^{k-1} b_i P_i$$

darstellbar.

Aus (58) folgt $a_j^{(k)} = 0$ für $j = 0, 1, \ldots, k-2$ und in Verbindung mit (57)

$$P_{k+1} - A_k I P_k = a_k^{(k)} P_k + a_{k-1}^{(k)} P_{k-1}.$$

Setzt man $B_k := a_k^{(k)}$, $C_k := a_{k-1}^{(k)}$, so ergibt sich (56).

Für $k = 1$ folgt unmittelbar aus (57)

$$P_2 = (A_1 I + a_1^{(1)}) P_1 + a_0^{(1)} P_0,$$

also wieder (56).

Bemerkung 5. Von besonderem Interesse sind für uns die mit einem Skalarprodukt ausgestatteten Räume $C_{(a,b)}$ und der Raum aller auf einer endlichen Menge $X = \{x_1, x_2, \ldots, x_m\}$ definierten Funktionen $f: X \to \mathbf{R}$ bezüglich des Skalarproduktes

$$(f, g) = \sum_{j=1}^{m} w_j f(x_j)\, g(x_j) \tag{59}$$

für fest gewählte $w_j \in \mathbf{R}_+^*$ und $f, g \in \{h: h: X \to \mathbf{R},\ X = \{x_1, x_2, \ldots, x_m\}\}$. Offensichtlich ist (55) für die Skalarprodukte (2), (50) und (59) erfüllt.

Als Anwendung des Satzes 8 leiten wir Rekursionsformeln für die Legendreschen und Tschebyscheffschen Polynome her.

Die Bestimmung der Konstanten A_k, B_k, C_k kann zum Beispiel durch einen Koeffizientenvergleich oder durch Einsetzen spezieller Werte erfolgen.

Der Leitkoeffizient $c_n^{(n)}$ des Legendreschen Polynoms P_n ist nach (36)

$$\begin{aligned} c_n^{(n)} &= \frac{1}{2^n n!}\, 2n(2n-1)\cdots(n+1) = \frac{(2n)!}{2^n (n!)^2} \\ &= \frac{(2n-1)\,(2n-3)\cdots 3 \cdot 1}{n!}. \end{aligned} \tag{60}$$

Wendet man die Leibnizsche Differentiationsformel

$$\frac{d^n}{dx^n}(uv) = (uv)^{(n)} = u^{(n)}v + \binom{n}{1} u^{(n-1)}v' + \binom{n}{2} u^{(n-2)}v'' + \cdots + uv^{(n)}$$

(Beweis durch vollständige Induktion) auf das Produkt $(x^2-1)^n = (x-1)^n (x+1)^n$ an, so ergibt sich auf Grund von (36)

$$\begin{aligned} P_n(1) &= \frac{1}{2^n n!}\, 2^n n! = 1, \\ P_n(-1) &= \frac{1}{2^n n!}\, 2^n n!\, (-1)^n = (-1)^n. \end{aligned} \tag{61}$$

Nun vergleichen wir in der Rekursionsformel (56) für die Legendreschen Polynome die Koeffizienten bei x^{k+1}. Man erhält mit Hilfe von (60)

$$\frac{(2k+1)(2k-1)\cdots 3\cdot 1}{(k+1)!} = \frac{(2k-1)(2k-3)\cdots 3\cdot 1}{k!} A_k,$$

also

$$A_k = \frac{2k+1}{k+1}.$$

Für $x = \pm 1$ folgt weiter aus (56) mit Beachtung von (61)

$$1 = \frac{2k+1}{k+1} + B_k + C_k,$$

$$1 = \frac{2k+1}{k+1} - B_k + C_k,$$

d. h.

$$B_k = 0, \qquad C_k = -\frac{k}{k+1}.$$

Die Legendreschen Polynome genügen also der Rekursionsformel

$$(k+1)P_{k+1}(x) - (2k+1)xP_k(x) + kP_{k-1}(x) = 0. \tag{62}$$

Für die Tschebyscheffschen Polynome liefert der Koeffizientenvergleich bei x^{k+1} in (56) auf Grund von Satz 6

$$A_k = 2,$$

und durch Einsetzen der Werte $x = \pm 1$ erhält man

$$1 = 2 + B_k + C_k,$$

$$1 = 2 - B_k + C_k,$$

also

$$B_k = 0, \qquad C_k = -1.$$

Die Rekursionsformel für die T_n lautet

$$T_{k+1}(x) - 2xT_k(x) + T_{k-1}(x) = 0. \tag{63}$$

Bemerkung 6. Da die lineare Hülle eines Systems $\{\varphi_i\}$ linear unabhängiger Elemente des unitären Raumes E mit der des daraus nach E. Schmidt gebildeten Orthogonalsystems $\{\psi_i\}$ übereinstimmt, kann die Konstruktion der besten Quadratmittelapproximation eines Elementes $f \in E$ in Gestalt einer Linearkombination der φ_i, also die Lösung des Problems (5), nach folgendem Algorithmus erfolgen:

1. Orthogonalisierung (Orthonormalisierung) der φ_i $(i = 1, 2, \ldots, n)$.

2. Bestimmung der Fourierkoeffizienten a_i^* von f bezüglich des konstruierten Orthogonalsystems ψ_i $(i = 1, 2, \ldots, n)$.

3. Umordnung von $\sum_{i=1}^{n} a_i^* \psi_i$ in die entsprechende Linearkombination der φ_i.

5.2.4. Polynomapproximation

Die weiteren Betrachtungen dieses Paragraphen befassen sich mit *Algorithmen zur Lösung spezieller Probleme.* Dabei werden stets Orthogonalsysteme zugrunde gelegt bzw. eigens konstruiert, um die gesuchte Bestapproximation durch die Berechnung von Fourierkoeffizienten bestimmen zu können. Dieser Abschnitt ist der Polynomapproximation von Funktionen gewidmet, die auf einer endlichen Menge $X = \{x_1, x_2, \ldots, x_m\}$ paarweise verschiedener $x_i \in \mathbf{R}$ definiert sind. Bezüglich des Skalarproduktes (59) und der üblichen Linearoperationen stellen diese in ihrer Gesamtheit einen unitären Raum dar, den wir (wie auch dessen Trägermenge) mit M_X bezeichnen. Wenn ein Polynom P als Element von M_X angesprochen wird, ist damit die Einschränkung $P|X$ im Sinne von MfL Bd. 4, 1.3.1., gemeint.

Wir untersuchen die optimale Quadratmittelapproximation einer Funktion $f \in M_X$ durch Linearkombinationen der Potenzen

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x, \quad \ldots, \quad \varphi_n(x) = x^n .^1) \tag{64}$$

Zunächst beweisen wir

Hilfssatz 1. *Die Funktionen* (64) *sind — als Elemente des Raumes M_X betrachtet — für $n \leqq m - 1$ linear unabhängig.*

Beweis. Es sei

$$\varphi = \sum_{i=0}^{m-1} \alpha_i \varphi_i$$

eine auf X identisch verschwindende Linearkombination, d. h., es gilt

$$\begin{aligned}
&\alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 + \cdots + \alpha_{m-1} x_1^{m-1} = 0,\\
&\alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \cdots + \alpha_{m-1} x_2^{m-1} = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\alpha_0 + \alpha_1 x_m + \alpha_2 x_m^2 + \cdots + \alpha_{m-1} x_m^{m-1} = 0.
\end{aligned}$$

Das ist ein lineares Gleichungssystem für die α_i, dessen Koeffizientendeterminante die mit den x_j, $j = 1, \ldots, m$, gebildete Vandermondesche Determinante ist. Diese verschwindet nicht, da die x_j paarweise verschieden sind. Mithin ist $\alpha_i = 0$ für $i = 0, 1, \ldots, m - 1$.

---

[1]) Dem Exponenten entsprechend werden die Funktionen (64) im Unterschied zu 5.1.(21) mit Null beginnend indiziert.

Es gibt also auf Grund von Satz 1 in der Menge der Polynome maximal n-ten Grades ($n \leqq m-1$) genau eins, das beste Quadratmittelapproximation einer beliebig vorgegebenen Funktion $f \in M_X$ ist. Seine Koeffizienten können durch Lösen der mit dem Skalarprodukt (59) gebildeten Normalgleichungen bestimmt werden. Wir bevorzugen jedoch den durch Bemerkung 6 vorgezeichneten Lösungsweg und konstruieren ein bezüglich (59) auf der Menge X orthogonales Polynomsystem $P_0, P_1, \ldots, P_n$. Dabei wird zunächst $w_j = 1$, $j = 1, 2, \ldots, m$, gesetzt; es ist dann vorteilhaft, die P_i gemäß

$$(P_i, P_k) = m\delta_{ik}, \qquad 0 \leqq i, k \leqq n \leqq m-1 \tag{65}$$

zu normieren. Die im folgenden beschriebene Methode wurde von R. Ludwig [31] für Zwecke der Ausgleichsrechnung entwickelt.

Wir bezeichnen die Koeffizienten von P_k mit c_{ik}:

$$P_k(x) = \sum_{i=0}^{k} c_{ik} x^i, \qquad k = 0, 1, \ldots, m-1. \tag{66}$$

Aus dem Orthogonalisierungsansatz (15) des E. Schmidtschen Verfahrens folgt unmittelbar, daß die P_k mit einem positiven Leitkoeffizienten konstruiert werden können. Ergänzen wir (65) durch diese Forderung, so erweisen sich diese Polynome als eindeutig bestimmt. Man findet zunächst

$$mc_{00}^2 = m, \quad \text{also} \quad P_0(x) = c_{00} = 1. \tag{67}$$

Weiter ist

$$(P_0, P_1) = c_{00} \sum_{j=1}^{m} (c_{01} + c_{11} x_j) = c_{00} \left(m c_{01} + c_{11} \sum_{j=1}^{m} x_j \right) = 0,$$

d. h.

$$m c_{01} + c_{11} \sum_{j=1}^{m} x_j = 0.$$

Damit folgt nach Multiplikation von

$$(P_1, P_1) = m c_{01}^2 + c_{11}^2 \sum_{j=1}^{m} x_j^2 + 2 c_{01} c_{11} \sum_{j=1}^{m} x_j = m$$

mit m

$$c_{11}^2 \left[\left(\sum_{j=1}^{m} x_j \right)^2 + m \sum_{j=1}^{m} x_j^2 - 2 \left(\sum_{j=1}^{m} x_j \right)^2 \right] = c_{11}^2 \left[m \sum_{j=1}^{m} x_j^2 - \left(\sum_{j=1}^{m} x_j \right)^2 \right] = m^2.$$

Mit der Abkürzung

$$N_1 := \sqrt{m \sum_{j=1}^{m} x_j^2 - \left(\sum_{j=1}^{m} x_j \right)^2} \tag{68}$$

erhält man nun

$$c_{11} = \frac{m}{N_1}, \qquad c_{01} = -\frac{\sum_{j=1}^{m} x_j}{N_1}. \tag{69}$$

Die Konstruktion der weiteren Polynome erfolgt nicht nach dem E. Schmidtschen Verfahren, sondern mit Hilfe von P_0 und P_1 auf Grund der Rekursionsformel (56). Skalare Multiplikation der Gleichung

$$P_k(x) = (A_{k-1}x + B_{k-1})\,P_{k-1}(x) + C_{k-1}P_{k-2}(x)$$

mit P_{k-1} und P_{k-2} liefert unter Beachtung von (65)

$$\begin{aligned} 0 &= A_{k-1}\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2 + B_{k-1}m, \\ 0 &= A_{k-1}\sum_{j=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j) + C_{k-1}m. \end{aligned} \tag{70}$$

Multipliziert man beide Seiten der Rekursionsformel skalar mit sich selbst, so resultiert

$$\begin{aligned} m = {} & A_{k-1}^2\sum_{j=1}^{m} x_j^2 P_{k-1}(x_j)^2 + 2A_{k-1}B_{k-1}\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2 \\ & + (B_{k-1}^2 + C_{k-1}^2)\,m + 2A_{k-1}C_{k-1}\sum_{j=1}^{m} x_k P_{k-1}(x_j)\,P_{k-2}(x_j). \end{aligned}$$

In dieser Gleichung werden B_{k-1} und C_{k-1} mit Hilfe von (70) eliminiert:

$$\begin{aligned} m^2 = {} & A_{k-1}^2\left\{m\sum_{j=1}^{m} x_j^2 P_{k-1}(x_j)^2 - 2\left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2\right]^2 + \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2\right]^2\right. \\ & \left. + \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j)\right]^2 - 2\left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j)\right]^2\right\} \\ = {} & A_{k-1}^2\left\{m\sum_{j=1}^{m} x_j^2 P_{k-1}(x_j)^2 - \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2\right]^2\right. \\ & \left. - \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j)\right]^2\right\}. \end{aligned}$$

Schließlich gewinnt man mit der Abkürzung

$$N_k := \sqrt{m\sum_{j=1}^{m} x_j^2 P_{k-1}(x_j)^2 - \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)^2\right]^2 - \left[\sum_{j=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j)\right]^2} \tag{71}$$

die Beziehungen

$$\begin{aligned} A_{k-1} &= \frac{m}{N_k}, \qquad B_{k-1} = -\frac{\sum_{i=1}^{m} x_j P_{k-1}(x_j)^2}{N_k}, \\ C_{k-1} &= -\frac{\sum_{i=1}^{m} x_j P_{k-1}(x_j)\,P_{k-2}(x_j)}{N_k}. \end{aligned} \tag{72}$$

Die beste Quadratmittelapproximation einer Funktion $f \in M_X$ durch ein Polynom vom Grade n ($n \leqq m-1$) oder — gleichbedeutend damit — durch eine Linearkombination

$$F_n = \sum_{k=0}^{n} a_k P_k, \qquad n \leqq m-1,$$

der mit Hilfe der Rekursionsformel und (72) konstruierten Orthogonalpolynome ergibt sich, wenn diese mit den Fourierkoeffizienten

$$a_i^* = \frac{\sum_{j=1}^{m} f(x_j)\, P_i(x_j)}{m}, \qquad i = 1, 2, \ldots, n, \tag{73}$$

gebildet wird. Die Approximationsgüte ist nach (20) und (65) durch

$$\|f - F_{n^*}\|^2 = \sum_{j=1}^{m} f(x_j)^2 - m \sum_{i=1}^{n} a_i^{*2} \tag{74}$$

bestimmt.

Um die mit den Koeffizienten (73) gebildeten Polynome F_n für ein Argument $x \neq x_j$ ($j = 1, 2, \ldots, m$) zu berechnen, wird man nach dem Hornerschen Schema verfahren und dazu die Linearkombination der P_k nach Potenzen von x oder einer anderen geeignet erscheinenden Entwicklungsgröße $x - x_0$ umordnen:

$$F_n(x) = \sum_{k=0}^{n} a_k P_k(x) = \sum_{i=0}^{n} C[i]\, x^i. \tag{75}$$

Zu diesem Zweck bestimmen wir zunächst mit Hilfe der Formel (56) rekursiv die Koeffizienten in (66). Für $k \geqq 2$ erhält man

$$\begin{aligned}
c_{0k} &= B_{k-1}c_{0,k-1} + C_{k-1}c_{0,k-2},\\
c_{1k} &= A_{k-1}c_{0,k-1} + B_{k-1}c_{1,k-1} + C_{k-1}c_{1,k-2},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
c_{ik} &= A_{k-1}c_{i-1,k-1} + B_{k-1}c_{i,k-1} + C_{k-1}c_{i,k-2},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
c_{k-2,k} &= A_{k-1}c_{k-3,k-1} + B_{k-1}c_{k-2,k-1} + C_{k-1}c_{k-2,k-2},\\
c_{k-1,k} &= A_{k-1}c_{k-2,k-1} + B_{k-1}c_{k-1,k-1},\\
c_{kk} &= A_{k-1}c_{k-1,k-1}.
\end{aligned} \tag{76}$$

Mit

$$\boldsymbol{c}_k = (c_{0k}, c_{1k}, \ldots, c_{kk})^{\mathsf{T}}, \qquad \boldsymbol{R}_k = (A_{k-1}, B_{k-1}, C_{k-1})^{\mathsf{T}} \tag{77}$$

läßt sich (76) als Matrizengleichung

$$\boldsymbol{c}_k = \begin{pmatrix} 0 & c_{k-1} & c_{k-2} \\ c_{k-1} & 0 & 0 \end{pmatrix} \boldsymbol{R}_k \tag{78}$$

schreiben. Aus $F_a = \sum_{k=0}^{n} a_k P_k$ folgt

$$C[i] = \sum_{k=i}^{n} c_{ik} a_k, \tag{79}$$

was nach Einführung der oberen $(n+1, n+1)$-Dreiecksmatrix

$$\boldsymbol{D} = \begin{pmatrix} c_0 & c_1 & \dots & c_n \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \ddots \end{pmatrix} = \begin{pmatrix} c_{00} & c_{01} & c_{02} & \dots & c_{0n} \\ 0 & c_{11} & c_{12} & \dots & c_{1n} \\ 0 & 0 & c_{22} & \dots & c_{2n} \\ & & & \ddots & \\ 0 & 0 & 0 & & c_{nn} \end{pmatrix} \tag{80}$$

in der Form

$$\boldsymbol{C} = \boldsymbol{D}\boldsymbol{a} \tag{81}$$

ausgedrückt werden kann, wenn

$$\boldsymbol{C} = (C[0], C[1], \dots, C[n])^{\mathrm{T}} \quad \text{und} \quad \boldsymbol{a} = (a_0, a_1, \dots, a_n)^{\mathrm{T}}$$

ist.

Im folgenden wird eine ALGOL-Prozedur *ORTPOL* beschrieben, welche bei gegebenen $y_j = f(x_j)$, $j = 1, 2, \dots, m$, die mit den Koeffizienten (73) gebildete beste Polynomapproximation F_{a^*} bestimmt, die Werte $F_{a^*}(x_j)$ und die Fehlergröße (74) berechnet und $F_{a^*} = \sum_{k=0}^{n} a_k^* P_k$ nach Potenzen einer vorgegebenen Entwicklungsgröße $x - x_0$ ordnet. Als formale Parameter treten auf:

$m, n, x0$:	diese Größen haben die oben eingeführte Bedeutung und werden in den Werteteil übernommen;
z:	reelle Variable, welcher der Wert $\|f - F_{a^*}\|^2$ zugeordnet wird;
nz:	Index $\leqq n$, bei dem die Berechnung der Orthogonalpolynome abgebrochen wird, sofern die Fehlergröße z auf Grund von Rundungsfehlern negativ wird;
X, Y, FA:	Felder zur Aufnahme der x_j, $y_j = f(x_j)$ und $F_{a^*}(x_j)$, $j = 1, 2, \dots, m$;
C:	Feld zur Speicherung der Koeffizienten des nach Potenzen von $x - x_0$ geordneten Polynoms der Bestapproximation von f.

Neben Hilfsgrößen werden lokal vereinbart:

r_1, r_2, r_3	reelle Variable in der Bedeutung der nach (72) zu berechnenden Koeffizienten A_{k-1}, B_{k-1}, C_{k-1} der Rekursionsformel (56);
$A[0:n]$	als Feld zur Speicherung der Fourierkoeffizienten (73);

$D[0:n, 0:n]$ } als Felder mit der Bedeutung des ersten Matrixfaktors der rechten
$CC[0:n, 1:3]$ } Seite von (81) bzw. (78);

$P[0:n, 1:m]$ als Feld zur Speicherung der $P_k(x_j)$, $k = 0, 1, \ldots, n$, $j = 1, 2, \ldots, m$.

Weitere Erklärungen zur Prozedur *ORTPOL* sind im ALGOL-Text enthalten. Das Einfügen solcher Erläuterungen hat nach bestimmten Regeln zu erfolgen und wird dann bei der Übersetzung eines Programms übergangen. Danach ist speziell folgender Kommentar zulässig:

Nach einem Semikolon erscheint das Grundsymbol **comment** gefolgt von ALGOL-Zeichen, deren letztes ein Semikolon ist, während im übrigen kein Semikolon in der Zeichenreihe auftritt. Mit Erläuterungen dieser Art wird in der Prozedur *ORTPOL* der Inhalt eines nachfolgenden Textstückes beschrieben:

```
procedure ORTPOL(m,n,x0,z,nz,X,Y,FA,C);
          value m,n,x0; integer m,n,nz; real x0,z;
          array X,Y,FA,C;
begin     integer i,j; real s1,s2,nen,r1,r2,r3;
          array A[0:n], CC[0:n,1:3], P[0:n,1:m], D[0:n,0:n];
comment   x0 wird Ursprung der Abszissen;
          if x0 ≠ 0 then for j := 1 step 1 until m do
          X[j] := X[j] − x0;
comment   Bestimmung von P0 und P1 gemäß (66) bis (69);
          s1 := s2 := z := A[0] := A[1] := 0;
          for j := 1 step 1 until m do begin
          s1 := s1 + X[j]; s2 := s2 + X[j] × X[j];
          z := z + Y[j] × Y[j]          end;
          D[0,0] := 1; nen := sqrt(m × s2 − s1 × s1);
          D[0,1] := −s1/nen; D[1,1] := m/nen;
comment   Berechnung der Fourierkoeffizienzen A[0], A[1];
          for j := 1 step 1 until m do                    begin
          P[0,j] := D[0,0]; P[1,j] := D[0,1] + D[1,1] × X[j];
          A[0] := A[0] + Y[j]; A[1] := A[1] + Y[j] × P[1,j] end;
          A[0] := A[0]/m; A[1] := A[1]/m;
comment   anteilige Berechnung der Fehlergröße z, der Werte FA[j] und von C[0],
          C[1] gemäß (81);
          z := z − (A[0] × A[0] + A[1] × A[1]) × m;
          for j := 1 step 1 until m do
          FA[j] := A[0] + A[1] × P[1,j];
          C[0] := A[0] + D[0,1] × A[1]; C[1] := D[1,1] × A[1];
comment   Bestimmung der höheren Orthogonalpolynome, Fourierkoeffizienten und
          C[i];
          if n ≠ 1 then begin for i := 2 step 1 until n do
          begin A[i] := C[i] := r1 := r2 := r3 := 0;
```

```
comment Berechnung der Koeffizienten in der Rekursionsformel (56) gemäß (72);
        for j := 1 step 1 until m do                    begin
        s1 := X[j] × P[i − 1, j]; s2 := s1 × P[i − 2, j];
        s1 := s1 × P[i − 1, j]; r1 := r1 + X[j] × s1;
        r2 := r2 + s1; r3 := r3 + s2;                   end
        nen := sqrt(m × r1 − r2 × r2 − r3 × r3);
        r1 := m/nen; r2 := −r2/nen; r3 := −r3/nen;

comment Berechnung der Werte des i-ten Orthogonalpolynoms an den Stellen x_j
        nach (56) und des Fourierkoeffizienten A[i] nach (73);
        for j := 1 step 1 until m do                    begin
        P[i,j] := (r1 × X[j] + r2) × P[i − 1, j] + r3 × P[i − 2, j];
        A[i] := A[i] + Y[j] × P[i, j]                   end;
        A[i] := A[i]/m;

comment Fortsetzung der Berechnung von FA[j];
        for j := 1 step 1 until m do
        FA[j] := FA[j] + A[i] × P[i,j];

comment anteilige Berechnung der C[j] für j von 0 bis i nach (78) und (81);
        for j := 0 step 1 until i − 1 do                begin
        CC[j + 1, 1] := CC[j,2] := D[j, i − 1];
        CC[j,3] := D[j, i − 2];                         end
        CC[0,1] := CC[i,2] := CC[i,3] := CC[i − 1, 3] := 0;
        for j := 0 step 1 until i do                    begin
        D[j,i] := r1 × CC[j,1] + r2 × CC[j,2] + r3 × CC[j,3];
        C[j] := C[j] + A[i] × D[j,i]                    end
        z := z − m × A[i] × A[i];
        if z < 0 then begin z := 0; nz := i; goto L end
                      end;
        end;
L: end
```

Bemerkung 7. Eine Transformation des Koordinatenursprungs empfiehlt sich, wenn der Mittelwert der x_j nicht in der Nähe von Null liegt. Sonst kann in diesem Fall — z. B. bei der Auswertung der arithmetischen Ausdrücke (72) — ein Verlust an wesentlichen Ziffern eintreten.

Bemerkung 8. Die im Anschluß an (65) dargelegte Konstruktion von Orthogonalpolynomen läßt sich mit geringfügigen Modifikationen auf den Fall eines beliebig gewichteten Skalarproduktes (59) übertragen. Es ist dann vorteilhaft, (65) durch die Normierung

$$(P_i, P_k) = W\delta_{ik}, \qquad 0 \leqq i, k \leqq n \leqq m - 1$$

mit

$$W := \sum_{j=1}^{m} w_j \tag{82}$$

zu ersetzen. Die Durchführung im einzelnen sei dem Leser als Übungsaufgabe überlassen.

Wenn die x_j äquidistant sind und

$$h := x_{j+1} - x_j,$$

gewinnt man mit Hilfe der linearen Transformation

$$t = \frac{x - x_1}{h}$$

an Stelle von X die Argumentmenge $T = \{0, 1, \ldots, m-1\}$. Für diese Normierung des Problems hat erstmalig TSCHEBYSCHEFF explizit Orthogonalpolynome bestimmt, die auch in modifizierter Form besonders für Anwendungen in der Statistik tabuliert wurden. Für weiterführende Studien sei der Leser auf die Literaturhinweise zu [31], Kap. 5, und auf [10], Kap. I, verwiesen.

Wir beschließen diesen Abschnitt mit einem Beispiel aus [10], das mit der Prozedur *ORTPOL* gerechnet wurde.

j	x_j	f_j	δ_j	
1	0,000	1,300	−0,003	$m = 10$
2	0,3	1,245	0,009	
3	0,6	1,095	−0,003	
4	0,9	0,855	−0,012	
5	1,2	0,514	0,000	
6	1,5	0,037	0,020	
7	1,8	−0,600	0,002	
8	2,1	−1,295	−0,033	
9	2,4	−1,767	0,025	
10	2,7	−1,914	−0,006	

Tabelle 5.1

Beispiel. Die mit Tabelle 5.1 gegebene Funktion soll durch ein Polynom fünften Grades approximiert werden. *ORTPOL* liefert für $x_0 = 1{,}35$ die Minimallösung

$$F_{a^*}(x) = 0{,}283 - 1{,}657(x - x_0) - 0{,}797(x - x_0)^2 + 0{,}100(x - x_0)^3 + 0{,}261(x - x_0)^4 + 0{,}086(x - x_0)^5$$

mit dem Normfehler

$$\|f - F_{a^*}\|^2 = 0{,}0024.$$

Die Abweichungen

$$\delta_j := f_j - F_{a^*}(x_j), \qquad j = 1(1)10,$$

findet man in der letzten Spalte der Tabelle 5.1. Abb. 5.7 zeigt den Graphen der Funktion F_{a^*} über dem Intervall $[\![0; 2{,}7]\!]$ und die Punkte (x_j, f_j), $j = 1(1)10$.

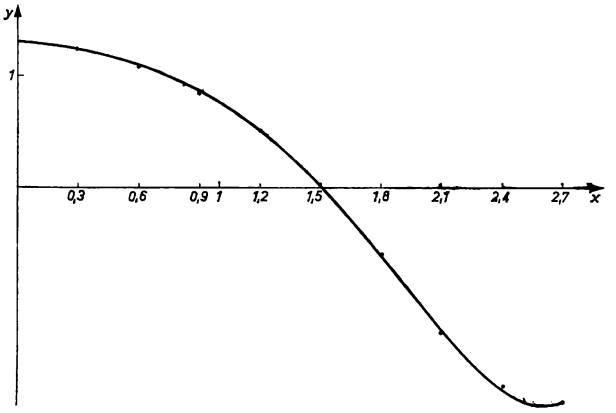


Abb. 5.7

5.2.5. Angenäherte Harmonische Analyse

Wir erörtern folgende Diskretisierung des dem Satz 3 zugrunde liegenden Sachverhaltes:

$$X = \{x_1, x_2, \ldots, x_m\}$$

sei die Menge der äquidistanten Teilpunkte

$$x_j = \frac{p}{m} j, \qquad j = 1, 2, \ldots, m, \tag{83}$$

des Intervalls $[\![0, p]\!]$, $p > 0$, und M_X die Gesamtheit der Funktionen

$$f: X \to \mathbf{R}.$$

M_X wird als unitärer Raum bezüglich des Skalarproduktes (59) mit $w_j = 1, j = 1(1)m$, betrachtet. Dann gilt

Satz 9. *Die auf X eingeschränkten Funktionen*

$$\begin{aligned} &\varphi_0: x \mapsto \frac{1}{2}, \qquad \varphi_1: x \mapsto \sin\frac{2\pi}{p}x, \qquad \varphi_2: x \mapsto \cos\frac{2\pi}{p}x, \\ &\varphi_3: x \mapsto \sin\frac{4\pi}{p}x, \qquad \varphi_4: x \mapsto \cos\frac{4\pi}{p}x, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\varphi_{2n-1}: x \mapsto \sin\frac{2\pi}{p}nx, \qquad \varphi_{2n}: x \mapsto \cos\frac{2\pi}{p}nx \end{aligned} \tag{84}$$

bilden unter der Voraussetzung

$$2n + 1 \leqq m \tag{85}$$

ein Orthogonalsystem mit

$$(\varphi_i, \varphi_j) = 0 \qquad \text{für } i \neq j,$$

$$(\varphi_i, \varphi_i) = \begin{cases} \dfrac{m}{4} & \text{für } i = 0, \\ \dfrac{m}{2} & \text{für } i \neq 0. \end{cases} \tag{86}$$

Die Fourierkoeffizienten a_k^*, b_k^* einer Funktion $f \in M_X$ zu φ_{2k} bzw. φ_{2k-1} sind

$$\begin{aligned} a_k^* &:= \frac{(f, \varphi_{2k})}{\|\varphi_{2k}\|^2} = \frac{2}{m}\sum_{j=1}^{m} f(x_j)\cos\frac{2\pi kj}{m}, \qquad k = 0, 1, 2, \ldots, \\ b_k^* &:= \frac{(f, \varphi_{2k-1})}{\|\varphi_{2k-1}\|^2} = \frac{2}{m}\sum_{j=1}^{m} f(x_j)\sin\frac{2\pi kj}{m}, \qquad k = 1, 2, \ldots. \end{aligned} \tag{87}$$

Unter allen Linearkombinationen der Funktionen (84) *ist*

$$F = \frac{a_0^*}{2} + \sum_{\nu=1}^{n}\left(a_\nu^* \cos\frac{2\pi\nu x}{p} + b_\nu^* \sin\frac{2\pi\nu x}{p}\right) \tag{88}$$

die beste Quadratmittelapproximation von f.

Beweis. Es genügt, die Orthogonalitätsrelationen (86) zu beweisen. Dazu werden die trigonometrischen Summen

$$\sum_{j=1}^{m}\cos\frac{2\pi h}{p}x_j = \sum_{j=1}^{m}\cos\frac{2\pi hj}{m},$$

$$\sum_{j=1}^{m}\sin\frac{2\pi h}{p}x_j = \sum_{j=1}^{m}\sin\frac{2\pi hj}{m},$$

$h = 1(1)2n$, berechnet. Mit Beachtung der Eulerschen Formel und der Summen-

formel für die geometrische Reihe ergibt sich

$$\sum_{j=1}^{m} \cos\frac{2\pi hx_j}{p} + i\sum_{j=1}^{m} \sin\frac{2\pi hx_i}{p} = \sum_{j=1}^{m} e^{i\frac{2\pi hx_j}{p}} = \sum_{j=1}^{m} e^{i\frac{2\pi hj}{m}}$$

$$= \frac{e^{i2\pi h} - 1}{e^{i\frac{2\pi h}{m}} - 1}\, e^{i\frac{2\pi h}{m}} = 0,$$

also

$$\begin{aligned} \sum_{j=1}^{m} \cos\frac{2\pi hx_j}{p} &= \sum_{j=1}^{m} \cos\frac{2\pi hj}{m} = 0, \\ \sum_{j=1}^{m} \sin\frac{2\pi hx_j}{p} &= \sum_{j=1}^{m} \sin\frac{2\pi hj}{m} = 0, \end{aligned} \tag{89}$$

$h = 1(1)2n$. Es sei bemerkt, daß wegen (85) $0 < \frac{h}{m} < 1$ und daher $e^{i\frac{2\pi h}{m}} \neq 1$ ist.

Auf Grund der beim Beweis von (30) benutzten trigonometrischen Formeln hat man

$$\begin{aligned} \sum_{j=1}^{m} \cos\frac{2\pi kx_j}{p} \sin\frac{2\pi lx_j}{p} &= \frac{1}{2}\sum_{j=1}^{m} \sin\frac{2\pi(k+l)\,x_j}{p} - \frac{1}{2}\sum_{j=1}^{m} \sin\frac{2\pi(k-l)\,x_j}{p}, \\ \sum_{j=1}^{m} \cos\frac{2\pi kx_j}{p} \cos\frac{2\pi lx_j}{p} &= \frac{1}{2}\sum_{j=1}^{m} \cos\frac{2\pi(k+l)\,x_j}{p} + \frac{1}{2}\sum_{j=1}^{m} \cos\frac{2\pi(k-l)\,x_j}{p}, \\ \sum_{j=1}^{m} \sin\frac{2\pi kx_j}{p} \sin\frac{2\pi lx_j}{p} &= \frac{1}{2}\sum_{j=1}^{m} \cos\frac{2\pi(k-l)\,x_j}{p} - \frac{1}{2}\sum_{j=1}^{m} \cos\frac{2\pi(k+l)\,x_j}{p}. \end{aligned} \tag{90}$$

Wegen (89) ergibt sich für $0 \leqq k,\ l \leqq n$ und $k \neq l$ auf der rechten Seite von (90) jedesmal Null, in der ersten Gleichung auch dann noch, wenn $k = l$. Falls $k = l \neq 0$ ist, verschwindet in den letzten beiden Gleichungen (90) je eine Summe auf der rechten Seite, während die andere den Wert m hat. Für $k = l = 0$ sind alle Kosinussummen auf der rechten Seite gleich m. Damit erhält man als Analogon zu (30) für $k, l \in \mathbf{N}$, $0 \leqq k, l \leqq n$:

$$\sum_{j=1}^{m} \cos\frac{2\pi kx_j}{p} \sin\frac{2\pi lx_j}{p} = 0,$$

$$\sum_{j=1}^{m} \cos\frac{2\pi kx_j}{p} \cos\frac{2\pi lx_j}{p} = \begin{cases} 0 & \text{für } k \neq l, \\ \frac{m}{2} & \text{für } k = l \neq 0, \\ m & \text{für } k = l = 0, \end{cases} \tag{91}$$

$$\sum_{j=1}^{m} \sin\frac{2\pi kx_j}{p} \sin\frac{2\pi lx_j}{p} = \begin{cases} 0 & \text{für } k \neq l, \\ \frac{m}{2} & \text{für } k = l \neq 0, \\ 0 & \text{für } k = l = 0, \end{cases}$$

woraus (86) folgt.

Für den bei der Approximation von f durch das trigonometrische Polynom (88) entstehenden Fehler berechnet man nach (20) und (86)

$$\|f - F\|^2 = \|f\|^2 - \frac{m}{2}\left[\frac{a_0^{*2}}{2} + \sum_{k=1}^{n}(a_k^{*2} + b_k^{*2})\right]. \tag{92}$$

Diesbezüglich heben wir den Spezialfall von (85)

$$2n + 1 = m \tag{93}$$

hervor. Hier gilt

Satz 10. *Unter der Voraussetzung* (93) *ist* $\|f - F\| = 0$, *d. h.*, (88) *löst die Aufgabe der trigonometrischen Interpolation*

$$f(x_j) = F(x_j) = \frac{a_0^*}{2} + \sum_{r=1}^{n}\left(a_r^* \cos\frac{2\pi r x_j}{p} + b_r^* \sin\frac{2\pi r x_j}{p}\right), \tag{94}$$
$$j = 1(1)m,$$

an den Stellen (83). *Sind umgekehrt die Gleichungen* (94) *für eine Linearkombination F der Funktionen* (84) *mit Koeffizienten a_r, b_r erfüllt, so müssen diese sämtlich mit den entsprechenden Fourierkoeffizienten* (87) *übereinstimmen.*

Beweis. Der zweite Teil des Satzes 10 folgt aus der im Zusammenhang mit (21) hervorgehobenen Minimaleigenschaft der Fourierkoeffizienten, da bei Erfülltsein von (94) $\|f - F\|^2$ verschwindet und die Abweichungsfunktion Z für das Koeffizientensystem a_r, b_r ihr absolutes Minimum annimmt.

Die Umkehrung wird durch Berechnung der Fehlergröße $\|f - F\|^2$ verifiziert. Dazu fassen wir in (92) zunächst die rein quadratischen Glieder zusammen, die sich aus (87) bei der Bildung von $a_k^{*2} + b_k^{*2}$, $k = 1(1)n$, und $\frac{a_0^{*2}}{2}$ ergeben. Nach Multiplikation mit $\frac{m}{2}$ findet man dafür unter Beachtung von (93) die Größe

$$\frac{m}{2}\left[\frac{2}{m^2}\sum_{j=1}^{m} f(x_j)^2 + n\frac{4}{m^2}\sum_{j=1}^{m} f(x_j)^2\right] = \frac{2n+1}{m}\sum_{j=1}^{m} f(x_j)^2 = \sum_{j=1}^{m} f(x_j)^2 = \|f\|^2.$$

Auf Grund von (92) ist also nur noch das Verschwinden der Gesamtheit aller gemischten Glieder bei der Berechnung von

$$\frac{a_0^{*2}}{2} + \sum_{k=1}^{n}(a_k^{*2} + b_k^{*2})$$

zu zeigen. Für zwei verschiedene Indizes i, j mit $1 \leqq i, j \leqq m = 2n + 1$ ergibt das entsprechende gemischte Glied von a_k^{*2} und b_k^{*2} zusammen

$$8\frac{f(x_i)\,f(x_j)}{(2n+1)^2}\left(\cos\frac{2\pi k i}{2n+1}\cos\frac{2\pi k j}{2n+1} + \sin\frac{2\pi k i}{2n+1}\sin\frac{2\pi k j}{2n+1}\right)$$
$$= 8\frac{f(x_i)\,f(x_j)}{(2n+1)^2}\cos\frac{2\pi k(i-j)}{2n+1}.$$

Auf Grund der durch vollständige Induktion leicht zu beweisenden Formel

$$\sum_{k=0}^{n} \cos kx = \frac{\sin \frac{n+1}{2} x \cos \frac{n}{2} x}{\sin \frac{x}{2}} \tag{95}$$

liefert die Summation dieser Terme von $k = 1$ bis $k = n$ mit $x := \frac{2\pi(i-j)}{2n+1}$

$$8 \frac{f(x_i) f(x_j)}{(2n+1)^2} \left[\frac{\sin \frac{(n+1)\pi(i-j)}{2n+1} \cos \frac{n\pi(i-j)}{2n+1}}{\sin \frac{\pi(i-j)}{2n+1}} - 1 \right]$$

$$= 8 \frac{f(x_i) f(x_j)}{(2n+1)^2} \left[\frac{\sin \pi(i-j) + \sin \frac{\pi(i-j)}{2n+1}}{2 \sin \frac{\pi(i-j)}{2n+1}} - 1 \right] = -\frac{4 f(x_i) f(x_j)}{(2n+1)^2}.$$

Das noch nicht berücksichtigte gemischte Glied zum Indexpaar i, j von $\frac{a_0^{*2}}{2}$ hat aber gerade den Wert $+\frac{4 f(x_i) f(x_j)}{(2n+1)^2}$, so daß sich die gemischten Glieder insgesamt tatsächlich annullieren.

Die weiteren Betrachtungen konzentrieren sich auf algorithmische Fragen im Zusammenhang mit der Berechnung der Fourierkoeffizienten. Die Ausdrücke (87) sind trigonometrische Summen der Form

$$C = \sum_{j=0}^{l} c_j \cos jx, \qquad S = \sum_{j=1}^{l} c_j \sin jx \tag{96}$$

mit gegebenen Werten für c_j und x. Man beachte, daß sich in (87) die Summanden für $j = m$ auch für $j = 0$ ergeben, wenn man

$$x_0 = 0 \quad \text{und} \quad f(x_0) = f(x_m)$$

setzt. Seit langem ist ein Verfahren von C. Runge in Gebrauch, das in Verbindung mit Hilfsmitteln wie Schablonen und Formularen zur Berechnung der Summen (96) von Hand entwickelt wurde. Dieses eignet sich auch gut für den Computereinsatz (vgl. etwa [31]), ist aber hinsichtlich der Rechenzeit nicht so günstig wie der folgende Algorithmus von G. Goertzel [17], der bei der Bestimmung aller Koeffizienten (87) mit einem zweimaligen Aufruf von Prozeduren zur Berechnung eines Kosinus- und Sinuswertes auskommt. Das Verfahren wird wesentlich durch den folgenden Satz beschrieben.

Satz 11. *Bezüglich der Summen* (96) *seien* $U_\varkappa$ *die rekursiv durch*

$$\begin{aligned} &U_{l+2} := U_{l+1} := 0, \\ &U_\varkappa := c_\varkappa + 2U_{\varkappa+1}\cos x - U_{\varkappa+2}, \qquad \varkappa = l(-1)1, \end{aligned} \tag{97}$$

bestimmten Größen. Dann ist

$$C = c_0 + U_1 \cos x - U_2, \tag{98}$$

$$S = U_1 \sin x. \tag{99}$$

Beweis. Wir setzen

$$\begin{aligned} &V_\varkappa := \sum_{j=\varkappa}^{l} c_j \sin(j - \varkappa + 1)\, x, \qquad \varkappa = 1(1)l, \\ &V_{l+1} := V_{l+2} := 0. \end{aligned} \tag{100}$$

Dafür gilt

$$\begin{aligned} &c_\varkappa \sin x + 2V_{\varkappa+1}\cos x - V_{\varkappa+2} \\ &= c_\varkappa \sin x + \sum_{j=\varkappa+1}^{l} c_j[2\cos x \sin(j-\varkappa)\, x - \sin(j-\varkappa-1)\, x] \\ &= c_\varkappa \sin x + \sum_{j=\varkappa+1}^{l} c_j \sin(j-\varkappa+1)\, x = V_\varkappa, \end{aligned} \tag{101}$$

letzteres wegen

$$\cos u \sin v = \frac{1}{2}\sin(u+v) - \frac{1}{2}\sin(u-v).$$

Bei der Herleitung von (101) ist zunächst $1 \leqq \varkappa \leqq l-2$ anzunehmen. Die Formel gilt aber auch für $\varkappa = l$ und $\varkappa = l-1$, denn es ist

$$c_l \sin x = V_l$$

und

$$\begin{aligned} c_{l-1}\sin x + 2V_l \cos x &= c_{l-1}\sin x + 2c_l \sin x \cos x \\ &= c_{l-1}\sin x + c_l \sin 2x = V_{l-1}. \end{aligned}$$

Mit Hilfe von (101) wird induktiv gezeigt, daß

$$V_\varkappa = U_\varkappa \sin x, \qquad \varkappa = 1(1)l, \tag{102}$$

ist. Für $\varkappa = l$ und $\varkappa = l-1$ folgt (102) nacheinander aus (101), (100) und (97). Nehmen wir nun an, daß (102) für $\varkappa \geqq k > 1$ gilt, dann ergibt sich auf Grund von (101), der Induktionsannahme und (97)

$$\begin{aligned} V_{k-1} &= c_{k-1}\sin x + 2V_k \cos x - V_{k+1} \\ &= (c_{k-1} + 2U_k \cos x - U_{k+1})\sin x = U_{k-1}\sin x, \end{aligned}$$

was zu beweisen war. Speziell ist

$$V_1 = S = U_1 \sin x.$$

Zum Beweis von (98) zeigen wir zunächst

$$\begin{aligned} & c_0 \sin x + V_1 \cos x - V_2 \\ & = c_0 \sin x + \sum_{j=1}^{l} c_j [\cos x \sin jx - \sin (j-1)x] \\ & = c_0 \sin x + \sum_{j=1}^{l} c_j \cos jx \sin x = C \sin x. \end{aligned} \tag{103}$$

Mit (102) folgt aus (103)

$$C \sin x = (c_0 + U_1 \cos x - U_2) \sin x$$

und nach Division durch $\sin x$ die Gleichung (98). Aus Stetigkeitsgründen gilt (98) auch für Argumente x, deren Sinus verschwindet.

Speziell findet man nach Satz 11 für die Fourierkoeffizienten (87) bei Beachtung von (85) mit $l = m - 1$

$$\begin{aligned} a_k^* &= \frac{2}{m} \left(f(x_0) + U_1^{(k)} \cos \frac{2\pi k}{m} - U_2^{(k)} \right) \\ b_k^* &= \frac{2}{m} U_1^{(k)} \sin \frac{2\pi k}{m}, \end{aligned} \tag{104}$$

wobei

$$\begin{aligned} & U_{m+1}^{(k)} := U_m^{(k)} := 0, \\ & U_x^{(k)} := f(x_x) + 2U_{x+1}^{(k)} \cos \frac{2\pi k}{m} - U_{x+2}^{(k)}, \\ & x = m-1(-1)1, \qquad k = 0(1)n. \end{aligned} \tag{105}$$

Wir zeigen noch, daß zur Ermittlung sämtlicher a_k^*, b_k^* nur ein Sinus- und ein Kosinuswert berechnet werden müssen. In der Tat gilt für die Größen

$$\varrho_p := \cos p\vartheta, \qquad \tau_p := \sin p\vartheta$$

auf Grund der trigonometrischen Additionstheoreme die Matrizengleichung

$$\begin{pmatrix} \varrho_{p+1} \\ \tau_{p+1} \end{pmatrix} = \begin{pmatrix} \varrho_1 & -\tau_1 \\ \tau_1 & \varrho_1 \end{pmatrix} \begin{pmatrix} \varrho_p \\ \tau_p \end{pmatrix}, \tag{106}$$

so daß die in (104) und (105) auftretenden Größen $\cos \frac{2\pi k}{m}$ und $\sin \frac{2\pi k}{m}$ nach (106) sämtlich aus $\cos \frac{2\pi}{m}$ und $\sin \frac{2\pi}{m}$ bestimmt werden können.

Der PAP in Abb. 5.8 stellt die Berechnung der Fourierkoeffizienten (87) nach (104) dar; anschließend wird diese in einer ALGOL-Prozedur *HARMON* zusammengefaßt. Im PAP sind die Strukturen von zwei ineinandergeschachtelten Laufanweisungen erkennbar, deren innere die Berechnung der zur Bestimmung von a_k^*, b_k^* benötigten Größen $U_1^{(k)}$, $U_2^{(k)}$ betrifft. Zur Speicherung der Fourierkoeffizienten und Funktionswerte $f(x_j)$ dienen die Felder $A[0{:}n]$, $B[0{:}n]$ bzw. $F[0{:}m-1]$.

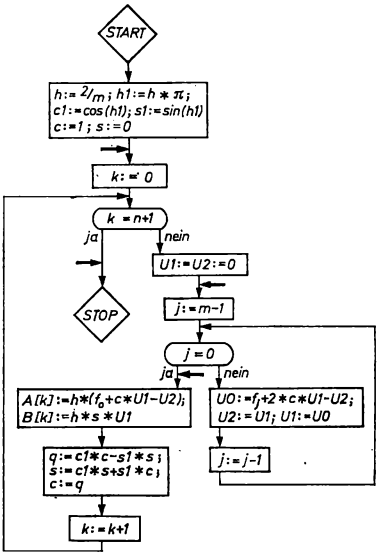


Abb. 5.8

Im k-Lauf werden die Formeln (104) ausgewertet und der für den nächsten Schritt dazu benötigte Kosinus- und Sinuswert bereitgestellt. Letzteres erfolgt gemäß (106), wobei den Größen ϱ, τ die Variablen c, s zuzuordnen sind:

```
procedure HARMON(m,n,F,A,B);
          value m,n; integer m,n; array F,A,B;
begin     integer k,j; real h,h1,u0,u1,u2,c,s,c1,s1,q;
          h := 2/m; h1 := h × 3.14159265;
          c1 := cos(h1); s1 := sin(h1); c := 1; s := 0;
          for k := 0 step 1 until n do begin
          u1 := u2 := 0;
          for j := m − 1 step −1 until 1 do begin
          u0 := F[j] + 2 × c × u1 − u2;
          u2 := u1; u1 := u0                    end;
          A[k] := h × (F[0] + c × u1 − u2); B[k] := h × s × u1;
          q := c1 × c − s1 × s; s := s1 × c + c1 × s;
          c := q                                end
end
```

Mit Hilfe der Prozedur *HARMON* wurden folgende Beispiele gerechnet:

1. Tabelle 5.2 enthält die in Volt gemessenen Werte für den zeitlichen Verlauf einer Kippspannung mit der Periode $p = \frac{1}{6}$ s an den $m = 12$ äquidistanten Stellen (83). Der Ungleichung (85) entsprechend wurden zur Approximation trigonometrische Polynome des Grades $n = 3$ und $n = 5$ gewählt; die Prozedur *HARMON* liefert dafür folgende Systeme von Fourierkoeffizienten:

$n = 3$

$a_0 = 8{,}85$

$a_1 = -2{,}41 \qquad b_1 = 3{,}21$

$a_2 = -1{,}20 \qquad b_2 = -0{,}04$

$a_3 = -0{,}20 \qquad b_3 = -0{,}08$

$n = 5$

$a_0 = 8{,}85$

$a_1 = -2{,}42 \qquad b_1 = 3{,}21$

$a_2 = -1{,}20 \qquad b_2 = -0{,}04$

$a_3 = -0{,}20 \qquad b_3 = -0{,}08$

$a_4 = -0{,}22 \qquad b_4 = 0{,}16$

$a_5 = -0{,}28 \qquad b_5 = 0{,}04$

j	1	2	3	4	5	6	7	8	9	10	11	12
y_j	3,85	6,45	8,85	9,10	7,30	5.80	4,45	3,25	2,20	1,30	0,55	0,00

Tabelle 5.2

Abb. 5.9 zeigt den Verlauf der Bestapproximation fünften Grades über einem Periodenintervall und die Abweichungen von den Meßwerten der Tabelle 5.2.

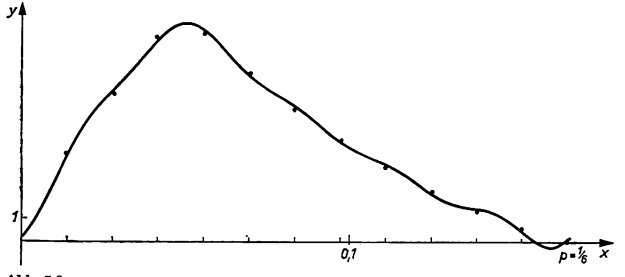


Abb. 5.9

2. Um bei optischen Bauelementen den Anteil des reflektierten und des durchgelassenen Lichtes durch Interferenz in gewünschter Weise zu beeinflussen, werden „dünne Schichten" geeigneter Substanzen aufgedampft. T (Transmissionsgrad) bedeutet den von der Vakuumwellenlänge λ abhängenden Bruchteil der durchgelassenen Lichtenergie bei senkrechtem Einfall.

Für einen Belag aus vier dielektrischen Schichten gleicher optischer Dicke Δ auf einem Glasträger ergaben sich die in Tab. 5.3 zusammengefaßten Werte für $F := T^{-1}$ in Abhängigkeit von der Größe $x := \frac{4\pi}{\lambda}\Delta$ an den $m = 9$ äquidistanten Stellen (83) des Intervalls $[\![0, p]\!] = [\![0, 2\pi]\!]$. Bestimmt man die Bestapproximation in der Klasse der trigonometrischen Polynome vierten Grades, so werden nach Satz 10 die Werte der Tabelle 5.3 im Rahmen der Rechengenauigkeit interpoliert.

j	1	2	3	4	5	6	7	8	9
T_j	1,192	1,319	1,112	1,031	1,031	1,112	1,319	1,192	1,042

Tabelle 5.3

Die Prozedur *HARMON* liefert für $m = 9$ und $n = 4$ folgendes System von Fourierkoeffizienten:

$$a_0 = 2{,}300 \qquad a_1 = 0{,}061 \qquad a_2 = -0{,}123$$
$$a_3 = -0{,}061 \qquad a_4 = 0{,}015;$$

die Sinuskoeffizienten verschwinden. Dieses Ergebnis ist in Übereinstimmung mit der Theorie der Interferenzschichtsysteme, wonach T^{-1} in dem betrachteten Fall ein Kosinuspolynom vierten Grades in x sein muß.

Wir betrachten noch zwei weitere Anwendungen des Satzes 11, zunächst die Berechnung sogenannter *Tschebyscheff-Reihen*, auf die wir in 5.3.2. zurückkommen werden. Zu berechnen sei die Linearkombination von Tschebyscheff-Polynomen 1. Art

$$f(x) = \sum_{j=0}^{l} c_j T_j(x). \tag{107}$$

Auf Grund von (44) handelt es sich dabei um eine Kosinussumme (96) mit dem Argument $\arccos x$. Nach Satz 11 ist

$$f(x) = c_0 + U_1 x - U_2 \tag{108}$$

mit

$$\begin{aligned} &U_{l+2} := U_{l+1} := 0 \\ &U_\varkappa := c_\varkappa + 2xU_{\varkappa+1} - U_{\varkappa+2}, \qquad \varkappa = l(-1)1. \end{aligned} \tag{109}$$

Abschließend erörtern wir einen Algorithmus zur Berechnung von Polynomen

$$P(z) = \sum_{j=0}^{l} c_j z^j$$

mit reellen Koeffizienten für komplexes Argument (vgl. [8])

$$z = x + iy = r(\cos\varphi + i\sin\varphi), \qquad r \geqq 0.$$

Es ist

$$P(z) = \sum_{j=0}^{l} c_j z^j = \sum_{j=0}^{l} c_j r^j \cos j\varphi + i \sum_{j=1}^{l} c_j r^j \sin j\varphi$$

die Zerlegung von $P(z)$ in Real- und Imaginärteil. Nach Satz 11 findet man dafür

$$\operatorname{Re} P(z) = c_0 + U_1 \cos\varphi - U_2, \qquad \operatorname{Im} P(z) = U_1 \sin\varphi,$$

wenn

$$U_{l+2} := U_{l+1} := 0,$$

$$U_\varkappa := c_\varkappa r^\varkappa + 2\cos\varphi\, U_{\varkappa+1} - U_{\varkappa+2}, \qquad \varkappa = l(-1)1,$$

ist. Mit der Transformation $U_\varkappa = r^\varkappa W_\varkappa$ folgt daraus für $r > 0$, daß mit

$$\begin{aligned} &W_{l+2} := W_{l+1} := 0, \\ &W_\varkappa := c_\varkappa + 2r\cos\varphi W_{\varkappa+1} - r^2 W_{\varkappa+2} = c_\varkappa + 2xW_{\varkappa+1} - r^2 W_{\varkappa+2}, \\ &\varkappa = l(-1)1, \end{aligned} \tag{110}$$

die Beziehungen

$$\begin{aligned} &\operatorname{Re} P(z) = c_0 + rW_1\cos\varphi - r^2 W_2 = c_0 + xW_1 - r^2 W_2, \\ &\operatorname{Im} P(z) = rW_1 \sin\varphi = yW_1 \end{aligned} \tag{111}$$

gelten.

Für die manuelle Berechnung von (111) nach der Rekursion (110) eignet sich das sogenannte *doppelzeilige Hornerschema*:

$$\begin{array}{ccccccccc} c_l & c_{l-1} & c_{l-2} & c_{l-3} & \dots & c_2 & c_1 & c_0 \\ & & c'_l s & c'_{l-1} s & \dots & c'_4 s & c'_3 s & c'_2 s \\ & c'_l t & c'_{l-1} t & c'_{l-2} t & \dots & c'_3 t & c'_2 t & \\ \hline c'_l & c'_{l-1} & c'_{l-2} & c'_{l-3} & \dots & c'_2 & c'_1 & c'_0 \end{array} \tag{112}$$

In (112) ist

$$s := -(x^2 + y^2), \qquad t := 2x, \tag{113}$$

und $c'_\varkappa$, $\varkappa = l(-1)0$, bezeichnet die Summe der über dieser Größe stehenden Werte. Diese lassen sich schrittweise mit $\varkappa = l$ beginnend berechnen. Man erkennt sofort,

daß die $c'_\varkappa$ für $\varkappa = l(-1)1$ mit den entsprechenden $W_\varkappa$ der Rekursion (110) übereinstimmen. Auf Grund dessen ist nach (111), (112)

$$\operatorname{Re} P(z) = xc'_1 + c'_0, \qquad \operatorname{Im} P(z) = yc'_1. \tag{114}$$

Wir wollen das Verfahren noch in einer ALGOL-Prozedur *POLKA* zusammenfassen. Formale Parameter derselben sind das Feld C der Koeffizienten, x, y als Real- bzw. Imaginärteil des Argumentes z und p, q als Real- bzw. Imaginärteil von $P(z)$; l steht für den Grad des Polynoms:

```
procedure POLKA(C,l,x,y,p,q);
          value x,y; integer l; real x,y,p,q; array C;
          begin
          integer j; real s,t,w,w1,w2;
          s := -(x × x + y × y); t := x + x; w1 := w2 := 0;
          for j := l step -1 until 1 do begin
          w := C[j] + t × w1 + s × w2;
          w2 := w1; w1 := w                end;
          p := C[0] + x × w1 + s × w2; q := y × w1
          end
```

5.3. Gleichmäßige Approximation

Dieser Abschnitt betrifft das mit der Norm

$$\|f\| = \max_{x\in\langle a,b\rangle} |f(x)| \tag{1}$$

über $C_{\langle a,b\rangle}$ formulierte lineare Approximationsproblem 5.1.2.(23). Nach der Erörterung grundlegender Begriffsbildungen werden einige Ergebnisse der Theorie (zum Teil ohne Beweis) mitgeteilt und angewendet.

5.3.1. Grundlegende Begriffe und Ergebnisse

Es sei daran erinnert, daß das eingangs formulierte Problem nach 5.1.2., Satz 2, eine Lösung besitzt, die jedoch auf Grund der Unitätsbetrachtung in 5.1.2. im Hinblick auf Satz 4 dieses Abschnittes nicht eindeutig bestimmt sein muß. Die Frage, unter welchen zusätzlichen Forderungen an das linear unabhängige Funktionensystem $\{\varphi_i\}$ genau eine Bestapproximation existiert, wurde von dem ungarischen Mathematiker Alfred Haar (1895—1933) geklärt. Ausgehend von Praxisproblemen der angenäherten Synthese von Mechanismen hat P. L. Tschebyscheff (1821—1894) etwa von 1850 bis zum Ende seines Lebens die Theorie der gleichmäßigen Approximation in zahlreichen Arbeiten untersucht und wesentlich begründet (vgl. [51, 52]). Es ist daher berechtigt, diese auch als *Tschebyscheff-Approximation* zu bezeichnen.

Folgende Definitionen und Sätze sind grundlegend:

Definition 1. n Funktionen $\varphi_i \in C_{[a,b]}$, $i = 1(1)n$, bilden ein *Tschebyscheff-System*[1]) bezüglich des Intervalls $[\![a, b]\!]$ genau dann, wenn jede nichttriviale Linearkombination

$$F_a = \sum_{i=1}^{n} a_i \varphi_i, \qquad a \neq 0,$$

auf $[\![a, b]\!]$ höchstens $n - 1$ Nullstellen besitzt.

Beispiel 1. Die Gesamtheit der Funktionen $\varphi_i : x \mapsto x^i$, $i = 0(1)n - 1$, ist bezüglich jedes Intervalls ein Tschebyscheff-System, da eine nichttriviale Linearkombination ein Polynom vom Grade $\leqq n - 1$ darstellt und nach dem Fundamentalsatz der Algebra höchstens $n - 1$ Nullstellen hat.

Beispiel 2. Die Funktionen

$$\varphi_0 : x \mapsto 1, \quad \varphi_1 : x \mapsto x, \quad \varphi_2 : x \mapsto \frac{1}{x}, \quad \varphi_3 : x \mapsto \frac{1}{x^2}$$

bilden bezüglich jedes Intervalls, das den Nullpunkt nicht enthält, ein Tschebyscheff-System. Der Beweis läßt sich wieder auf den Fundamentalsatz der Algebra zurückführen und sei dem Leser als Übung empfohlen.

Bemerkung 1. Wenn für alle φ_i die Beziehung $\varphi_i(a) = \varphi_i(b)$ gilt, d. h. die φ_i als stetige Funktionen über die Zahlgerade periodisch fortgesetzt werden können, dann sind bei der Bestimmung eines Tschebyscheff-Systems Nullstellen an den Enden des Intervalls $[\![a, b]\!]$ nur einmal zu zählen.

Beispiel 3. Auf die Funktionen 5.2.5.(84) trifft die Bemerkung 1 zu. Diese bilden im Sinne der für periodische Funktionen modifizierten Definition 1 ein Tschebyscheff-System bezüglich des Intervalls $[\![0, p]\!]$. Im Sinne eines indirekten Beweises sei angenommen, daß das durch nichttriviale Linearkombination entstehende trigonometrische Polynom

$$F(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n} \left(a_\nu \cos \frac{2\pi\nu x}{p} + b_\nu \sin \frac{2\pi\nu x}{p} \right) \tag{2}$$

wenigstens $2n + 1$ Nullstellen im Intervall $]\!]0, p[\!]$ besitzt; diese seien $x_1 < x_2 < \cdots < x_{2n+1}$. Mit i als imaginärer Einheit gilt

$$\cos \frac{2\pi\nu x}{p} = \frac{e^{i\frac{2\pi\nu x}{p}} + e^{-i\frac{2\pi\nu x}{p}}}{2}, \qquad \sin \frac{2\pi\nu x}{p} = \frac{e^{i\frac{2\pi\nu x}{p}} - e^{-i\frac{2\pi\nu x}{p}}}{2i} \tag{3}$$

und — wenn

$$c_0 := \frac{a_0}{2}, \quad c_\nu := \frac{1}{2}(a_\nu - ib_\nu), \quad c_{-\nu} := \frac{1}{2}(a_\nu + ib_\nu), \quad \nu = 1(1)n, \tag{4}$$

[1]) Diese Bezeichnung wurde von S. N. Bernstein eingeführt. — Wenn es zweckmäßig erscheint, wird auch eine mit Null beginnende Indizierung der φ_i verwendet.

gesetzt wird —

$$F(x) = \sum_{\nu=-n}^{n} c_\nu e^{i\frac{2\nu x}{p}}$$
$$= e^{-i\frac{2nx}{p}}\left[c_{-n} + c_{-n+1}e^{i\frac{2x}{p}} + \cdots + c_0 e^{i\frac{2nx}{p}} + \cdots + c_n e^{i\frac{4nx}{p}}\right]$$
$$= e^{-i\frac{2nx}{p}} P_{2n}\left(e^{i\frac{2x}{p}}\right), \tag{5}$$

wobei P_{2n} ein Polynom vom Grade $\leqq 2n$ bedeutet, dessen Koeffizienten $c_{-n}, c_{-n+1}, \ldots, c_n$ sind. Da die Exponentialfunktion für kein Argument verschwindet, besitzt P_{2n} die Nullstellen

$$\zeta_j = e^{i\frac{2nx_j}{p}}, \qquad j = 1(1)2n+1.$$

Diese sind paarweise verschieden. Aus $\zeta_j = \zeta_k$, $1 \leqq j, k \leqq 2n+1$, $j < k$, würde nämlich

$$e^{i\frac{2n(x_k - x_j)}{p}} = 0$$

und damit

$$x_k - x_j = mp \quad \text{mit} \quad m \in \mathbf{N}^*$$

folgen. Dem widerspricht aber $x_k - x_j < p$. Das Polynom P_{2n} besitzt also mindestens $2n+1$ Nullstellen und ist nach dem Fundamentalsatz der Algebra das Nullpolynom. Auf Grund von (4) verschwinden dann auch sämtliche a_ν, b_ν, und (2) wäre im Widerspruch zur Annahme die triviale Linearkombination des Funktionensystems 5.2.5.(84).

Bemerkung 2. Man kann allgemein zeigen, daß die Anzahl der in einem periodischen Tschebyscheff-System enthaltenen Funktionen ungerade ist (vgl. [6], 2.10.).

Satz 1. *Die Funktionen eines Tschebyscheff-Systems sind linear unabhängig.*

Beweis. Es sei $\{\varphi_i\}$, $i = 1(1)n$, ein Tschebyscheff-System bezüglich $[\![a, b]\!]$ und $\sum_{i=1}^{n} a_i \varphi_i(x) \equiv 0$ auf diesem Intervall. Dann müssen sämtliche a_i verschwinden, da andernfalls die betrachtete Linearkombination höchstens $n-1$ Nullstellen auf $[\![a, b]\!]$ hätte, dort also nicht identisch verschwinden könnte.

Es gibt jedoch linear unabhängige Funktionen, die kein Tschebyscheff-System bilden:

Beispiel 4. Die Funktionen $\varphi_1 : x \mapsto x$, $\varphi_2 : x \mapsto e^x$ sind bezüglich jedes Intervalls linear unabhängig. In der Tat folgt aus

$$a_1 x + a_2 e^x \equiv 0$$

durch zweimalige Differentiation $a_2 e^x \equiv 0$, also $a_2 = 0$ und dann auch $a_1 = 0$. Die Funktionen φ_1, φ_2 bilden aber z. B. bezüglich des Intervalls $[\![0, 3]\!]$ kein Tschebyscheff-System. Die spezielle Linearkombination $F(x) = 4x - e^x$ ist negativ bei

$x = 0$ und $x = 3$, positiv bei $x = 1$. Damit existieren nach dem Zwischenwertsatz für stetige Funktionen mindestens 2 ($= n$) Nullstellen von F im Intervall $[\![0, 3]\!]$.

Im folgenden werden zwei Hilfssätze über Tschebyscheff-Systeme bewiesen, aus denen sich u. a. ein wichtiges Resultat der Interpolationstheorie ableiten läßt.

Hilfssatz 1. *Es sei* $\{\varphi_i\}$, $i = 1(1)n$, *ein Tschebyscheff-System bezüglich* $[\![a, b]\!]$, *und für gewisse* $x_i, x_{i+1}, \ldots, x_k$, $1 \leqq i \leqq k < n$, *dieses Intervalls gelte*

$$\begin{vmatrix} \varphi_i(x_i) & \varphi_{i+1}(x_i) & \cdots & \varphi_k(x_i) \\ \varphi_i(x_{i+1}) & \varphi_{i+1}(x_{i+1}) & \cdots & \varphi_k(x_{i+1}) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_i(x_k) & \varphi_{i+1}(x_k) & \cdots & \varphi_k(x_k) \end{vmatrix} \neq 0. \tag{6}$$

Dann gibt es für jedes q, $k < q \leqq n$, *Punkte* $x_{k+1}, x_{k+2}, \ldots, x_q$ *auf* $[\![a, b]\!]$ *derart, daß*

$$\begin{vmatrix} \varphi_i(x_i) & \varphi_{i+1}(x_i) & \cdots & \varphi_k(x_i) & \cdots & \varphi_q(x_i) \\ \varphi_i(x_{i+1}) & \varphi_{i+1}(x_{i+1}) & \cdots & \varphi_k(x_{i+1}) & \cdots & \varphi_q(x_{i+1}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \varphi_i(x_k) & \varphi_{i+1}(x_k) & \cdots & \varphi_k(x_k) & \cdots & \varphi_q(x_k) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \varphi_i(x_q) & \varphi_{i+1}(x_q) & \cdots & \varphi_k(x_q) & \cdots & \varphi_q(x_q) \end{vmatrix} \neq 0 \tag{7}$$

ist.

Beweis. Die durch

$$F(x) = \begin{vmatrix} \varphi_i(x_i) & \cdots & \varphi_k(x_i) & \varphi_{k+1}(x_i) \\ \varphi_i(x_{i+1}) & \cdots & \varphi_k(x_{i+1}) & \varphi_{k+1}(x_{i+1}) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_i(x_k) & \cdots & \varphi_k(x_k) & \varphi_{k+1}(x_k) \\ \varphi_i(x) & \cdots & \varphi_k(x) & \varphi_{k+1}(x) \end{vmatrix} \tag{8}$$

auf $[\![a, b]\!]$ definierte Funktion F ist eine nichttriviale Linearkombination der Funktionen $\varphi_i, \ldots, \varphi_k, \varphi_{k+1}$ und damit auch eine solche des in der Voraussetzung gegebenen Tschebyscheff-Systems $\{\varphi_i\}$. Um das einzusehen, braucht man (8) nur nach der letzten Zeile zu entwickeln und (6) zu beachten. F besitzt höchstens $n - 1$ Nullstellen auf $[\![a, b]\!]$, und es muß daher ein von den $x_i, \ldots, x_k$ verschiedenes Argument x_{k+1} dieses Intervalls geben, für welches $F(x_{k+1}) \neq 0$ ist. Indem man erneut eine Determinante der Form (8) mit den Punkten $x_i, \ldots, x_k, x_{k+1}$ und den Funktionen $\varphi_i, \ldots, \varphi_k, \varphi_{k+1}, \varphi_{k+2}$ bildet, gelangt man so weiterschließend zu dem Resultat (7).

Hilfssatz 2. *Es sei* $\{\varphi_i\}$, $i = 1(1)n$, *ein Tschebyscheff-System bezüglich* $[\![a, b]\!]$ *und* $x_1, x_2, \ldots, x_k$ ($k \leqq n$) *eine beliebige Menge paarweise verschiedener Punkte dieses Inter-*

valls. Dann hat die Matrix

$$\begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \dots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \dots & \varphi_n(x_2) \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \varphi_1(x_k) & \varphi_2(x_k) & \dots & \varphi_n(x_k) \end{pmatrix} \tag{9}$$

den Rang k.

Beweis. Wir beweisen die Behauptung induktiv. Für $k = 1$ ist zu zeigen, daß nicht alle der Werte $\varphi_j(x_1)$, $j = 1(1)n$, verschwinden. Indirekt schließend betrachte man Punkte $y_2, y_3, \dots, y_n$ des Intervalls $[\![a, b]\!]$, für welche

$$\begin{vmatrix} \varphi_2(y_2) & \varphi_3(y_2) & \dots & \varphi_n(y_2) \\ \varphi_2(y_3) & \varphi_3(y_3) & \dots & \varphi_n(y_3) \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \varphi_2(y_n) & \varphi_3(y_n) & \dots & \varphi_n(y_n) \end{vmatrix} \neq 0 \tag{10}$$

ist. Deren Existenz läßt sich so erkennen: Auf Grund von Satz 1 gibt es ein $y_2 \in [\![a, b]\!]$, für welches $\varphi_2(y_2) \neq 0$, und nach Hilfssatz 1 weitere Punkte $y_3, \dots, y_n$, die (10) verifizieren. Die durch

$$F(x) = \begin{vmatrix} \varphi_1(x) & \varphi_2(x) & \dots & \varphi_n(x) \\ \varphi_1(y_2) & \varphi_2(y_2) & \dots & \varphi_n(y_2) \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \varphi_1(y_n) & \varphi_2(y_n) & \dots & \varphi_n(y_n) \end{vmatrix} \tag{11}$$

definierte Funktion F ist eine nichttriviale Linearkombination der Funktionen $\varphi_1, \varphi_2, \dots, \varphi_n$, da der Koeffizient bei φ_1 der von Null verschiedene Determinantenwert (10) ist. Diese hätte bei x_1 eine Nullstelle, sofern $\varphi_j(x_1) = 0$, $j = 1(1)n$, ist, und außerdem gilt $F(x) = 0$ für $x = y_j$, $j = 2(1)n$, da dann zwei Zeilen in (11) übereinstimmen. Mit Rücksicht auf Definition 1 kann F aber nicht n Nullstellen haben, und daher ist $\varphi_j(x_1) \neq 0$ für mindestens ein $j \in \{1, 2, \dots, n\}$.

Wir nehmen nun an, daß die Behauptung des Hilfssatzes 2 für $k = 1, 2, \dots, m-1$ ($m \leqq n$) gilt, und betrachten die Matrix (9) für $x_1, x_2, \dots, x_m$. Durch Umnumerierung der φ_j kann dann stets erreicht werden, daß

$$\begin{vmatrix} \varphi_2(x_2) & \dots & \varphi_m(x_2) \\ \dots\dots\dots\dots\dots\dots \\ \varphi_2(x_m) & \dots & \varphi_m(x_m) \end{vmatrix} \neq 0$$

ist. Auf Grund des Hilfssatzes 1 existieren Punkte $y_{m+1}, y_{m+2}, \dots, y_n$ derart, daß

$$\begin{vmatrix} \varphi_2(x_2) & \dots & \varphi_m(x_2) & \varphi_{m+1}(x_2) & \dots & \varphi_n(x_2) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \varphi_2(x_m) & \dots & \varphi_m(x_m) & \varphi_{m+1}(x_m) & \dots & \varphi_n(x_m) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \varphi_2(y_n) & \dots & \varphi_m(y_n) & \varphi_{m+1}(y_n) & \dots & \varphi_n(y_n) \end{vmatrix} \neq 0$$

gilt. Somit ist die durch

$$\begin{vmatrix} \varphi_1(x) & \varphi_2(x) & \cdots & \varphi_n(x) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_n(x_2) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \varphi_1(x_m) & \varphi_2(x_m) & \cdots & \varphi_n(x_m) \\ \varphi_1(y_{m+1}) & \varphi_2(y_{m+1}) & \cdots & \varphi_n(y_{m+1}) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \varphi_1(y_n) & \varphi_2(y_n) & \cdots & \varphi_n(y_n) \end{vmatrix} \tag{12}$$

definierte Funktion eine nichttriviale Linearkombination der φ_i. Diese besitzt offenbar die $n-1$ Nullstellen

$$x_2, x_3, \ldots, x_m, y_{m+1}, \ldots, y_n.$$

Würde (9) bezüglich der Argumente $x_1, x_2, \ldots, x_m$ nicht den Rang m haben, so verschwände (12) auch für x_1, was man sofort durch Entwicklung dieser Determinante nach den ersten m Zeilen erkennt. Die Existenz von n Nullstellen für (12) widerspricht aber der Eigenschaft von $\{\varphi_i\}$, ein Tschebyscheff-System zu sein. Damit ist die Induktionsbehauptung bewiesen.

Wir können nun ein allgemeines Theorem der Interpolationstheorie ableiten, das zwei früher behandelte Spezialfälle einschließt.

Satz 2. *$\{\varphi_i\}$, $i = 1(1)n$, sei ein Tschebyscheff-System bezüglich $[\![a, b]\!]$ und $x_1, x_2, \ldots, x_n$ eine Menge paarweise verschiedener Punkte dieses Intervalls. Dann gibt es für beliebige $y_j \in \mathbf{R}$, $j = 1(1)n$, genau eine Linearkombination*

$$F_a = \sum_{i=1}^{n} a_i \varphi_i, \tag{13}$$

für welche

$$F_a(x_j) = \sum_{i=1}^{n} a_i \varphi_i(x_j) = y_j, \qquad j = 1(1)n, \tag{14}$$

ist, d. h., es gibt in der Menge der Linearkombinationen der φ_i eine wohlbestimmte, welche die Interpolationsaufgabe bezüglich der Knoten x_i und dafür gegebener Ordinaten löst.

Beweis. (14) ist ein lineares Gleichungssystem für die in (13) auftretenden a_i mit der Koeffizientendeterminante

$$\begin{vmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_n(x_2) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix}.$$

Diese ist nach Hilfssatz 2 von Null verschieden, und das System (14) besitzt genau eine Lösung.

Die Bestimmung von (13) für das in Beispiel 1 betrachtete Tschebyscheff-System führt auf die in MfL Bd. 9, 4.2.1., erörterte Interpolation mit ganzen rationalen Funktionen. (14) besitzt in diesem Falle eine Vandermondesche Koeffizientendeterminante, deren Nichtverschwinden auf Grund der bekannten Produktdarstellung offensichtlich ist.

Bezüglich des periodischen Tschebyscheff-Systems aus Beispiel 3 und der Knoten 5.2.5.(83) erhält man eine spezielle Aufgabe der trigonometrischen Interpolation, deren eindeutige Lösbarkeit bereits nach 5.2.5., Satz 10, erkannt ist.

In beiden Fällen verfügen wir über Algorithmen zur Bestimmung von (13), die unabhängig von der Lösung des linearen Gleichungssystems (14) sind.

Der Erörterung des oben angesprochenen Einzigkeitsproblems für die gleichmäßige Approximation schicken wir das folgende notwendige Kriterium voraus:

Hilfssatz 3. *$\{\varphi_i\}$, $i = 1(1)n$, sei ein Tschebyscheff-System bezüglich $[\![a, b]\!]$, $f \in C_{(a,b)}$ und $F_a = \sum_{i=1}^{n} a_i\varphi_i$. Dann gilt:*

$$\begin{aligned}&F_a \textit{ ist Bestapproximation (im Tschebyscheffschen Sinne) von } f \Rightarrow \textit{die}\\ &\textit{Mächtigkeit der Menge } \{x: x \in [\![a, b]\!] \wedge |f(x) - F_a(x)| = \|f - F_a\|\}\\ &\textit{ist größer oder gleich } n.\end{aligned} \tag{15}$$

Beweis. Wir beweisen (15) in der kontraponierten Form und nehmen an, daß $\{x_j\}$, $x_j \in [\![a, b]\!]$, $j = 1(1)m$, $m < n$, die Menge sei, auf der

$$|f(x_j) - F_a(x_j)| = \|f - F_a\|$$

gilt. Das lineare Gleichungssystem

$$\sum_{i=1}^{n} \xi_i\varphi_i(x_j) = f(x_j) - F_a(x_j), \qquad j = 1(1)m, \tag{16}$$

ist nach Hilfssatz 2 lösbar, wobei die ξ_i nicht alle verschwinden können, da gewiß $f \neq F_a$ und folglich $|f(x_j) - F_a(x_j)| = \|f - F_a\| = Z(a) > 0$ ist. Es sei

$$R: x \mapsto f(x) - F_a(x); \tag{17}$$

R ist auf $[\![a, b]\!]$ stetig, und $|R|$ nimmt an den Stellen x_j den positiven Wert $Z(a)$ an. Diese werden durch paarweise disjunkte Mengen $U_j = I_j \cap [\![a, b]\!]$ überdeckt, wobei I_j ein offenes Intervall der Form $|x - x_j| < r_j$ ist. Die r_j seien so klein gewählt, daß die U_j offene (bzw. für $x_j \in \{a, b\}$ halboffene) Intervalle sind, auf denen

$$\mu_j := \inf_{x \in U_j} |R(x)| > 0 \tag{18}$$

und — wenn ξ einen Lösungsvektor von (16) bezeichnet —

$$\inf_{x \in U_j} |F_\xi(x)| \geqq \frac{Z(a)}{2} \tag{19}$$

gilt. Auf der abgeschlossenen Menge $U^* = [\![a, b]\!] \setminus U_1 \setminus U_2 \setminus \cdots \setminus U_m$ bedeutet

$$M := \max_{x \in U^*} |F_\xi(x)| \tag{20}$$

und

$$Z^*(a) := \max_{x \in U^*} |R(x)|. \tag{21}$$

Ferner sei

$$M_j := \sup_{x \in U_j} |F_\xi(x)|. \tag{22}$$

Da $|R(x)| < Z(a)$ auf U^* ist und das Maximum von $|R(x)|$ auf U^* als Funktionswert dort angenommen wird, ist offenbar

$$\mu := Z(a) - Z^*(a) > 0. \tag{23}$$

Nun wählen wir ε so, daß

$$0 < \varepsilon < \min\left\{\frac{\mu}{M}, \frac{\mu_1}{M_1}, \ldots, \frac{\mu_m}{M_m}\right\} \tag{24}$$

ist, und bilden den Parametervektor

$$a' := a + \varepsilon\xi. \tag{25}$$

Für diesen gilt

$$|f(x) - F_{a'}(x)| = |f(x) - F_a(x) - \varepsilon F_\xi(x)| = |R(x) - \varepsilon F_\xi(x)|. \tag{26}$$

Damit findet man auf U_j, $j = 1(1)m$, unter Beachtung von (19)

$$|f(x) - F_{a'}(x)| = |R(x)|\left(1 - \varepsilon\frac{F_\xi(x)}{R(x)}\right) = |R(x)|\left(1 - \varepsilon\frac{|F_\xi(x)|}{|R(x)|}\right)$$

$$\leqq Z(a)\left(1 - \varepsilon\frac{\frac{Z(a)}{2}}{Z(a)}\right) = Z(a)\left(1 - \frac{\varepsilon}{2}\right) < Z(a).$$

Bezüglich dieser Abschätzung sei darauf hingewiesen, daß R und F_ξ auf U_j das gleiche Vorzeichen haben. In der Tat nehmen diese Funktionen gemäß (16) und (17) bei x_j denselben Wert an und verschwinden wegen (18) und (19) auf U_j nicht. Damit folgt die Behauptung aus Stetigkeitsgründen. Ferner ist wegen (18), (22) und (24) auf U_j

$$1 - \varepsilon\frac{|F_\xi(x)|}{|R(x)|} \geqq 1 - \varepsilon\frac{M_j}{\mu_j} > 0.$$

Auf U^* folgt aus (26), (21), (20), (24) und (23)

$$|f(x) - F_{a'}(x)| \leqq |R(x)| + \varepsilon|F_\xi(x)| \leqq Z^*(a) + \varepsilon M < Z^*(a) + \mu = Z(a).$$

Durchweg gilt also auf $[\![a, b]\!]$

$$Z(a') < Z(a),$$

d. h., F_a ist nicht Bestapproximation von f.

Der folgende Satz von A. Haar bringt zum Ausdruck, daß die Einzigkeit der gleichmäßigen Bestapproximation durch Linearkombinationen aus linear unabhängigen Funktionen damit äquivalent ist, daß diese ein Tschebyscheff-System bilden.

Satz 3. *$\{\varphi_i\}$, $i = 1(1)n$, $\varphi_i \in C_{\{a,b\}}$, sei ein System bezüglich des Intervalls $[\![a, b]\!]$ linear unabhängiger Funktionen. Dann gilt:*

$$\bigwedge_f \left\{f \in C_{\{a,b\}} \Rightarrow \begin{array}{l}\textit{Bestapproximation im Tschebyscheffschen Sinne}\\ \textit{durch Linearkombinationen der } \varphi_i \textit{ eindeutig bestimmt}\end{array}\right\} \tag{27}$$

$\Leftrightarrow$ *$\{\varphi_i\}$ ist ein Tschebyscheff-System bezüglich $[\![a, b]\!]$.*

Beweis. Die eine der in (27) enthaltenen Implikationen beweisen wir in der kontraponierten Form:

Wenn $\{\varphi_i\}$ kein Tschebyscheff-System bezüglich $[\![a, b]\!]$ ist, dann gibt es für mindestens ein $f \in C_{\{a,b\}}$ mehrere Bestapproximationen.

Im Sinne der Voraussetzung sei $F_{\tilde{a}}$ eine nichttriviale Linearkombination der φ_i mit den paarweise verschiedenen Nullstellen $x_1, x_2, \ldots, x_n$. Für diese gilt

$$\begin{vmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_n(x_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \varphi_1(x_n) & \varphi_2(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} = 0, \tag{28}$$

da das homogene lineare Gleichungssystem (14) die Komponenten von $\tilde{a}$ als nichttriviale Lösung besitzt. Aus (28) folgt die Existenz von Zahlen

$$c_1, c_2, \ldots, c_n, \qquad \sum_{j=1}^{n} c_j^2 > 0, \tag{29}$$

derart, daß

$$\sum_{j=1}^{n} c_j \varphi_i(x_j) = 0 \qquad \text{für} \quad i = 1(1)n. \tag{30}$$

Mit (30) gewinnt man für jede Linearkombination F_a der φ_i

$$\sum_{j=1}^{n} c_j F_a(x_j) = \sum_{j=1}^{n} c_j \sum_{i=1}^{n} a_i \varphi_i(x_j) = \sum_{i=1}^{n} a_i \sum_{j=1}^{n} c_j \varphi_i(x_j) = 0. \tag{31}$$

$\lambda \neq 0$ sei so gewählt, daß

$$\max_{x \in \{a,b\}} |\lambda F_{\tilde{a}}(x)| < 1, \tag{32}$$

und $g \in C_{\{a,b\}}$ so, daß

$$|g(x)| \leqq 1 \text{ auf } [\![a, b]\!] \quad \text{und} \quad g(x_j) = \operatorname{sgn} c_j \tag{33}$$

gilt, sofern $c_j \neq 0$ ist. Eine derartige Funktion kann leicht mit Hilfe eines Streckenzuggraphen konstruiert werden. Offensichtlich hat auch die auf $[\![a, b]\!]$ stetige Funktion

$$f(x) = g(x)\,(1 - |\lambda F_{\tilde{a}}(x)|) \tag{34}$$

die Eigenschaft (33). Auf Grund dessen gilt für einen beliebigen Parametervektor a

$$Z(a) = \max_{x \in \{a,b\}} |f(x) - F_a(x)| \geqq 1, \tag{35}$$

Andernfalls wäre speziell für die x_j, für welche $c_j \neq 0$ ist,

$$|f(x_j) - F_a(x_j)| < 1$$

und wegen $|f(x_j)| = 1$

$$\operatorname{sgn} F_a(x_j) = \operatorname{sgn} f(x_j) = \operatorname{sgn} c_j.$$

Danach würde $\sum_{j=1}^{n} c_j F_a(x_j)$ positiv sein, im Widerspruch zu (31). Nunmehr ergibt sich für $|\varepsilon| \leqq 1$ auf Grund von (34), (32) und (33)

$$\begin{aligned} |f(x) - \varepsilon\lambda F_{\tilde{a}}(x)| &\leqq |f(x)| + |\varepsilon\lambda F_{\tilde{a}}(x)| \\ &= |g(x)|\,(1 - |\lambda F_{\tilde{a}}(x)|) + |\varepsilon\lambda F_{\tilde{a}}(x)| \\ &\leqq 1 - |\lambda F_{\tilde{a}}(x)|\,(1 - |\varepsilon|) \leqq 1, \end{aligned}$$

also im Hinblick auf (35)

$$Z(\varepsilon\lambda\tilde{a}) = \|f - \varepsilon\lambda F_{\tilde{a}}(x)\| = 1.$$

Das besagt: Alle Linearkombinationen $\varepsilon\lambda F_{\tilde{a}}$, $|\varepsilon| \leqq 1$, sind Bestapproximationen der Funktion (34). Die Konklusio der betrachteten Implikation ist damit verifiziert.

Um den Beweis von (27) zu vollenden, bleibt zu zeigen, daß die Eigenschaft von $\{\varphi_i\}$, ein Tschebyscheff-System auf $[\![a, b]\!]$ zu sein, für die Einzigkeit der Bestapproximation irgendeiner Funktion aus $C_{\{a,b\}}$ hinreichend ist. Im Sinne eines indirekten Beweises sei angenommen, daß für eine gewisse Funktion $f \in C_{\{a,b\}}$ zwei Linearkombinationen F_a, F_b $(a \neq b)$ der φ_i Bestapproximation realisieren. Dann folgt aus

$$\left|f(x) - F_{\frac{a+b}{2}}\right| \leqq \frac{1}{2}|f(x) - F_a(x)| + \frac{1}{2}|f(x) - F_b(x)|,$$

$$\left|f(x) - F_{\frac{a+b}{2}}\right| \leqq \frac{1}{2}\|f - F_a\| + \frac{1}{2}\|f - F_b\|, \tag{36}$$

$$\left\|f - F_{\frac{a+b}{2}}\right\| \leqq \frac{1}{2}\|f - F_a\| + \frac{1}{2}\|f - F_b\|,^{1})$$

daß auch $F_{\frac{a+b}{2}}$ diese Eigenschaft hat. Nach Hilfssatz 3 besitzt die Gleichung

$$\left|f(x) - F_{\frac{a+b}{2}}\right| = Z\left(\frac{a+b}{2}\right) = Z(a) = Z(b) = Z$$

wenigstens n Lösungen $x_1, x_2, \ldots, x_n$, d. h., es gilt

$$f(x_j) - F_{\frac{a+b}{2}}(x_j) = \frac{1}{2}(f(x_j) - F_a(x_j)) + \frac{1}{2}(f(x_j) - F_b(x_j)) = \pm Z, \quad j = 1(1)n. \tag{37}$$

Auf Grund von (36) ist aber auch

$$|f(x_j) - F_a(x_j)| = |f(x_j) - F_b(x_j)| = Z$$

und wegen (37) sogar

$$f(x_j) - F_a(x_j) = f(x_j) - F_b(x_j) = \pm Z, \qquad j = 1(1)n, \tag{38}$$

mit dem dort bei Z gegebenen Vorzeichen. Aus (38) folgt

$$F_{a-b}(x_j) = 0,$$

d. h., die nichttriviale Linearkombination F_{a-b} der φ_i besitzt mindestens n Nullstellen, was der Voraussetzung widerspricht, daß $\{\varphi_i\}$ ein Tschebyscheff-System ist.

Wir beschließen diesen Abschnitt mit Betrachtungen zum *Tschebyscheffschen Alternantensatz*, der ein notwendiges und hinreichendes Kriterium dafür enthält, daß eine aus Funktionen eines Tschebyscheff-Systems gebildete Linearkombination F_a eine Funktion $f \in C_{\{a,b\}}$ am besten gleichmäßig approximiert. Die Formulierung bezieht sich auf folgende Definitionen:

Definition 2. Es sei $f \in C_{\{a,b\}}$ und $X \subseteq [\![a, b]\!]$, $X = \{x_j\}$, $j = 0(1)k$, $x_0 < x_1 < \cdots < x_k$.

$$f \text{ alterniert über } X: \Leftrightarrow f(x_j) = (-1)^j \zeta \max_{a \leqq x \leqq b} |f(x)|, \; j = 0(1)k,$$

wobei $\zeta \in \{-1, 1\}$ ist. X wird auch als *Alternante* bezeichnet.

[1]) Tatsächlich muß in dieser Abschätzung das Gleichheitszeichen gelten.

Definition 3. $f \in C_{(a,b)}$ alterniert auf $[\![a, b]\!]$ k-mal :$\Leftrightarrow$ es gibt eine Menge $X \subseteq [\![a, b]\!]$ der Mächtigkeit $k+1$, über der f alterniert, und jede Alternante $Y \subseteq [\![a, b]\!]$ besitzt eine Mächtigkeit $\leqq k+1$.

Satz 4. *Es sei* $\{\varphi_i\}$, $i = 1(1)n$, *ein Tschebyscheff-System bezüglich* $[\![a, b]\!]$, $f \in C_{(a,b)}$ *und* $F_a = \sum_{i=1}^{n} a_i \varphi_i$. *Dann gilt:*

F_a *Bestapproximation von* $f \Leftrightarrow f - F_a$ *alterniert auf* $[\![a, b]\!]$ *mindestens* n*-mal.*

Hiernach gibt es auf Grund von Satz 3 genau eine Linearkombination der φ_i, welche die Bedingung des Satzes 4 erfüllt. Deren Hinlänglichkeit ist leicht einzusehen: Angenommen, F_a besitzt eine Alternante X mit $n+1$ Elementen und ist nicht Bestapproximation von f, sondern dies trifft für $F_{\tilde{a}}$ mit $a \neq \tilde{a}$ zu. Dann wechselt die Linearkombination

$$F_{\tilde{a}} - F_a = (f - F_a) - (f - F_{\tilde{a}})$$

beim Übergang von einem Punkt x_j aus X zum nächstfolgenden das Vorzeichen, denn nach Voraussetzung ist

$$\|f - F_{\tilde{a}}\| < \|f - F_a\|, \qquad |f(x_j) - F_a(x_j)| = \|f - F_a\|$$

und folglich

$$\operatorname{sgn}\big(F_{\tilde{a}}(x_j) - F_a(x_j)\big) = \operatorname{sgn}\big(f(x_j) - F_a(x_j)\big).$$

Nach dem Zwischenwertsatz für stetige Funktionen besitzt $F_{\tilde{a}} - F_a$ also mindestens n Nullstellen auf $[\![a, b]\!]$. Daraus ergibt sich ein Widerspruch, da die φ_i ein Tschebyscheff-System bilden und $F_{\tilde{a}} - F_a$ wegen $a \neq \tilde{a}$ eine nichttriviale Linearkombination dieser Funktionen ist.

Der Beweis der Notwendigkeit der Bedingung des Satzes 4 ist weitläufiger und soll hier nicht ausgeführt werden.

Um zu zeigen, daß man mit Hilfe des Satzes 4 unter Umständen die Minimallösung auch konstruieren kann, betrachten wir die beste Tschebyscheff-Approximation einer Funktion $f \in C_{(a,b)}$, die auf $[\![a, b]\!]$ eine zweite Ableitung konstanten Vorzeichens besitzt, durch ein lineares Polynom

$$F_a(x) = a_0 + a_1 x. \tag{39}$$

Ohne Beschränkung der Allgemeinheit kann $f'' > 0$ angenommen werden. Bei den folgenden Überlegungen wollen wir uns auf die bewiesene Hinlänglichkeit des Alternantenkriteriums stützen. Danach ist (39) Minimallösung, wenn für einen inneren Punkt ξ des Intervalls $[\![a, b]\!]$ die Gleichungen

$$f(a) - a_0 - a_1 a = L, \tag{40a}$$

$$f(\xi) - a_0 - a_1 \xi = -L, \tag{40b}$$

$$f(b) - a_0 - a_1 b = L \tag{40c}$$

mit $L := \|f - F_a\|$ erfüllt sind. Aus (40a), (40c) folgt durch Subtraktion zunächst

$$a_1 = \frac{f(b) - f(a)}{b - a}, \tag{41}$$

d. h., die der Minimallösung entsprechende Gerade ist zur Sehne der Kurve $y = f(x)$ über dem Intervall $[\![a, b]\!]$ parallel. Auf Grund des Mittelwertsatzes gibt es wegen der Monotonie von f' genau ein $\xi \in]a, b[$, für welches

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}$$

ist. Bedeutet weiterhin in (40) ξ diesen Wert, so ist wegen (41)

$$\frac{d}{dx}(f - F_a)\Big|_{x=\xi} = f'(\xi) - a_1 = 0$$

und außerdem

$$\frac{d^2}{dx^2}(f - F_a)\Big|_{x=\xi} = f''(\xi) > 0,$$

d. h., die Abweichung $f - F_a$ nimmt auf $[\![a, b]\!]$ bei ξ ein relatives Minimum an, sofern a_1 gemäß (41) bestimmt ist. Da die Ableitung von $f - F_a$ außer bei ξ nirgends im Intervall $]a, b[$ verschwindet, wird das absolute Minimum von $f - F_a$ bei ξ oder an den Intervallenden a, b angenommen. Daraus folgt: $|f - F_a|$ wird bei ξ oder a, b maximal. Gilt also

$$|f(\xi) - F_a(\xi)| = |f(a) - F_a(a)| = |f(b) - F_a(b)|, \tag{42}$$

so sind diese Größen der Tschebyscheff-Abweichung L gleich.

Nunmehr wird a_0 aus (40a), (40b) zu

$$a_0 = \frac{1}{2}[f(a) + f(\xi) - a_1(a + \xi)] \tag{43}$$

bestimmt. Dann ist für die entsprechende Funktion (39)

$$\begin{aligned} f(\xi) - F_a(\xi) &= f(\xi) - \frac{1}{2}[f(a) + f(\xi) - a_1(a + \xi)] - a_1\xi \\ &= \frac{1}{2}\left\{f(\xi) - f(a) - \frac{f(b) - f(a)}{b - a}(\xi - a)\right\} \\ &= -\big(f(a) - F_a(a)\big) \\ &= \frac{1}{2}\left\{f(\xi) - f(b) - \frac{f(b) - f(a)}{b - a}(\xi - b)\right\} \\ &= -\big(f(b) - F_a(b)\big). \end{aligned}$$

Es gilt also (42) und (40) mit L als Tschebyscheff-Abweichung von f und $F_{\boldsymbol{a}}$. Die mit (41) und (43) gebildete Funktion (39) ist folglich Minimallösung. Man verifiziert leicht, daß deren Graph durch die in Abb. 5.1 angegebene Konstruktion bestimmt ist.

Zur Beurteilung der mit der besten gleichmäßigen Approximation über der Menge der Linearkombinationen aus Funktionen eines Tschebyscheff-Systems erreichbaren Annäherung ist folgende *Ungleichung von de la Vallée-Poussin* nützlich.

Satz 5. *Es sei* $\{\varphi_i\}$, $i = 1(1)n$, *ein Tschebyscheff-System bezüglich* $[\![a, b]\!]$, $f \in C_{[a,b]}$, $F_{\boldsymbol{a}} = \sum_{i=1}^{n} a_i\varphi_i$ *und speziell* $F_{\boldsymbol{a}^*}$ *die Bestapproximation von* f *in der Menge dieser Linearkombinationen. Dann gilt:*

Nimmt die Differenz $f - F_{\boldsymbol{a}}$ *über* $n + 1$ *Argumenten*

$$x_1 < x_2 < \cdots < x_{n+1} \text{ des Intervalls } [\![a, b]\!] \tag{44}$$

Werte mit wechselndem Vorzeichen an, d. h. ist

$$[f(x_j) - F_{\boldsymbol{a}}(x_j)]\,[f(x_{j+1}) - F_{\boldsymbol{a}}(x_{j+1})] < 0, \qquad j = 1(1)n,$$

so folgt

$$Z(\boldsymbol{a}) \geqq Z(\boldsymbol{a}^*) \geqq m \tag{45}$$

mit $Z(\boldsymbol{a}) = \|f - F_{\boldsymbol{a}}\|$ *und*

$$m := \min_{j=1(1)n} \{|f(x_j) - F_{\boldsymbol{a}}(x_j)|\}. \tag{46}$$

Beweis. Offensichtlich ist $Z(\boldsymbol{a}) \geqq Z(\boldsymbol{a}^*)$. Im Sinne eines indirekten Beweises wird nun angenommen, daß

$$Z(\boldsymbol{a}^*) < m \tag{47}$$

ist. Aus (47) folgt $\boldsymbol{a}^* \neq \boldsymbol{a}$ und

$$|f(x_j) - F_{\boldsymbol{a}^*}(x_j)| < |f(x_j) - F_{\boldsymbol{a}}(x_j)| \quad \text{für } j = 1(1)n.$$

Auf Grund dessen wechselt die nichttriviale Linearkombination

$$F_{\boldsymbol{a}^*} - F_{\boldsymbol{a}} = (f - F_{\boldsymbol{a}}) - (f - F_{\boldsymbol{a}^*})$$

über der Menge der Punkte (44) das Vorzeichen und besitzt daher auf $[\![a, b]\!]$ mindestens n Nullstellen. Dem widerspricht, daß $\{\varphi_i\}$ ein Tschebyscheff-System ist.

5.3.2. Anwendungen und Beispiele

1. *Polynome, die über einem Intervall am wenigsten von Null abweichen.* Wir betrachten sämtliche Polynome P n-ten Grades mit dem Leitkoeffizienten 1 und stellen diese in der Form $P(x) = x^n - Q_{n-1}(x)$ dar, wobei Q_{n-1} ein Polynom maximal $(n - 1)$-ten

Grades bedeutet. Gefragt wird nach der Existenz einer Funktion in dieser Menge, welche im Sinne der Norm 5.3.(1) über dem Intervall $[\![-1, 1]\!]$ am wenigsten von Null abweicht. Das Problem läßt sich äquivalent so formulieren: In der n-parametrigen Schar der Polynome

$$Q_{n-1}(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$$

ist die beste gleichmäßige Approximation der Potenz $f: x \mapsto x^n$ gesucht. Da Q_{n-1} Linearkombination der Funktionen des in Beispiel 1 betrachteten Tschebyscheff-Systems ist, hat die Aufgabe nach Satz 3 genau eine Lösung. Um diese zu bestimmen, wird das normierte Tschebyscheff-Polynom n-ten Grades $\hat{T}_n$ (vgl. 5.2.3., Beispiel 4) mit dem Leitkoeffizienten 1 betrachtet. Auf Grund von 5.2.3., Satz 6, ist

$$\hat{T}_n = \frac{1}{2^{n-1}} T_n, \tag{48}$$

und $\hat{T}_n$ nimmt an den Stellen 5.2.3.(49)

$$\bar{x}_l^{(n)} = \cos\frac{l\pi}{n}, \qquad l = 0, 1, \ldots, n,$$

die Werte $\frac{(-1)^l}{2^{n-1}}$ als absolute Extrema über $[\![-1, 1]\!]$ an, d. h., $\hat{T}_n$ alterniert auf diesem Intervall n-mal. Wird

$$\hat{T}_n(x) = x^n - Q_{n-1}^*(x)$$

gesetzt, so erweist sich Q_{n-1}^* auf Grund von 5.3.1., Satz 4, als die gesuchte Bestapproximation von x^n. Damit gleichbedeutend ist

Hilfssatz 4. *$\hat{T}_n$ ist in der Menge der Polynome n-ten Grades mit dem Leitkoeffizienten 1 das über dem Intervall $[\![-1, 1]\!]$ am wenigsten von Null abweichende Polynom.*

Wir verallgemeinern die Problemstellung und suchen in der Menge der Polynome n-ten Grades mit dem Leitkoeffizienten A dasjenige, welches über dem Intervall $[\![a, b]\!]$ am wenigsten von Null abweicht. Durch die lineare Transformation

$$t = \frac{b-a}{2}x + \frac{a+b}{2}, \qquad x = \frac{2t-(a+b)}{b-a} \tag{49}$$

wird eine eineindeutige Abbildung der Intervalle $-1 \leqq x \leqq 1$ und $a \leqq t \leqq b$ aufeinander vermittelt. Der nach Potenzen von t entwickelte Ausdruck $\hat{T}_n\left(\frac{2t-(a+b)}{b-a}\right)$ stellt daher ein Polynom n-ten Grades mit dem Leitkoeffizienten $\frac{2^n}{(b-a)^n}$ dar, das auf $[\![a, b]\!]$ n-mal alterniert;

$$\hat{T}_n: t \mapsto A\,\frac{(b-a)^n}{2^n}\,\hat{T}_n\left(\frac{2t-(a+b)}{b-a}\right) \tag{50}$$

ist ein solches Polynom mit dem Leitkoeffizienten A. Wie in dem Spezialfall des Hilfssatzes 4 ($A = 1$, $[\![a, b]\!] = [\![-1, 1]\!]$) begründet man die Aussage

Satz 6. *Unter allen Polynomen n-ten Grades mit dem Leitkoeffizienten A ist* (50) *das auf* $[\![a, b]\!]$ *am wenigsten von Null abweichende. Dabei ist*

$$\|\hat{T}_n\| = \max_{t\in[a,b]} |\hat{T}(t)| = A\,\frac{(b-a)^n}{2^{2n-1}}. \tag{51}$$

(51) folgt mit (49) und (50) aus $\max\limits_{x\in(-1,1)} |\hat{T}_n(x)| = \dfrac{1}{2^{n-1}}$.

2. *Trigonometrische Polynome minimaler Abweichung von Null.* Gegeben seien reelle Zahlen A, B, die nicht zugleich verschwinden. Wir betrachten sämtliche trigonometrischen Polynome der Form

$$\begin{aligned} TP(x) &= A\cos nx + B\sin nx \\ &\quad - \Big[a_{n-1}\cos(n-1)x + b_{n-1}\sin(n-1)x + \cdots + a_1\cos x \\ &\quad + b_1\sin x + \frac{a_0}{2}\Big] \end{aligned} \tag{52}$$

und suchen unter diesen dasjenige, welches über einem Periodenintervall (und damit auf der ganzen x-Achse) am wenigsten von Null abweicht. Damit gleichbedeutend ist die Forderung, für

$$f(x) = A\cos nx + B\sin nx$$

die beste gleichmäßige Approximation in der Menge der trigonometrischen Polynome

$$F(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n-1}(a_\nu\cos\nu x + b_\nu\sin\nu x) \tag{53}$$

zu bestimmen. Die eindeutige Lösbarkeit der Aufgabe folgt aus Satz 3, da (53) Linearkombination der Funktionen des in Beispiel 3 ($p = 2\pi$) untersuchten Tschebyscheff-Systems ist. Die Funktion $f(x)$ gestattet die Darstellung

$$\begin{aligned} f(x) &= A\cos nx + B\sin nx \\ &= \sqrt{A^2+B^2}\left[\frac{A}{\sqrt{A^2+B^2}}\cos nx + \frac{B}{\sqrt{A^2+B^2}}\sin nx\right] \\ &= \sqrt{A^2+B^2}\,[\cos n\alpha\cos nx + \sin n\alpha\sin nx] \\ &= \sqrt{A^2+B^2}\,\cos n(x-\alpha), \end{aligned} \tag{54}$$

wenn α gemäß

$$\cos n\alpha = \frac{A}{\sqrt{A^2+B^2}}, \qquad \sin n\alpha = \frac{B}{\sqrt{A^2+B^2}}$$

im Intervall $0 < \alpha \leqq \frac{2\pi}{n}$ bestimmt wird. Die Funktion (54) nimmt an den $2n$ Stellen

$$x_k = \alpha + \frac{\pi}{n} k, \qquad k = 0, \pm 1, \pm 2, \ldots, \pm(n-1), n, \tag{55}$$

des halboffenen Periodenintervalls $]-\pi + \alpha, \pi + \alpha]$ absolute Extremwerte vom Betrag $\sqrt{A^2 + B^2}$ mit wechselndem Vorzeichen an, alterniert also über diesem mindestens $(2n-1)$-mal. Daraus folgt nach Satz 4, daß die mit durchweg verschwindenden Koeffizienten gebildete Linearkombination (53) die Funktion f am besten approximiert oder:

Satz 7. *Unter allen trigonometrischen Polynomen* (52) *weicht*

$$f(x) = A \cos nx + B \sin nx, \qquad A^2 + B^2 > 0,$$

über der Zahlgeraden am wenigsten von Null ab. Auf Grund von (54) *ist*

$$\|f\| = \sqrt{A^2 + B^2}.$$

3. *Minimierung des Restgliedes bei der Interpolation mit ganzen rationalen Funktionen.* Es sei f eine auf $[\![a, b]\!]$ n-mal stetig differenzierbare Funktion, deren $(n+1)$-te Ableitung in $]a, b[$ existiert und beschränkt ist. P bedeute das über den Abszissen $x_j \in [\![a, b]\!]$, $j = 0(1)n$, mit den Werten von f konstruierte Interpolationspolynom maximal n-ten Grades. Für das durch

$$f(x) = P(x) + R_n(x)$$

definierte Restglied R_n gilt nach MfL Bd. 9, 4.2.4.(29),

$$R_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\, \omega_n(x),$$

wobei

$$\omega_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

und

$$\min\{x, x_0, x_1, \ldots, x_n\} < \xi_x < \max\{x, x_0, x_1, \ldots, x_n\}$$

ist. Mit M_{n+1} als einer oberen Schranke für $|f^{(n+1)}(x)|$ auf $]a, b[$ folgt daraus

$$|R_n(x)| \leqq \frac{M_{n+1}}{(n+1)!}\, |\omega_n(x)|. \tag{56}$$

Im Hinblick auf den Faktor $|\omega_n(x)|$ in dieser Abschätzung wird nun nach einer Verteilung der Interpolationsknoten x_j, $j = 0(1)n$, über dem Intervall $[\![a, b]\!]$ gefragt, bei welcher der Betrag des Restgliedes minimal wird, d. h. $\omega_n(x)$ am wenigsten von Null abweicht. Die Antwort lautet auf Grund von Satz 6 ($A = 1$):

Satz 8. *Die rechte Seite der Restgliedabschätzung* (56) *wird genau dann minimal, wenn die Interpolationsknoten als die Nullstellen des Polynoms* $\hat{T}_{n+1}$ (vgl. (50)) *gewählt werden.*

4. *Minimierung des Abbruchfehlers bei Potenzreihen.* Wir erläutern diese Problematik wie folgt: Für eine gegebene Funktion f sei über dem Intervall $[\![a, b]\!]$

$$\begin{aligned} f(x) &= P_n(x) + R_n(x), \\ P_n(x) &= a_0 + a_1 x + \cdots + a_n x^n, \qquad a_n \neq 0, \end{aligned} \tag{57}$$

d. h., R_n bezeichnet den Fehler bei der Ersetzung von f durch das Polynom P_n. Um den Rechenaufwand zu verringern, wird ein Polynom niedrigeren Grades gesucht, das an Stelle von P_n eine möglichst geringe Verschlechterung der Approximationsgüte verursacht.

Der Einfachheit halber sei $[\![a, b]\!] = [\![-1, 1]\!]$. Dann ergibt sich die Aufgabe, P_n auf $[\![-1, 1]\!]$ durch ein Polynom maximal $(n-1)$-ten Grades möglichst gut gleichmäßig zu approximieren. Offenbar kann P_n in der Form

$$P_n(x) = b_0 + b_1 x + \cdots + b_{n-1} x^{n-1} + c_n T_n(x), \qquad c_n \neq 0, \tag{58}$$

dargestellt werden, wobei T_n das Tschebyscheff-Polynom 5.2.(44) bezeichnet, welches nach 5.2., Satz 6, den Leitkoeffizienten 2^{n-1} besitzt. Mit

$$T_n(x) = \sum_{j=0}^{n} d_{nj} x^j, \qquad d_{nn} = 2^{n-1},$$

ergibt sich offenbar

$$c_n = \frac{a_n}{d_{nn}} = \frac{a_n}{2^{n-1}}, \qquad b_i = a_i - c_n d_{ni}, \qquad i = 0(1)n-1. \tag{59}$$

Das in (58) auftretende Polynom

$$P(x) = b_0 + b_1 x + \cdots + b_{n-1} x^{n-1} \tag{60}$$

leistet das Gewünschte, da

$$P_n - P = c_n T_n$$

auf $[\![-1, 1]\!]$ n-mal alterniert und P somit nach Satz 4 Bestapproximation von P_n in der Menge der Polynome vom Grade $\leqq n-1$ ist.[1])

Die Nützlichkeit der Methode sei an einer [37] entnommenen Aufgabe erläutert: Mit Hilfe eines aus der Taylorentwicklung von

$$f\colon x \mapsto \sin \frac{\pi}{2} x$$

gemäß (57) bis (60) hergeleiteten Polynoms sollen Sinuswerte mit 7 gültigen Ziffern berechnet werden.

---

[1]) Die Ersetzung von P_n durch P gemäß (58) und (60) wird in der englischsprachigen Literatur gelegentlich „telescoping a polynomial" genannt.

Auf Grund der Periodizitäts- und Symmetrieeigenschaften von f genügt es, das Intervall $0 \leqq x \leqq 1$ zu betrachten. Es ist

$$\sin\frac{\pi}{2}x = \frac{\pi}{2}x - \frac{1}{3!}\left(\frac{\pi}{2}\right)^3 x^3 + \frac{1}{5!}\left(\frac{\pi}{2}\right)^5 x^5 \pm \cdots,$$

und bei Abbruch nach der Potenz x^{2k-1} gilt für das Restglied nach MfL Bd. 4, 2.2.3., Satz 2,

$$|R| \leqq \frac{\left(\frac{\pi}{2}\right)^{2k+1} |x|^{2k+1}}{(2k+1)!}.$$

Wird $\sin\frac{\pi}{2}x$ näherungsweise mit Hilfe des Polynoms

$$P_{2k-1}(x) = \frac{\pi}{2}x - \frac{1}{3!}\left(\frac{\pi}{2}\right)^3 x^3 \pm \cdots + (-1)^{k+1}\frac{1}{(2k-1)!}\left(\frac{\pi}{2}\right)^{2k-1} x^{2k-1} \tag{61}$$

berechnet, so ergibt sich für den Betrag des verfahrensbedingten relativen Fehlers δ gemäß MfL Bd. 9, 2.5.(3),

$$|\delta(x)| \leqq \frac{\left(\frac{\pi}{2}\right)^{2k+1} x^{2k+1}}{(2k+1)!\sin\frac{\pi}{2}x}, \qquad -1 \leqq x \leqq 1.$$

Man verifiziert leicht, daß $1 \leqq \frac{x}{\sin x} \leqq \frac{\pi}{2}$ für $x \in \left[\!\left[0, \frac{\pi}{2}\right]\!\right]$ gilt, und erhält schließlich

$$|\delta(x)| \leqq \frac{\left(\frac{\pi}{2}\right)^{2k+1}}{(2k+1)!}, \qquad -1 \leqq x \leqq 1. \tag{62}$$

Aus (62) folgt für $k = 7$

$$|\delta(x)| < \frac{1}{2} \cdot 10^{-8},$$

und gemäß MfL Bd. 9, 2.5.(10), sind dann bei Vernachlässigung von Rundungsfehlern acht Ziffern des Näherungswertes gültig. Betrachtet man an Stelle von P_{13} gemäß (61) das Näherungspolynom P_{11}, so ist etwa für $x = 1$ bereits die siebente Dezimale nicht mehr sicher. In der Tat gilt für das vernachlässigte Glied

$$\frac{1}{13!}\left(\frac{\pi}{2}\right)^{13} \approx 5{,}7 \cdot 10^{-8}. \tag{63}$$

Nun reduzieren wir P_{13} gemäß (58) durch Abspaltung des Tschebyscheff-Termes und bestimmen etwa mit Hilfe der Rekursionsformel 5.2.(63) und 5.2.(43)

$$T_{13}(x) = 4096x^{13} - 13312x^{11} + 16640x^9 - 9984x^7 + 2912x^5 - 364x^3 + 13x.$$

Damit liefert (59)

$$c_{13} = 1{,}38969 \cdot 10^{-11},$$
$$b_0 = b_2 = \cdots = b_{12} = 0,$$
$$b_1 = 1{,}5707963, \quad b_3 = -0{,}64596410, \quad b_5 = 0{,}079692580,$$
$$b_7 = -0{,}0046816154, \quad b_9 = 0{,}00016020995, \quad b_{11} = -0{,}0000034138477;$$

die Koeffizienten b_i bis b_{11} sind mit acht wesentlichen Ziffern angegeben, die alle gültig sind. Wir vergleichen das gemäß (60) mit den Koeffizienten b_i, $i = 1(2)11$, gebildete Polynom P mit dem Polynom P_{11}, das aus P_{13} durch Weglassen des höchsten Gliedes entsteht. Wegen

$$P_{13} - P = c_{13}T_{13} \quad \text{und} \quad |T_{13}(x)| \leqq 1 \quad \text{für } x \in [\![-1, 1]\!]$$

ist

$$\|P_{13} - P\| := \max_{x \in [-1,1]} |P_{13}(x) - P(x)| \leqq 1{,}39 \cdot 10^{-11},$$

d. h., diese Abweichung ist auf Grund von (63) kleiner als $\frac{1}{4000}\|P_{13} - P_{11}\|$. Die gewünschte siebenstellige Genauigkeit bei der Berechnung von $\sin\frac{\pi}{2}x$ ist also mit dem Polynom P zu erreichen.

Das Abspalten von Tschebyscheff-Termen gemäß (57) und (58) läßt sich unter Beachtung einer Genauigkeitsforderung unter Umständen wiederholt vornehmen. Bei einem Polynom

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

so oft wie möglich angewandt, führt dieser Prozeß auf die Umordnung von $P_n(x)$ nach Tschebyscheff-Polynomen:

$$P_n = c_n T_n + c_{n-1}T_{n-1} + \cdots + c_1 T_1 + c_0. \tag{64}$$

Auf Grund der Gleichung (64) ist $\sum_{i=0}^{n} c_i T_i$ natürlich auch bezüglich des Skalarproduktes 5.2.(50) die beste Quadratmittelapproximation von P_n in der Menge aller Linearkombinationen der Polynome $T_0, T_1, \ldots, T_n$, wenn diese als Elemente des Raumes $C_{[-1,1]}$ aufgefaßt werden. Nach 5.2.3. sind die c_i, $i = 0(1)n$, eindeutig als die Fourierkoeffizienten 5.2.(54) der Funktion $F(t) = P_n(\cos t)$ bestimmt. Aus algorithmischen Gründen ist es zweckmäßig, diese durch Fourierentwicklung der in P_n auftretenden Potenzen x^j nach den T_i zu konstruieren. Setzt man

$$x^j = \frac{1}{2}c_{j0} + \sum_{i=1}^{j} c_{ji}T_i, \tag{65}$$

so gilt nach 5.2.(54) einheitlich

$$c_{ji} = \frac{2}{\pi}\int_0^{\pi} \cos^j t \cos it \, dt. \tag{66}$$

Mit

$$\cos it = \cos(i-1)t\cos t - \sin(i-1)t\sin t$$

gewinnt man daraus

$$c_{ji} = \frac{2}{\pi}\int_0^{\pi} \cos^{j+1} t \cos(i-1)t \, dt - \frac{2}{\pi}\int_0^{\pi} \cos^j t \sin t \sin(i-1)t \, dt.$$

Partielle Integration des zweiten Integrals liefert mit

$$\frac{dv(t)}{dt} = \cos^j t \sin t, \qquad v = -\frac{1}{j+1}\cos^{j+1} t, \qquad u(t) = \sin(i-1)t$$

die Beziehung

$$\int_0^{\pi} \cos^j t \sin t \sin(i-1)t \, dt = \frac{i-1}{j+1}\int_0^{\pi} \cos^{j+1} t \cos(i-1)t \, dt.$$

Insgesamt ergibt sich damit die Rekursionsformel

$$c_{ji} = \frac{2}{\pi}\left[1 - \frac{i-1}{j+1}\right]\int_0^\pi \cos^{j+1} t \cos(i-1)\,t\,dt = \left[1 - \frac{i-1}{j+1}\right] c_{j+1,i-1}.$$

Wenn c_{j0} für beliebiges j bekannt ist, lassen sich damit alle benötigten c_{ji} berechnen. Man findet mit Hilfe partieller Integration für $j \geqq 2$

$$\int_0^\pi \cos^j t\,dt = \int_0^\pi \cos^{j-1} t \cos t\,dt = (j-1)\int_0^\pi \cos^{j-2} t \sin^2 t\,dt$$
$$= -(j-1)\int_0^\pi \cos^j t\,dt + (j-1)\int_0^\pi \cos^{j-2} t\,dt,$$

also

$$\int_0^\pi \cos^j t\,dt = \frac{j-1}{j}\int_0^\pi \cos^{j-2} t\,dt$$

und gemäß (66)

$$c_{j0} = \frac{j-1}{j}\,c_{j-2,0}.$$

Offenbar ist $c_{00} = 2$ und $c_{10} = 0$. Damit haben wir folgendes Rekursionsschema zur Berechnung der c_{ji}:

$$\begin{aligned} &c_{00} = 2, \qquad c_{10} = 0, \\ &c_{j0} = \frac{j-1}{j}\,c_{j-2,0}, \qquad j = 2(1)2n, \\ &c_{ji} = \left(1 - \frac{i-1}{j+1}\right) c_{j+1,i-1}, \qquad j = i(1)2n - i. \end{aligned} \tag{67}$$

Durch Einsetzen von (65) in das Polynom P_n findet man

$$P_n(x) = \sum_{j=0}^{n} a_j x^j = \sum_{j=0}^{n} a_j\left[\frac{1}{2}c_{j0} + \sum_{i=1}^{j} c_{ji} T_i(x)\right]$$
$$= \frac{1}{2}\sum_{j=0}^{n} a_j c_{j0} + \sum_{i=1}^{j}\left(\sum_{j=0}^{n} a_j c_{ji}\right) T_i(x),$$

also im Vergleich mit (64)

$$\begin{aligned} c_0 &= \frac{1}{2}\sum_{j=0}^{n} a_j c_{j0}, \\ c_i &= \sum_{j=0}^{n} a_j c_{ji} = \sum_{j=1}^{n} a_j c_{ji}, \qquad i = 1(1)n. \end{aligned} \tag{68}$$

Wir fassen die Berechnung der c_i in einer ALGOL-Prozedur *CHEBY(A,C,n)* zusammen, deren formale Parameter folgende Bedeutung haben:

A Feld der Koeffizienten des nach den T_i umzuordnenden Polynoms,
n dessen Grad und
C Feld der in (64) auftretenden Koeffizienten.

Lokal wird ein Feld $B[0:2n]$ zur Berechnung der c_{ji} vereinbart. Diese erfolgt gemäß (67) spaltenweise nach dem Schema der Tabelle 5.4. Die in dem schraffierten Teil des Schemas liegenden Spaltenabschnitte gehen in die Koeffizientenberechnung (68) ein.

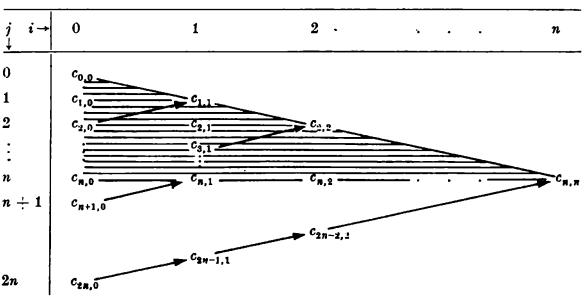


Tabelle 5.4

```
procedure CHEBY(A,C,n); value n; integer n; array A,B;
begin
  integer nn;
  nn := n + n;
    begin
      integer i,j,i1,j1; array B[0:nn];
      B[0] := 2; B[1] := 0; C[0] := 0;
      for j := 2 step 1 until n do B[j] := B[j − 2] × (j − 1)/j;
      for j := 0 step 1 until n do C[0] := C[0] + A[j] × B[j];
      C[0] := C[0] × 0.5;
      for i := 1 step 1 until n do  begin C[i] := 0; i1 := i − 1;
      for j := i step 1 until nn − i do
      begin j1 := j + 1; B[j] := (1 − i1/j1) × B[j1] end;
      for j := i step 1 until n do
      C[i] := C[i] + B[j] × A[j]    end;
    end
end
```

Bei der automatischen Berechnung kann das Reduktionsverfahren in der folgenden Weise durchgeführt werden:

1. Umordnung eines gegebenen Approximationspolynoms nach Tschebyscheff-Polynomen mit Hilfe der Prozedur *CHEBY*.
2. Weglassen „höherer Glieder", soweit das mit der Genauigkeitsanforderung verträglich ist. Wegen
$$\max_{x\in(-1,1)} |T_i(x)| = 1, \qquad i = 0, 1, 2, \ldots,$$
ist dafür allein die Größenordnung der Koeffizienten in (64) maßgebend.
3. Die Berechnung des reduzierten Polynoms erfordert keine Umordnung nach Potenzen von x. Auf die verbliebene Partialsumme einer Tschebyscheff-Reihe kann der Goertzel-Algorithmus in der Form 5.2.(108), (109) angewendet werden.

5.4. Zu Theorie und Anwendung von Splinefunktionen

5.4.1. Einführende Betrachtungen

Splinefunktionen (kurz auch *Splines* genannt) sind ein spezieller Gegenstand der Approximationstheorie, dessen Bearbeitung wesentlich nach dem zweiten Weltkrieg einsetzte und seitdem sehr ergebnisreich verlaufen ist. Die Untersuchungen konzentrierten sich zunächst auf Interpolationsprobleme und damit zusammenhängende Anwendungen (vgl. [45], 2.6.). Dafür entwickelte Lösungsalgorithmen sind in zahlreiche Verfahren der numerischen Mathematik eingegangen; eine Auswahl für den Spezialfall kubischer Splines (s. u.) wird im folgenden erörtert. Zur Einführung sei an die Interpolation mit ganzen rationalen Funktionen und die damit zusammenhängende Fehlerproblematik erinnert (vgl. MfL Bd. 9, 4.2.4.).

Dem Vorzug, beliebig oft differenzierbar zu sein, steht der Nachteil eines Interpolationspolynoms entgegen, mit wachsendem Grad, d. h. wachsender Knotenzahl, stark zu schwanken und so unter Umständen große Abweichungen von den Werten der zu interpolierenden Funktion zwischen den Knoten zu verursachen. (Man studiere in dieser Hinsicht noch einmal das Beispiel der Abb. 4.12 in MfL Bd. 9, 4.2.4.) Es liegt daher nahe, die Interpolationsaufgabe für ein System von Knoten x_i

$$a < x_1 < x_2 < \cdots < x_n < b$$

und eine Funktion $f: [\![a, b]\!] \to \mathbf{R}$ nicht durch ein Polynom zu lösen, sondern dafür — nach einer Zerlegung von $[\![a, b]\!]$ in Teilintervalle — über diesen Teilintervallen Polynome entsprechend niedrigeren Grades anzusetzen und die Glattheit eines Interpolationspolynoms eingeschränkt durch die Forderung nachzubilden, daß für diese stückweise polynomiale Funktion alle Ableitungen bis zu einer gewissen Ordnung stetig sein sollen. Aus solchen Vorstellungen hat sich der Begriff der Splinefunktion herausgebildet. Die Bezeichnung stammt von I. J. Schoenberg — einem Begründer der Theorie — und ist der englische Name eines Gerätes zum Zeichnen von Kurven, bei welchem die Anpassung an ein System zu verbindender Punkte durch ein an diskreten Stellen belastetes Stahlband realisiert wird. Die Kurve der Durchbiegung erweist sich als stückweise aus kubischen Parabeln zusammengesetzt. Wir präzisieren die Begriffsbildung in der folgenden

Definition 1. $S: \mathbf{R} \to \mathbf{R}$ ist eine *Splinefunktion m-ten Grades* mit den Knoten $x_i \in \mathbf{R}$, $i = 1(1)n$, $x_1 < x_2 < \cdots < x_n$, genau dann, wenn mit $x_0 = -\infty$ und $x_{n+1} = \infty$

a) S in jedem der Intervalle $]\!]x_i, x_{i+1}[\![$, $i = 0(1)n$, durch ein Polynom maximal m-ten Grades darstellbar ist und

b) S auf $\mathbf{R}$ stetige Ableitungen bis zur Ordnung $m - 1$ besitzt.

Bemerkung 1. Wenn $m = 0$ ist, wird b) gegenstandslos. Eine Splinefunktion nullten Grades ist eine im allgemeinen in den Knoten unstetige stückweise konstante Funktion.

Bemerkung 2. Man beachte, daß S in den Intervallen $]x_i, x_{i+1}[$ im allgemeinen durch verschiedene Polynome dargestellt wird. Dennoch genügen auch alle Polynome vom Grade $\leqq m$ der Definition 1 und sind spezielle Splinefunktionen.

Im Zusammenhang mit Interpolationsproblemen werden sogenannte *natürliche Splinefunktionen* eine Rolle spielen, die durch folgende Spezialisierung der Definition 1 charakterisiert sind:

Definition 2. Eine der Definition 1 genügende Funktion $s: \mathbf{R} \to \mathbf{R}$[1]) ist eine *natürliche Splinefunktion* genau dann, wenn der Grad m ungerade ist und s für $m = 2k - 1$ in den Intervallen $]-\infty, x_1[$ und $]x_n, \infty[$ durch je ein Polynom vom Grade $\leqq k - 1$ dargestellt wird.

Von grundsätzlicher Bedeutung für die Theorie sind die modifizierten Potenzen

$$x_+^m := \begin{cases} x^m & \text{für } x > 0,^{2}) \\ 0 & \text{für } x \leqq 0, \end{cases} \qquad m \in \mathbf{N}. \tag{1}$$

Offenbar ist $F: x \mapsto x_+^m$ eine Splinefunktion m-ten Grades mit dem Knoten $x_1 = 0$. Die Ableitungen bis zur Ordnung $m - 1$ lassen sich nach der Differentiationsregel für die Potenz x^m bilden. Beispielsweise ist für $m > 1$

$$\frac{dF(x)}{dx} = m x_+^{m-1}. \tag{2}$$

Für $x \neq 0$ gewinnt man

$$\frac{d^m F(x)}{dx^m} = m! x_+^0, \tag{3}$$

wobei x_+^0 die in Abb. 5.10 dargestellte sogenannte *Heaviside-Funktion* ist. Die Funktion

$$F_c: x \mapsto (x - c)_+^m, \qquad c \in \mathbf{R}, \tag{4}$$

erweist sich entsprechend als Splinefunktion m-ten Grades mit dem Knoten c.

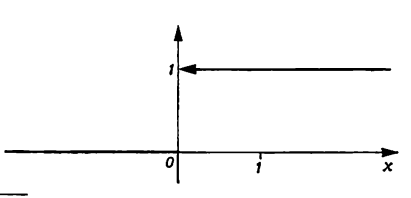


Abb. 5.10

[1]) Für natürliche Splinefunktionen wird im folgenden stets ein Kleinbuchstabe als Funktionssymbol gewählt.

[2]) In der englischsprachigen Literatur als „truncated-power"-Größe bezeichnet.

Es wird im folgenden Abschnitt zu zeigen sein, daß sich jede Splinefunktion m-ten Grades als eine Linearkombination von Funktionen des Typs (4) mit verschiedenen Werten für c darstellen läßt. Wegen dieses für die Theorie grundlegenden Resultates werden die Funktionen (4) auch *elementare Splinefunktionen* genannt.

5.4.2. Darstellung durch elementare Splinefunktionen

Es sei $\mathfrak{S}_m(x_1, x_2, \ldots, x_n)$ die Gesamtheit der Splinefunktionen m-ten Grades mit den Knoten x_i, $i = 1(1)n$, $x_1 < x_2 < \cdots < x_n$, und $\mathfrak{P}_m$ die Klasse der Polynome vom Grade $\leqq m$. Dann gilt

Satz 1. *Für jedes $S \in \mathfrak{S}_m(x_1, x_2, \ldots, x_n)$ gibt es genau ein System nichtverschwindender reeller Zahlen $c_1, c_2, \ldots, c_n$ und genau ein Polynom $P_m \in \mathfrak{P}_m$ derart, daß für alle $x \in \mathbf{R}$*

$$S(x) = P_m(x) + \sum_{i=1}^{n} c_i (x - x_i)_+^m \tag{5}$$

gilt. P_m ist die Polynomdarstellung von S im Intervall $]-\infty, x_1[$.

Die rechte Seite von (5) definiert, wenn $P_m \in \mathfrak{P}_m$ und $x_i, c_i \in \mathbf{R}$, $i = 1(1)n$, $x_1 < x_2 < \cdots < x_n$, beliebig gewählt werden, eine Splinefunktion der Klasse $\mathfrak{S}_m(x_1, x_2, \ldots, x_n)$.

Folgerung. *Jede natürliche Splinefunktion $s \in \mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$ gestattet eine wohlbestimmte Darstellung der Form*

$$s(x) = P_{k-1}(x) + \sum_{i=1}^{n} c_i (x - x_i)_+^{2k-1} \tag{6}$$

mit $P_{k-1} \in \mathfrak{P}_{k-1}$.

Beweis. Es sei $P_{mi} \in \mathfrak{P}_m$ die Polynomdarstellung von S im Intervall $]x_i, x_{i+1}[$, $i = 0(1)n$, die für $m > 0$ aus Stetigkeitsgründen auch noch in den Knotenendpunkten gilt. Weiterhin wird dies angenommen und der Trivialfall $m = 0$ dem Leser zur gesonderten Betrachtung überlassen. Dann ist $P_{mi} - P_{m,i-1}$, $i = 1(1)n$, ein Polynom der Klasse $\mathfrak{P}_m$, das auf Grund von Definition 1 b) bei x_i eine Nullstelle der Ordnung m besitzt. Somit gilt für alle x

$$P_{mi}(x) - P_{m,i-1}(x) = c_i (x - x_i)^m. \tag{7}$$

Aus der Identität

$$P_{ml} - P_{m0} = (P_{ml} - P_{m,l-1}) + (P_{m,l-1} - P_{m,l-2}) + \cdots + (P_{m1} - P_{m0}),$$

$l = 1, 2, \ldots, n$, folgt mit (7)

$$P_{ml}(x) = P_{m0}(x) + \sum_{i=1}^{l} c_i (x - x_i)^m \tag{8}$$

und weiter

$$S(x) = P_{m0}(x) + \sum_{i=1}^{n} c_i (x - x_i)^m. \tag{9}$$

In der Tat: Für $-\infty < x < x_1$ reduziert sich die rechte Seite von (9) auf $P_{m0}(x)$ und stimmt daher in diesem Intervall mit $S(x)$ überein. Entsprechend gewinnt man für $x \in [\![x_l, x_{l+1}[\![$

$$P_{m0}(x) + \sum_{i=1}^{l} c_i(x - x_i)^m,$$

und wegen (8) stellt dieser Ausdruck wieder $S(x)$ dar. Nach Umbenennung des Polynoms P_{m0} in P_m geht (9) in (5) über, und P_m ist die Polynomdarstellung von S im Intervall $]\!]-\infty, x_1[\![$.

Wir betrachten nun die durch die rechte Seite von (5) definierte Funktion, wenn $P_m \in \mathfrak{P}_m$ und $x_i, c_i \in \mathbf{R}$, $i = 1(1)n$, $x_1 < x_2 < \cdots < x_n$, beliebig gewählt wurden. Diese sei wieder mit S bezeichnet. Auf Grund von (1) erkennt man sofort, daß S in den Intervallen $]\!]x_i, x_{i+1}[\![$, $i = 0(1)n$, ein Polynom der Klasse $\mathfrak{P}_m$ darstellt; speziell ist $S(x) = P_m(x)$ in $]\!]-\infty, x_1[\![$. Gemäß Definition 1 bleibt noch zu zeigen, daß S auf $-\infty < x < \infty$ $(m-1)$-mal stetig differenzierbar ist. Dazu wird eine der Argumentstellen x_j, $j = 1, 2, \ldots, n$, betrachtet. Es ist

$$S(x) = P_m(x) + \sum_{i=1}^{j-1} c_i(x - x_i)^m \qquad \text{in} \quad x_{j-1} < x < x_j \tag{10a}$$

und

$$S(x) = P_m(x) + \sum_{i=1}^{j} c_i(x - x_i)^m \qquad \text{in} \quad x_j < x < x_{j+1}. \tag{10b}$$

Daraus folgt

$$\lim_{x \to x_j - 0} S(x) = P_m(x_j) + \sum_{i=1}^{j-1} c_i(x_j - x_i)^m \tag{11a}$$

und

$$\lim_{x \to x_j + 0} S(x) = P_m(x_j) + \sum_{i=1}^{j-1} c_i(x_j - x_i)^m. \tag{11b}$$

Für das Verschwinden des letzten Summanden in (10b) beim Grenzübergang ist die Voraussetzung $m \geqq 1$ wesentlich.

Unter Beachtung von

$$S(x_j) = P_m(x_j) + \sum_{i=1}^{j-1} c_j(x_j - x_i)^m \tag{12}$$

folgt aus (11) die Stetigkeit von S auf $-\infty < x < \infty$.

Nun betrachten wir S' in den an x_j anschließenden Intervallen. Gemäß (10) und (11) ist

$$S'(x) = P'_m(x) + m \sum_{i=1}^{j-1} c_i(x - x_i)^{m-1} \qquad \text{in} \quad x_{j-1} < x < x_j \tag{13a}$$

und

$$S'(x) = P'_m(x) + m \sum_{i=1}^{j} c_i(x - x_i)^{m-1} \qquad \text{in} \quad x_j < x < x_{j+1}. \tag{13b}$$

Im Fall $m - 1 \geqq 1$ ergibt sich aus (13) analog zu (11)

$$\lim_{x \to x_j \pm 0} S'(x) = P'_m(x_j) + m \sum_{i=1}^{j-1} c_i (x_j - x_i)^{m-1}. \tag{14}$$

Mit Hilfe des Mittelwertsatzes folgt aus (14) auch die Differenzierbarkeit von S bei x_j selbst. In der Tat ist

$$\begin{aligned} S'(x_j) &= \lim_{x \to x_j} \frac{S(x) - S(x_j)}{x - x_j} = \lim_{x \to x_j} S'(\xi_x) \\ &= P'_m(x_j) + m \sum_{i=1}^{j-1} c_i (x_j - x_i)^{m-1} \qquad (x \dots \xi_x \dots x_j). \end{aligned} \tag{15}$$

(14) und (15) ergeben die Stetigkeit von S' auf $-\infty < x < \infty$. So fortfahrend kann man zeigen, daß alle Ableitungen von S bis zur Ordnung $m - 1$ einschließlich stetig sind. Beiläufig ergibt sich dabei

$$S^{(l)}(x) = P_m^{(l)}(x) + \binom{m}{l} l! \sum_{i=1}^{n} c_i (x - x_i)_+^{m-l}, \quad 0 \leqq l \leqq m - 1. \tag{16}$$

Die Unitätsaussage des Satzes 1 folgt aus der Bemerkung, daß die *Parameter der Splinefunktion* (5), das sind die Koeffizienten von P_m und die c_i, $i = 1(1)n$, eindeutig durch S bestimmt sind: Es ist $P_m(x) = S(x)$ für $x < x_1$; auf Grund des Fundamentalsatzes der Algebra würde bereits $P_m(\xi_k) = S(\xi_k)$ für $m + 1$ Argumente ξ_k das Polynom P_m charakterisieren. Ferner gewinnt man auf Grund von (16) durch abermalige Differentiation der $(m - 1)$-ten Ableitung zwischen aufeinanderfolgenden Knoten

$$\lim_{x \to x_j - 0} S^{(m)}(x) = P_m^{(m)}(x_j) + m! \sum_{i=1}^{j-1} c_i,$$

$$\lim_{x \to x_j + 0} S^{(m)}(x) = P_m^{(m)}(x_j) + m! \sum_{i=1}^{j} c_i,$$

also

$$c_j = \frac{1}{m!} \left(\lim_{x \to x_j + 0} S^{(m)}(x) - \lim_{x \to x_j - 0} S^{(m)}(x) \right). \tag{17}$$

Das heißt, die Größen $m! c_j$ sind die Sprünge der m-ten Ableitung von S an den Knoten x_j.

5.4.3. Lineare Approximation durch Splinefunktionen

Wie in 5.4.2. betrachten wir auch weiterhin nur Splines mit festen Knoten

$$a < x_1 < x_2 < \dots < x_n < b. \tag{18}$$

$[\![a, b]\!]$ wird ein *Einschließungsintervall* der Knoten genannt, und es werden, wo es zweckmäßig erscheint, a und b auch mit x_0 bzw. x_{n+1} bezeichnet. Satz 1, Gleichung (5),

lehrt, daß dann $\mathfrak{S}_m(x_1, \ldots, x_n)$ für gegebenes m aus den Linearkombinationen der Funktionen

$$\varphi_j: x \mapsto x^j, \qquad j = 0(1)m,$$

und

$$\varphi_{m+j}: x \mapsto (x - x_j)_+^m, \qquad j = 1(1)n, \tag{19}$$

besteht.

Satz 2. *Die $m + n + 1$ Funktionen* (19) *sind über jedem Einschließungsintervall $[\![a, b]\!]$ linear unabhängig.*

Beweis. Wir betrachten eine auf $[\![a, b]\!]$ identisch verschwindende Linearkombination der Funktionen (19):

$$\sum_{\mu=0}^{m} a_\mu x^\mu + \sum_{i=1}^{n} c_i (x - x_i)_+^m \equiv 0. \tag{20}$$

Für $a \leqq x \leqq x_1$ reduziert sich (20) auf $\sum_{\mu=0}^{m} a_\mu x^\mu \equiv 0$, woraus nach dem Fundamentalsatz der Algebra $a_\mu = 0$ für $\mu = 0(1)m$ folgt. Im Intervall $x_1 < x \leqq x_2$ liefert (20) dann $c_1 (x - x_1)^m \equiv 0$, also $c_1 = 0$. Indem man weiter sukzessive die Intervalle $x_2 < x \leqq x_3$, $x_3 < x \leqq x_4$, ..., $x_{n-1} < x \leqq x_n$ und $x_n < x \leqq b$ betrachtet, ergibt sich so $c_i = 0$ für $i = 1(1)n$. Aus (20) folgt also das Verschwinden aller Kombinationskoeffizienten, d. h. die lineare Unabhängigkeit des Systems (19).

Satz 3. *Die Funktionen* (19) *bilden auf keinem Einschließungsintervall ein Tschebyscheff-System.*

Beweis. Man betrachte eine Linearkombination

$$S(x) = \sum_{\mu=0}^{m} a_\mu x^\mu + \sum_{i=1}^{n} c_i (x - x_i)_+^m, \tag{21}$$

in der $a_\mu = 0$ für $\mu = 0(1)m$ und mindestens eines der c_i verschieden von Null ist. $S(x)$ verschwindet dann auf $[\![a, x_1]\!]$ identisch, was nach 5.3., Definition 1, der Eigenschaft eines Tschebyscheff-Systems widerspricht.

Betrachten wir nun in $C_{(a,b)}$ das lineare Approximationsproblem 5.1.(23) bezüglich des Funktionensystems (19) und einer ausgezeichneten Norm, dann gilt nach 5.1., Satz 2, daß dieses mindestens eine Lösung besitzt. Im Falle der Tschebyscheff-Norm

$$\|f\| := \max_{x \in (a,b)} |f(x)| \tag{22}$$

können wir jedoch die Einzigkeit einer Bestapproximation im Hinblick auf den zuletzt bewiesenen Satz und 5.3., Satz 3, nicht behaupten.

Bei der weiteren Erörterung des Sachverhaltes beziehen wir uns auf folgendes notwendige und hinreichende Kriterium von L. L. Schumaker [45] zur Charakterisierung einer Bestapproximation in $\mathfrak{S}_m(x_1, \ldots, x_n)$, das dem Tschebyscheffschen Alternantensatz für die Linear-

kombinationen eines Tschebyscheff-Systems entspricht. $[\![a, b]\!]$ bedeutet darin ein Einschließungsintervall für die Knoten x_i im Sinne von (18).

Satz 4. *$S \in \mathfrak{S}_m(x_1, \ldots, x_n)$ ist gleichmäßige Bestapproximation einer stetigen Funktion $f \in C_{(a,b)}$ genau dann, wenn eine Alternante* (vgl. 5.3., Definition 2) *X von $f - S$ bezüglich $[\![a, b]\!]$ und ein Intervall $[\![x_i, x_{i+j+1}]\!]$, $0 \leq i \leq i + j + 1 \leq n + 1$, derart existieren, daß die Mächtigkeit von $X \cap [\![x_i, x_{i+j+1}]\!]$ größer oder gleich $i + j + 2$ ist.*

Mit Hilfe des Satzes 4, auf dessen Beweis hier nicht eingegangen werden kann, lassen sich stetige Funktionen konstruieren, die mehrere Bestapproximationen in $\mathfrak{S}_m(x_1, \ldots, x_n)$ besitzen. Wir betrachten ein [45] entnommenes Beispiel.

Es sei $[\![a, b]\!] = [\![-1, 1]\!]$, $m = n = 1$, ferner $\mathfrak{S}_m(x_1, \ldots, x_n) = \mathfrak{S}_1(0)$ und die zu approximierende Funktion durch

$$f(x) = \begin{cases} 1 & \text{für } -1 \leq x \leq -\frac{1}{2}, \\ -2x & \text{für } -\frac{1}{2} \leq x \leq 0, \\ 2x & \text{für } 0 \leq x \leq 1 \end{cases}$$

gegeben. Abb. 5.11 zeigt die Graphen von f und der zu $\mathfrak{S}_1(0)$ gehörenden Funktionen

$$S_c(x) = \frac{1}{4} - x + cx_+^1$$

für $c = \frac{5}{2}$ und $c = 3$. Bei der Anwendung von Satz 4 haben wir die Intervalle ($x_0 = -1$, $x_1 = 0$, $x_2 = 1$)

$$[\![x_0, x_1]\!], \quad [\![x_0, x_2]\!] \quad \text{und} \quad [\![x_1, x_2]\!]$$

zu betrachten. $X_3 = \left\{-1, -\frac{1}{2}, 0\right\}$ und $X_{5/2} = \left\{-1, -\frac{1}{2}, 0, 1\right\}$ sind Alternanten für S_3

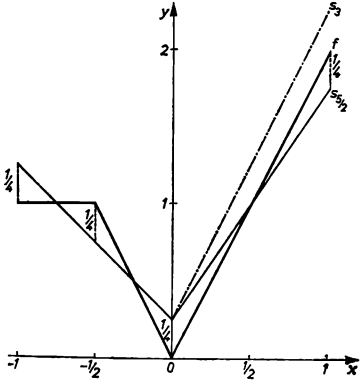


Abb. 5.11

bzw. $S_{5/2}$ auf $[\![-1, 1]\!]$, die dem Kriterium des Satzes bezüglich des Intervalls $[\![x_0, x_1]\!]$ genügen. Diese Funktionen sind also gleichmäßige Bestapproximationen von f in $\mathfrak{S}_1(0)$, und dies trifft offenbar auch auf alle S_c mit $\frac{5}{2} \leqq c \leqq 3$ zu.

Im Hinblick auf Satz 3 dieses Abschnittes und 5.3., Satz 2, ist nicht zu erwarten, daß die Interpolationsaufgabe für ein beliebiges System von $m + n + 1$ Knoten eines Einschließungsintervalls (18) und vorgegebene Ordinaten durch Linearkombinationen der Funktionen (19), d. h. in $\mathfrak{S}_m(x_1, \ldots, x_n)$ lösbar ist. Das hat zur Formulierung spezieller Interpolationsaufgaben der Splinetheorie geführt; eine derselben betrachten wir im folgenden Abschnitt.

5.4.4. Interpolation mit natürlichen Splinefunktionen

Wir kommen auf die Bemerkungen zur Polynominterpolation in den einführenden Betrachtungen zurück und versuchen, der Forderung nach Verringerung der Abweichung der interpolierenden Funktion und Wahrung der Stetigkeit für gewisse ihrer Ableitungen dadurch zu entsprechen, daß einerseits die Klasse der zur Lösung einer Interpolationsaufgabe zugelassenen Funktionen erweitert und andererseits eine die Glattheit derselben charakterisierende Größe minimiert wird. Im Sinne dieser Vorstellungen formulieren wir folgendes *Interpolationsproblem.* Es sei

$$x_1 < x_2 < \cdots < x_n \tag{23}$$

ein System von Interpolationsknoten und y_i der bei x_i $(i = 1(1)n)$ vorgeschriebene Funktionswert. Gesucht wird eine auf $[\![a, b]\!]$ k-mal stetig differenzierbare Funktion f, eine Funktion der Klasse $C^k_{[a,b]}$, für die

$$f(x_i) = y_i \qquad (i = 1(1)n) \tag{24}$$

ist und

$$\int_a^b [f^{(k)}(x)]^2 \, dx \tag{25}$$

minimal wird. Bei der Lösung dieser Extremalaufgabe sind also nur Funktionen aus $C^k_{[a,b]}$ zum Vergleich zugelassen, die (24) erfüllen. $[\![a, b]\!]$ bedeutet ein Einschließungsintervall für die Knoten im Sinne von (18).

Dieses Problem hat für $k = n$ genau eine Lösung, nämlich das in der Klasse $\mathfrak{P}_{n-1}$ wohlbestimmte Interpolationspolynom P. In der Tat: Da $P^{(k)}$ identisch verschwindet und so (25) annulliert, ist P Lösung der Aufgabe. Für eine weitere Lösung f müßte

$$\int_a^b [f^{(n)}(x)]^2 \, dx = 0,$$

d. h.

$$f^{(n)} \equiv 0 \text{ auf } [\![a, b]\!]$$

gelten. Daraus folgt aber $f \in \mathfrak{P}_{n-1}$ und im Hinblick auf (24) und MfL Bd. 9, 4.2., Satz 1,

$$f = P \quad \text{für} \quad x \in [\![a, b]\!].$$

Wenn $k > n$ ist, existieren offenbar unendlich viele Lösungen, z. B. kommen alle Interpolationspolynome in Betracht, die über (24) hinaus Wertvorgaben an $k - n$ weiteren Argumentstellen realisieren.

Für $1 < k < n$ werden wir sehen, daß in Verallgemeinerung des Falles $k = n$ das Interpolationsproblem genau eine Lösung besitzt, und zwar eine natürliche Splinefunktion des Grades $2k - 1$ mit den Knoten (23). Dieses Ergebnis ist eine Konsequenz der folgenden Sätze 5 und 6, deren Formulierung und Beweis wir einen Hilfssatz vorausschicken.

Hilfssatz 1. *Es sei* $s \in \mathfrak{S}_{2k-1}(x_1, \ldots, x_n)$ *eine natürliche Splinefunktion, ferner*

$$s(x) = P_{k-1}(x) + \sum_{i=1}^{n} c_i (x - x_i)_+^{2k-1}, \qquad P_{k-1} \in \mathfrak{P}_{k-1}, \tag{26}$$

die nach Satz 1 wohlbestimmte Darstellung durch elementare Splinefunktionen und

$$\sigma := \int_a^b [s^{(k)}(x)]^2 \, dx.$$

$[\![a, b]\!]$ *bedeutet ein Einschließungsintervall für die Knoten* x_i. *Dann gilt*

$$\sigma = (-1)^k (2k - 1)! \sum_{i=1}^{n} c_i s(x_i). \tag{27}$$

Beweis. Wir setzen $k \geqq 2$ voraus; die geringfügigen Modifikationen des Beweises für $k = 1$ seien dem Leser überlassen. Da s in den Intervallen $-\infty < x \leqq x_1$ und $x_n \leqq x < \infty$ durch je ein Polynom der Klasse $\mathfrak{P}_{k-1}$ dargestellt wird, ist daselbst $s^{(l)}(x) \equiv 0$ für $l = k, k + 1, \ldots, 2k - 2$. Unter Berücksichtigung, daß die Ableitungen bis zur Ordnung $2k - 2$ überall stetig sind, findet man durch partielle Integration

$$\begin{aligned}
\sigma &= \int_a^b [s^{(k)}(x)]^2 \, dx \\
&= \int_a^{x_1} [s^{(k)}(x)]^2 \, dx + \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} [s^{(k)}(x)]^2 \, dx + \int_{x_n}^{b} [s^{(k)}(x)]^2 \, dx \\
&= \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} [s^{(k)}(x)]^2 \, dx \\
&= \sum_{i=1}^{n-1} \left\{ [s^{(k-1)}(x) \, s^{(k)}(x)]_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} s^{(k-1)}(x) \, s^{(k+1)}(x) \, dx \right\} \\
&= -\sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(k-1)}(x) \, s^{(k+1)}(x) \, dx.
\end{aligned}$$

Durch Wiederholung dieser Schlußweise ergibt sich nach $k-1$ Schritten

$$\sigma = (-1)^{k-1} \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s'(x)\, s^{(2k-1)}(x)\, dx.$$

$s^{(2k-1)}(x)$ ist in den Intervallen $]\!]x_i, x_{i+1}[\![$, $i = 1(1)n-1$, einer Konstanten γ_i gleich, die in der Form

$$\gamma_i = s^{(2k-1)}(x_i + 0) = s^{(2k-1)}(x_{i+1} - 0) \tag{28}$$

bestimmt werden kann. Damit gilt weiter

$$\begin{aligned}\sigma &= (-1)^{k-1} \sum_{i=1}^{n-1} \gamma_i \int_{x_i}^{x_{i+1}} s'(x)\, dx \\ &= (-1)^{k-1} \sum_{i=1}^{n-1} \gamma_i [s(x_{i+1}) - s(x_i)] \\ &= (-1)^k \left\{ \gamma_1 s(x_1) + \sum_{i=2}^{n-1} (\gamma_i - \gamma_{i-1})\, s(x_i) - \gamma_{n-1} s(x_n) \right\},\end{aligned}$$

mit Beachtung von (28) also

$$\begin{aligned}\sigma = (-1)^k \Big\{ &s^{(2k-1)}(x_1 + 0)\, s(x_1) + \sum_{i=2}^{n-1} \left(s^{(2k-1)}(x_i + 0) - s^{(2k-1)}(x_i - 0) \right) s(x_i) \\ &- s^{(2k-1)}(x_n - 0)\, s(x_n) \Big\}.\end{aligned} \tag{29}$$

Nach $(2k-1)$-maliger Differentiation gewinnt man aus (26) an jeder von den Knoten verschiedenen Argumentstelle (vgl. (16))

$$s^{(2k-1)}(x) = (2k-1)! \sum_{i=1}^{n} c_i (x - x_i)_+^0.$$

Folglich ist

$$\begin{aligned} s^{(2k-1)}(x) &= (2k-1)! \sum_{j=1}^{i-1} c_j \quad \text{in }]\!]x_{i-1}, x_i[\![, \qquad i = 2(1)n, \\ \text{und} \qquad s^{(2k-1)}(x) &= (2k-1)! \sum_{j=1}^{i} c_j \quad \text{in }]\!]x_i, x_{i+1}[\![, \qquad i = 1(1)n-1. \end{aligned} \tag{30}$$

Damit kann für (29)

$$\sigma = (-1)^k (2k-1)! \left\{ c_1 s(x_1) + \sum_{i=2}^{n-1} c_i s(x_i) - s(x_n) \sum_{i=1}^{n-1} c_i \right\}$$

geschrieben werden.

Für $x > x_n$ folgt aus (26)

$$s(x) = P_{k-1}(x) + \sum_{i=1}^{n} c_i (x - x_i)^{2k-1}. \tag{31}$$

Da s als natürliche Splinefunktion des Grades $2k-1$ in diesem Argumentbereich aber durch ein Polynom aus $\mathfrak{P}_{k-1}$ dargestellt wird, müssen in (31) alle Potenzen von x verschwinden, deren Exponent größer als $k-1$ ist. Daraus ergibt sich

Hilfssatz 2. *Die Parameter c_i, $i = 1(1)n$, einer natürlichen Splinefunktion $s \in \mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$ in* (20) *genügen den Gleichungen*

$$\sum_{i=1}^{n} c_i x_i^j = 0 \quad \text{für} \quad j = 0, 1, \ldots, k-1. \tag{32}$$

Umgekehrt folgt auf Grund von Satz 1 aus dem Erfülltsein von (32), daß (26) eine natürliche Splinefunktion der Klasse $\mathfrak{S}_{2k-1}(x_1, \ldots, x_n)$ darstellt. Für $j = 0$ liefert (32)

$$c_n = -\sum_{j=1}^{n-1} c_j,$$

womit der zuletzt erhaltene Ausdruck für σ die Gestalt (27) annimmt.

Wir beweisen nun

Satz 5. *Es sei $k, n \in \mathbf{N}^*$ und $k \leqq n$. Dann gibt es für ein beliebiges Knotensystem* (23) *genau eine natürliche Splinefunktion $s \in \mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$, welche an den Stellen x_i vorgegebene Werte y_i, $i = 1(1)n$, annimmt.*

Beweis. Legen wir für s die Darstellung (26) zugrunde, wobei

$$P_{k-1}(x) = \sum_{\varkappa=0}^{k-1} a_\varkappa x^\varkappa$$

gesetzt sei, so bedeutet die Interpolationsanforderung die Erfüllung des Gleichungssystems

$$\sum_{\varkappa=0}^{k-1} a_\varkappa x_j^\varkappa + \sum_{i=1}^{n} c_i (x_j - x_i)_+^{2k-1} = y_j, \qquad j = 1(1)n. \tag{33}$$

Zusammen mit den nach Hilfssatz 2 von jeder natürlichen Splinefunktion aus $\mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$ zu erfüllenden Gleichungen (32) sind dies $n + k$ lineare Gleichungen für die Parameter von s. Offenbar ist die Behauptung von Satz 5 mit der eindeutigen Lösbarkeit dieses linearen Systems äquivalent. Das ist bewiesen, wenn wir zeigen können, daß das zugehörige homogene System nur die triviale Lösung besitzt. Wir schließen indirekt und nehmen für das homogene System die Existenz einer nichttrivialen Lösung

$$\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_{k-1}, \quad \bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n$$

an. Die mit diesen Parametern gemäß (26) gebildete natürliche Splinefunktion $\bar{s} \in \mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$ löst die Interpolationsaufgabe

$$\bar{s}(x_i) = 0, \qquad i = 1(1)n. \tag{34}$$

Nach Hilfssatz 1 ist dann

$$\tilde{\sigma} := \int_a^b [\tilde{s}^{(k)}(x)]^2 \, dx = 0$$

und für $k > 1$ aus Stetigkeitsgründen

$$\tilde{s}^{(k)}(x) \equiv 0 \quad \text{auf } [\![a, b]\!]; \tag{35}$$

wenn $k = 1$ ist, gilt (35) eventuell mit Ausnahme der Knotenargumente x_i. In diesem Fall folgt aus der Stetigkeit von $\tilde{s}$ und dem Umstand, daß $\tilde{s}$ auf $[\![a, x_1]\!]$ und $[\![x_n, b]\!]$ durch ein Polynom nullten Grades dargestellt wird, das bei x_1 und x_n verschwindet,

$$\tilde{s}(x) \equiv 0 \quad \text{auf } [\![a, b]\!]. \tag{36}$$

Für $k > 1$ gewinnt man (36) durch folgende Überlegung: Auf Grund von (35) wird $\tilde{s}$ auf $[\![a, b]\!]$ durch ein Polynom aus $\mathfrak{P}_{k-1}$ dargestellt. Dieses verschwindet gemäß (34) an n $(> k - 1)$ Stellen und ist daher das Nullpolynom. Mit Satz 2 folgt nun aus (36) das Verschwinden aller Parameter $\tilde{a}_\nu$, $\tilde{c}_i$, im Widerspruch zu unserer Annahme über diese Größen.

Mit dem folgenden Satz 6 werden wir erkennen, daß die nach Satz 5 in $\mathfrak{S}_{2k-1}(x_1, x_2, \ldots, x_n)$ wohlbestimmte natürliche Splinefunktion nicht nur die Forderung (24) des zu Anfang dieses Abschnitts formulierten Interpolationsproblems erfüllt, sondern für $k > 1$ auch einzige Lösung der mit (25) verknüpften Extremalaufgabe ist.

Satz 6. *Es sei $k, n \in \mathbf{N}^*$, $k \leqq n$ und s die zu dem Knotensystem* (23) *und vorgegebenen Ordinaten y_i, $i = 1(1)n$, gemäß Satz* 5 *wohlbestimmte natürliche Splinefunktion. Dann ist bezüglich jedes Einschließungsintervalls $[\![a, b]\!]$*

$$\int_a^b [f^{(k)}(x)]^2 \, dx \geqq \int_a^b [s^{(k)}(x)]^2 \, dx \tag{37}$$

für alle $f \in C^k_{[a,b]}$, die (24) *erfüllen. Im Fall $k > 1$ gilt das Gleichheitszeichen in* (37) *genau dann, wenn $f \equiv s$ auf $[\![a, b]\!]$ ist.*

Beweis. Ausgangspunkt unserer Überlegungen ist folgende Umformung:

$$\begin{aligned} \int_a^b [f^{(k)}(x)]^2 \, dx &= \int_a^b \left[s^{(k)}(x) + \left(f^{(k)}(x) - s^{(k)}(x)\right)\right]^2 dx \\ &= \int_a^b [s^{(k)}(x)]^2 \, dx + \int_a^b \left(f^{(k)}(x) - s^{(k)}(x)\right)^2 dx \\ &\quad + 2 \int_a^b s^{(k)}(x) \left(f^{(k)}(x) - s^{(k)}(x)\right) dx. \end{aligned} \tag{38}$$

Auf Grund von (38) ergibt sich (37), wenn gezeigt werden kann, daß

$$R := \int_a^b s^{(k)}(x)\left(f^{(k)}(x) - s^{(k)}(x)\right) dx = 0 \tag{39}$$

ist. Dazu wird das Integrationsintervall von (39) in die durch die Knoten (23) bestimmten Teilintervalle zerlegt. Unter Beachtung, daß s in $[\![a, x_1]\!]$ und $[\![x_n, b]\!]$ durch je ein Polynom aus $\mathfrak{P}_{k-1}$ dargestellt wird, folgt

$$\begin{aligned} R &= \int_a^b s^{(k)}(x)\left[f^{(k)}(x) - s^{(k)}(x)\right] dx \\ &= \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(k)}(x)\left[f^{(k)}(x) - s^{(k)}(x)\right] dx. \end{aligned} \tag{40}$$

Im Falle $k = 1$ ergibt sich aus (39) weiter mit Benutzung der durch (28) bestimmten Konstanten γ_i

$$\begin{aligned} R &= \sum_{i=1}^{n-1} \gamma_i \int_{x_i}^{x_{i+1}} \left[f'(x) - s'(x)\right] dx \\ &= \sum_{i=1}^{n-1} \gamma_i \left[\left(f(x_{i+1}) - s(x_{i+1})\right) - \left(f(x_i) - s(x_i)\right)\right], \end{aligned}$$

also $R = 0$ auf Grund der Voraussetzung, daß f und s die Interpolationsforderung (24) erfüllen. Für $k > 1$ gelangt man zu diesem Resultat durch partielle Integration in (40). Zunächst ist

$$\begin{aligned} R = \sum_{i=1}^{n-1} \Bigg\{ &\left[s^{(k)}(x)\left(f^{(k-1)}(x) - s^{(k-1)}(x)\right)\right]_{x_i}^{x_{i+1}} \\ &- \int_{x_i}^{x_{i+1}} s^{(k+1)}(x)\left[f^{(k-1)}(x) - s^{(k-1)}(x)\right] dx \Bigg\} \end{aligned}$$

und weiter wegen der Stetigkeit der im integralfreien Term auftretenden Funktionen und des Verschwindens von $s^{(k)}(x_1)$ und $s^{(k)}(x_n)$

$$R = -\sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(k+1)}(x)\left[f^{(k-1)}(x) - s^{(k-1)}(x)\right] dx.$$

Durch abermalige partielle Integration gewinnt man daraus auf entsprechende Weise

$$R = \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(k+2)}(x)\left[f^{(k-2)}(x) - s^{(k-2)}(x)\right] dx$$

und so fortfahrend nach $k-1$ Schritten

$$R = (-1)^{k-1} \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(2k-1)}(x)\,[f'(x) - s'(x)]\,dx.$$

$s^{(2k-1)}(x)$ ist in den Intervallen $]x_i, x_{i+1}[$ gleich den gemäß (28) bestimmten Konstanten γ_i, und von hier aus kann der Beweis wie im Falle $k = 1$ geführt werden.

Bei Gleichheit von

$$\int_a^b [f^{(k)}(x)]^2\,dx \quad \text{und} \quad \int_a^b [s^{(k)}(x)]^2\,dx$$

ergibt sich auf Grund von (38)

$$\int_a^b [f^{(k)}(x) - s^{(k)}(x)]^2\,dx = 0. \tag{41}$$

Im Fall $k > 1$ ist der Integrand stetig, und man gewinnt aus (41)

$$f^{(k)}(x) - s^{(k)}(x) \equiv 0 \text{ auf } [a, b],$$

d. h., $f - s$ ist auf $[a, b]$ durch ein Polynom aus $\mathfrak{P}_{k-1}$ darstellbar. Da f und s nach Voraussetzung (24) erfüllen, verschwindet dieses an mindestens n $(> k-1)$ Stellen und daher identisch auf $[a, b]$. Damit ist der Satz bewiesen.

Wie bei der Interpolation mit ganzen rationalen Funktionen ergeben sich auch bei der Spline-Interpolation in verschiedener Hinsicht interessante Spezialisierungen, wenn man die Knoten als äquidistant annimmt. Man vergleiche dazu etwa [18], Abschnitt 6; [33] und [44], Abschnitt I.5.

5.4.5. Kubische Splines

Unter einem *kubischen Spline* verstehen wir eine natürliche Splinefunktion s dritten Grades. Sind deren Knoten durch (23) gegeben, so ist s'' eine durchweg stetige Funktion, die in $]-\infty, x_1]$ und $[x_n, \infty[$ identisch verschwindet und in den Intervallen $]x_i, x_{i+1}[$, $i = 1(1)n-1$, durch ein Polynom der Klasse $\mathfrak{P}_1$ dargestellt wird. Ziel der folgenden Betrachtungen ist ein Algorithmus zur Konstruktion des durch die Sätze 5 und 6 charakterisierten interpolierenden Splines in dem hier zu erörternden Spezialfall.

T. N. E. Greville [19] folgend, gehen wir von der Newtonschen Darstellungsformel (MfL Bd. 9, 4.3.2.(17), (19)) aus. Danach gilt für beliebiges $x \neq x_i, x_{i+1}$ $(i = 1(1)\,n-1)$

$$s(x) = s(x_i) + (x - x_i)\,s[x_i, x_{i+1}] + (x - x_i)\,(x - x_{i+1})\,s[x, x_i, x_{i+1}], \tag{42}$$

wobei die mit eckigen Klammern gebildeten Terme Steigungen von s bedeuten; die Definition dieser Größen findet man in MfL Bd. 9, 4.2.2.

Algorithmen für Aufgaben mit kubischen Splines beruhen wesentlich auf dem Umstand, daß man die im Restglied von (42) auftretende Steigung $s[x, x_i, x_{i+1}]$ durch eine Linearkombination der Werte $s''(x_i)$ und $s''(x_{i+1})$ ausdrücken kann. Das ist — wie jetzt gezeigt werden soll — eine Konsequenz der oben hervorgehobenen Besonderheit der zweiten Ableitung von s: Auf Grund des Taylorschen Satzes ist für $x \in]x_i, x_{i+1}[$

$$s(x) = s(x_i) + s'(x_i)\,(x - x_i) + \frac{1}{2}(x - x_i)^2\, s''(x_i) + \frac{1}{6}(x - x_i)^3\, s'''(\xi_i),$$

$$x_i < \xi_i < x, \tag{43}$$

und folglich

$$s[x, x_i] = s'(x_i) + \frac{1}{2}(x - x_i)\, s''(x_i) + \frac{1}{6}(x - x_i)^2\, s'''(\xi_i). \tag{44a}$$

Da s''' in $]x_i, x_{i+1}[$ konstant ist und in diesem Intervall den Wert

$$s'''(x) = \frac{s''(x_{i+1}) - s''(x_i)}{x_{i+1} - x_i} = s''[x_i, x_{i+1}]$$

hat, nimmt (44a) die Form

$$s[x, x_i] = s'(x_i) + \frac{1}{2}(x - x_i)\, s''(x_i) + \frac{1}{6}(x - x_i)^2\, s''[x_i, x_{i+1}] \tag{44b}$$

an. Läßt man in dieser Gleichung x gegen x_{i+1} streben, so resultiert

$$s[x_i, x_{i+1}] = s'(x_i) + \frac{1}{2}(x_{i+1} - x_i)\, s''(x_i) + \frac{1}{6}(x_{i+1} - x_i)^2\, s''[x_i, x_{i+1}] \tag{45}$$

und, indem man (44b) von (45) subtrahiert und durch $x_{i+1} - x$ dividiert,

$$\begin{aligned} s[x, x_i, x_{i+1}] &= \frac{1}{2}\, s''(x_i) + \frac{1}{6}\,\frac{x_{i+1}^2 - x^2 - 2x_i(x_{i+1} - x)}{x_{i+1} - x}\, s''[x_i, x_{i+1}] \\ &= \frac{1}{2}\, s''(x_i) + \frac{1}{6}(x_{i+1} - x_i + x - x_i)\, s''[x_i, x_{i+1}] \\ &= \frac{1}{2}\, s''(x_i) + \frac{1}{6}\left(s''(x_{i+1}) - s''(x_i)\right) + \frac{1}{6}(x - x_i)\, s''[x_i, x_{i+1}]. \end{aligned}$$

Da s'' im Intervall $[x_i, x_{i+1}]$ durch ein Polynom aus $\mathfrak{P}_1$ dargestellt wird, kann die Steigung des letzten Terms durch $s''[x, x_i]$ ausgedrückt werden. Mit Beachtung dessen

findet man schließlich

$$s[x, x_i, x_{i+1}] = \frac{1}{2} s''(x_i) + \frac{1}{6}\left(s''(x_{i+1}) - s''(x_i)\right) + \frac{1}{6}\left(s''(x) - s''(x_i)\right)$$

$$= \frac{1}{6}\left[s''(x_i) + s''(x) + s''(x_{i+1})\right] \tag{46}$$

und erkennt, daß (46) eine Linearkombination der Werte $s''(x_i)$ und $s''(x_{i+1})$ ist. Im Hinblick auf (42) können wir demnach feststellen:

Der durch Satz 5 charakterisierte kubische Interpolationsspline ist berechenbar, wenn außer den Funktionswerten

$$s(x_i) = y_i, \qquad i = 1(1)n, \tag{47}$$

noch die Werte $s''(x_i)$ an den Knoten (23) bekannt sind.

Weil s'' überall stetig und $s''(x) \equiv 0$ für $-\infty < x < x_1$ und $x_n < x < \infty$ ist, gilt zunächst

$$s''(x_1) = s''(x_n) = 0. \tag{48}$$

Im weiteren wird gezeigt, daß $s''(x_2), s''(x_3), \ldots, s''(x_{n-1})$ als Lösung eines linearen Gleichungssystems bestimmt werden können. Wegen

$$s''[x_i, x_{i+1}] = \frac{s''(x_{i+1}) - s''(x_i)}{x_{i+1} - x_i}$$

folgt aus (45)

$$s[x_i, x_{i+1}] = s'(x_i) + (x_{i+1} - x_i)\left\{\frac{1}{3} s''(x_i) + \frac{1}{6} s''(x_{i+1})\right\}. \tag{49}$$

Durch Anwendung des Taylorschen Satzes auf das Intervall $]x_{i-1}, x_i[$ mit x_i als Bezugspunkt der Entwicklung gewinnt man der Herleitung von (49) entsprechend

$$s[x_{i-1}, x_i] = s'(x_i) - (x_i - x_{i-1})\left\{\frac{1}{3} s''(x_i) + \frac{1}{6} s''(x_{i-1})\right\}. \tag{50}$$

Subtraktion der Gleichung (50) von (49) liefert

$$(x_i - x_{i-1})\, s''(x_{i-1}) + 2(x_{i+1} - x_{i-1})\, s''(x_i) + (x_{i+1} - x_i)\, s''(x_{i+1})$$
$$= 6\{s[x_i, x_{i+1}] - s[x_{i-1}, x_i]\}, \qquad i = 2(1)n - 1. \tag{51}$$

(51) stellt in Verbindung mit (48) ein System von $n - 2$ linearen Gleichungen zur Bestimmung der Größen $s''(x_i)$, $i = 2(1)n - 1$, dar, dessen Koeffizienten und rechte Seite mit den Knoten (23) und den zu interpolierenden Werten $s(x_i) = y_i$, $i = 1(1)n$, gegeben sind. In 6.1.2. werden wir zeigen, daß (51) für paarweise verschiedene x_i

und beliebige y_i genau eine Lösung besitzt, und eine Interpolationsaufgabe durchrechnen. Abb. 5.12 veranschaulicht schematisch die Struktur der Koeffizientenmatrix, die nur in der Hauptdiagonalen und der darüber und darunter liegenden Schräglinie von Null verschiedene Elemente enthält. Deshalb bezeichnet man diese als (spezielle) *Bandmatrix* oder *tridiagonale Matrix*.

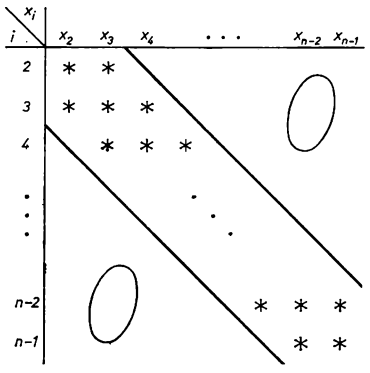


Abb. 5.12

Das in der Einleitung zu MfL Bd. 9, 4.3., formulierte Prinzip zur Gewinnung von Näherungsformeln der numerischen Differentiation und Integration mit Hilfe von Interpolationspolynomen kann auch für interpolierende Splinefunktionen übernommen werden. Wir führen das mit dem gemäß (42), (46) und (51) bestimmten kubischen Spline s durch.

Differentiation von (42) liefert mit Beachtung von (46) und der Konstanz von s''' zwischen zwei aufeinanderfolgenden Knoten

$$s'(x) = s[x_i, x_{i+1}] + (2x - x_i - x_{i+1})\, s[x, x_i, x_i] + \frac{1}{6}(x - x_i)(x - x_{i+1})\, s''[x_i, x_{i+1}]. \tag{52}$$

Man erkennt an (52), daß auch s' aus den in (47) beschriebenen Daten berechenbar ist.

Bei der Bestimmung von $\int_{x_1}^{x_n} s(x)\,dx$ wird das Integrationsintervall in die durch die Knoten berandeten Teilintervalle zerlegt. Die Integrale

$$\int_{x_i}^{x_{i+1}} s(x)\,dx, \qquad i = 1(1)n-1,$$

lassen sich auf Grund von MfL Bd. 9, 4.3.(31), *ohne Verfahrensfehler* mit der Keplerschen Faßregel berechnen. Dabei benötigt man die Werte von s an den Intervallmitten

$$\xi_i := \frac{1}{2}(x_i + x_{i+1}), \qquad i = 1(1)n-1, \tag{53}$$

die mit Hilfe von (42) berechnet werden. Wegen der Linearität von s'' auf $[\![x_i, x_{i+1}]\!]$ ist gemäß (46)

$$\begin{aligned} s[\xi_i, x_i, x_{i+1}] &= \frac{1}{6}\left\{s''(x_i) + \frac{s''(x_i) + s''(x_{i+1})}{2} + s''(x_{i+1})\right\} \\ &= \frac{1}{4}\left(s''(x_i) + s''(x_{i+1})\right) \end{aligned}$$

und folglich

$$\begin{aligned} s(\xi_i) &= s(x_i) + \frac{x_{i+1} - x_i}{2}\,\frac{s(x_{i+1}) - s(x_i)}{x_{i+1} - x_i} - \frac{(x_{i+1} - x_i)^2}{16}\left(s''(x_i) + s''(x_{i+1})\right) \\ &= \frac{1}{2}\left(s(x_i) + s(x_{i+1})\right) - \frac{(x_{i+1} - x_i)^2}{16}\left(s''(x_i) + s''(x_{i+1})\right). \end{aligned} \tag{54}$$

Man findet nun auf die beschriebene Weise unter Beachtung von MfL Bd. 9, 4.3.(24),

$$\begin{aligned} \int_{x_1}^{x_n} s(x)\,dx &= \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s(x)\,dx = \frac{1}{6}\sum_{i=1}^{n-1}(x_{i+1} - x_i)\left\{s(x_i) + 4s(\xi_i) + s(x_{i+1})\right\} \\ &= \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)\left(s(x_i) + s(x_{i+1})\right) \\ &\quad - \frac{1}{24}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^3\left(s''(x_i) - s''(x_{i+1})\right). \end{aligned} \tag{55}$$

Der erste Term der rechten Seite von (55) entspricht der Trapezregel zur Berechnung von $\int_{x_1}^{x_n} s(x)\,dx$. Für die Auswertung der Quadraturformel sind wieder die in (47) charakterisierten Daten maßgebend.

Allgemein (d. h. für beliebiges k) gewinnt man bei der näherungsweisen Berechnung von $\int_a^b f(x)\,dx$ durch Integration des bezüglich eines Knotensystems (23) gemäß Satz 5 zu $y = f(x)$ bestimmten interpolierenden Splines s eine Quadraturformel der Gestalt

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} A_i f(x_i) \tag{56}$$

mit von f unabhängigen Gewichten A_i. Das folgt aus der Bemerkung, daß man die Parameter der natürlichen Splinefunktion s bei der Lösung des linearen Systems (33) als Linearkombination der Werte $y_j = f(x_j)$ gewinnt, deren Koeffizienten nur von den Knoten abhängen.

Auf Grund einer von A. SARD entwickelten Theorie über die Bestapproximation linearer Funktionale und eines dafür grundlegenden Satzes von I. J. SCHOENBERG gelangt man zu der Einsicht, daß die mit den nach Satz 5 bestimmten natürlichen Splinefunktionen konstruierten numerischen Differentiations- und Quadraturformeln (speziell also (52) und (55)) in einem gewissen Sinne optimal sind. Im Rahmen dieser Einführung kann auf diesen Problemkreis nicht näher eingegangen werden, und wir müssen daher auf Fehlerbetrachtungen zu (52) und (55) verzichten. Für ein weiterführendes Studium sei auf [19, 20, 44] verwiesen.

6. Lösung von Gleichungen

In diesem Kapitel werden die Betrachtungen von MfL Bd. 9, 4.1., weitergeführt und auf *Gleichungssysteme* bzw. Gleichungen in mehrdimensionalen Räumen ausgedehnt. Ausführlicher gehen wir dabei auf lineare Systeme ein unter besonderer Berücksichtigung der Spezialfälle, die sich bei der Erörterung von Approximationsproblemen ergeben haben. Der zweite Teil des Kapitels ist nichtlinearen Gleichungen gewidmet. Nach allgemeiner Erörterung iterativer Methoden werden speziell algebraische Gleichungen behandelt, wobei mit Rücksicht auf die Schule die Lokalisierung von Polynomnullstellen verhältnismäßig breit dargestellt ist.

6.1. Numerische Lösung linearer Gleichungssysteme

6.1.1. Allgemeine Bemerkungen und Ergänzungen zur linearen Algebra

Wir befassen uns mit der numerischen Lösung des linearen Gleichungssystems

$$\begin{aligned} a_{11}x_1 &+ a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 &+ a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 &+ a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{aligned} \tag{1}$$

im Körper der reellen Zahlen. Mit der Koeffizientenmatrix

$$\boldsymbol{A} = (a_{ij})$$

vom Typ $m \times n$ und

$$\boldsymbol{b} = (b_1, b_2, \ldots, b_m)^{\mathsf{T}}, \qquad \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathsf{T}},$$

$$a_{ij}, b_i, x_j \in \mathbf{R}, \qquad i = 1(1)m, \quad j = 1(1)n,$$

nimmt (1) die Gestalt

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

an.

Von einem rein theoretischen Standpunkt aus betrachtet lassen sich die mit (2) zusammenhängenden algorithmischen Probleme so beschreiben:

Es sind Verfahren zu entwickeln, welche die Lösbarkeit von (2) *zu entscheiden gestatten und mit denen gegebenenfalls sämtliche Lösungen dieser Gleichung berechnet werden können.*

Grundlage für die Bestimmung der Struktur der Lösungsgesamtheit von (2) ist der in MfL Bd. 3, 5.4., behandelte *Gaußsche Algorithmus*. Dabei ist es wesentlich zu wissen, ob gewisse bei dem Eliminationsprozeß gebildete Koeffizienten und Störglieder (Komponenten der rechten Seite) verschwinden oder nicht. Offensichtlich ist das beim Rechnen mit gerundeten Größen und im Hinblick darauf, daß verschiedene EDVA im allgemeinen unterschiedlich runden, ein kritischer Punkt (vgl. 6.1.2.). Aus diesem Grunde wird bei der Entwicklung numerischer Verfahren zur Lösung von linearen Gleichungssystemen meist vorausgesetzt, daß die Matrix $\boldsymbol{A}$ in (2) regulär, also das System quadratisch ist ($m = n$). Diese Verfahren lassen sich wesentlich in zwei Klassen einteilen:

a) die sogenannten *exakten* oder *direkten*, welche die Lösung mit endlich vielen arithmetischen Operationen bei Abwesenheit von Rundungsfehlern exakt bestimmen, und

b) die *iterativen*, mit deren Hilfe eine Folge von Vektoren des $\mathbf{R}^n$ konstruiert wird, die gegen die Lösung konvergieren.

Mehrere der a) zuzuordnenden Methoden sind Weiterentwicklungen des Gaußschen Algorithmus. Das einfachste Iterationsverfahren ist das der *sukzessiven Approximation(en)*. Es beruht auf einer geeigneten Umformung von (2) ($m = n$) in eine Gleichung zweiter Art (vgl. MfL Bd. 9, Einleitung von 4.1.).

In den Anwendungen ergeben sich lineare Gleichungssysteme häufig aus Messungen, und die zu berechnenden Lösungen sind auf diese Weise im allgemeinen überbestimmt. Zur Erläuterung betrachten wir das folgende [31] entnommene Beispiel.

Für vier Punkte A, B, C, D einer Geraden (vgl. Abb. 6.1) wurden die Abstände in überschüssiger Anzahl gemessen. Dabei erhielt man

$$\begin{aligned} x_1 &:= (A, B) = 117{,}34 \text{ m}, \\ x_2 &:= (B, C) = 68{,}45 \text{ m}, \\ x_3 &:= (C, D) = 41{,}27 \text{ m}, \\ (A, C) &= 185{,}81 \text{ m}, \\ (B, D) &= 109{,}70 \text{ m}, \\ (A, D) &= 227{,}05 \text{ m}. \end{aligned} \tag{3}$$

Aus (3) resultiert folgendes lineare Gleichungssystem zur Bestimmung von x_1, x_2, x_3:

$$\begin{aligned} x_1 &= 117{,}34, \\ x_2 &= 68{,}45, \\ x_3 &= 41{,}27, \\ x_1 + x_2 &= 185{,}81, \\ x_2 + x_3 &= 109{,}70, \\ x_1 + x_2 + x_3 &= 227{,}05. \end{aligned} \tag{4}$$

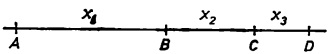


Abb. 6.1

Lineare Gleichungssysteme, die vergleichbar dem System (4) aufgestellt werden, sind wegen möglicher Meßungenauigkeiten im allgemeinen nicht lösbar. Andererseits wäre es unvernünftig, korrekt durchgeführte überschüssige Messungen als Hinderungsgrund für die Bestimmbarkeit der beobachteten Größen zu akzeptieren. In solchen Fällen hilft die sogenannte *Gaußsche Transformation* weiter, die aus der Umformung des Gleichungsproblems in eine spezielle Aufgabe der Quadratmittelapproximation resultiert.

Bezeichnet man die Spaltenvektoren der Matrix $\boldsymbol{A}$ mit $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$, so lassen sich (1), (2) in der Form

$$x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n = \boldsymbol{b} \tag{5}$$

schreiben. An Stelle von (5) betrachten wir folgende Aufgabe in dem unitären Raum E, der aus dem $\mathbf{R}^m$ durch Einführung des Skalarproduktes 5.2.(1) entsteht:

Gesucht wird diejenige Linearkombination der Vektoren $\boldsymbol{a}_i$, $i = 1(1)n$, die im Sinne von 5.2.(5) am wenigsten von $\boldsymbol{b}$ abweicht.

Nach der in 5.2. entwickelten Theorie genügen die Koeffizienten x_i dieser Bestapproximation den Normalgleichungen 5.2.(8). Mit den notwendigen Ersetzungen erhält man für diese im vorliegenden Fall

$$\sum_{i=1}^{n} x_i\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle = \langle \boldsymbol{b}, \boldsymbol{a}_j\rangle, \qquad j = 1(1)n. \tag{6}$$

In Matrixschreibweise kann (6) in der Form

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} \tag{7}$$

ausgedrückt werden. Offensichtlich ist in (7) die Koeffizientenmatrix symmetrisch $\left((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{\mathrm{T}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{TT}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)$. Den Übergang von (1) zu (7) bezeichnet man als *Gaußsche Transformation*. Auf Grund von 5.2., Satz 2, gilt:

Satz 1. *Sind die Spaltenvektoren $\boldsymbol{a}_i$, $i = 1(1)n$, linear unabhängig, so ist die Matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ regulär, und (7) besitzt für beliebiges $\boldsymbol{b} \in \mathbf{R}^m$ genau eine Lösung.*

Die Voraussetzung des Satzes 1 ist offenbar für das System (4) unseres Beispiels erfüllt. Das transformierte System (7) lautet:

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 530{,}20, \\ 2x_1 + 4x_2 + 2x_3 &= 591{,}01, \\ x_1 + 2x_2 + 3x_3 &= 378{,}02. \end{aligned} \tag{8}$$

Satz 2. *Ist in* (2) $\boldsymbol{A}$ *regulär, so sind die Systeme* (2) *und* (7) *äquivalent und eindeutig lösbar.*

Beweis. Die eindeutige Lösbarkeit folgt aus Satz 1, die Äquivalenz der Gleichungssysteme durch linksseitige Multiplikation von (2) mit $\boldsymbol{A}^{\mathsf{T}}$ bzw. von (7) mit $(\boldsymbol{A}^{\mathsf{T}})^{-1}$.

Für die Durchführung des Gaußschen Algorithmus ist es von Interesse, ob die sogenannten *Abschnittsdeterminanten* von $\boldsymbol{A}$

$$\Delta_1 := a_{11}, \qquad \Delta_2 := \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \ldots, \quad \Delta_n := \begin{vmatrix} a_{11} & \ldots & a_{1n} \\ \ldots\ldots\ldots\ldots \\ a_{n1} & \ldots & a_{nn} \end{vmatrix} \tag{9}$$

ungleich Null sind (s. u. Satz 5). Im Zusammenhang damit heben wir folgenden Spezialfall hervor:

Definition 1. Eine reelle symmetrische Matrix $\boldsymbol{A} = (a_{ij})$ vom Typ $n \times n$ heißt *positiv definit* genau dann, wenn ihre Abschnittsdeterminanten (9) positiv sind.

Lineare Gleichungssysteme mit positiv definiter Koeffizientenmatrix treten in den Anwendungen häufig auf, und einige Verfahren zur numerischen Lösung von linearen Gleichungen machen von dieser Eigenschaft wesentlich Gebrauch. Im Hinblick auf die Probleme der Quadratmittelapproximation (5.2.) beweisen wir:

Satz 3. $\varphi_1, \varphi_2, \ldots, \varphi_n$ *seien linear unabhängige Elemente eines unitären Raumes. Dann ist die Koeffizientenmatrix der Normalgleichungen* 5.2.(8) *positiv definit.*

Beweis. Wir bemerken zunächst, daß Δ_n die Gramsche Determinante der Vektoren φ_i, $i = 1(1)n$, ist. Um zu zeigen, daß die Δ_i, $i = 1(1)n$, in (9) positiv sind, wird die formale Determinante

$$\begin{vmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) & \ldots & (\varphi_1, \varphi_{n-1}) & \varphi_1 \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) & \ldots & (\varphi_2, \varphi_{n-1}) & \varphi_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (\varphi_n, \varphi_1) & (\varphi_n, \varphi_2) & \ldots & (\varphi_n, \varphi_{n-1}) & \varphi_n \end{vmatrix}$$

betrachtet, die bei Entwicklung nach der letzten Spalte eine gewisse Linearkombination

$$\psi_n := \alpha_1\varphi_1 + \alpha_2\varphi_2 + \cdots + \alpha_{n-1}\varphi_{n-1} + \Delta_{n-1}\varphi_n, \quad \alpha_i \in \mathbf{R}, \quad i = 1(1)\,n-1, \tag{10}$$

der φ_i darstellt. ψ_n ist nicht der Nullvektor, da Δ_{n-1} sonst auf Grund der linearen Unabhängigkeit der φ_i, $i = 1(1)n$, verschwinden müßte. Δ_{n-1} ist aber als Gramsche

Determinante des linear unabhängigen Systems $\varphi_1, \varphi_2, \ldots, \varphi_{n-1}$ nach 5.2., Satz 2, ungleich Null. Offenbar ist für $k = 1(1)n$

$$(\psi_n, \varphi_k) = \begin{vmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) & \cdots & (\varphi_1, \varphi_{n-1}) & (\varphi_1, \varphi_k) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) & \cdots & (\varphi_2, \varphi_{n-1}) & (\varphi_2, \varphi_k) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (\varphi_n, \varphi_1) & (\varphi_n, \varphi_2) & \cdots & (\varphi_n, \varphi_{n-1}) & (\varphi_n, \varphi_k) \end{vmatrix}$$

und folglich

$$(\psi_n, \varphi_k) = \begin{cases} 0 & \text{für } k \neq n, \\ \Delta_n & \text{für } k = n. \end{cases} \tag{11}$$

Skalare Multiplikation der Gleichung (10) mit ψ_n liefert unter Beachtung von (11)

$$(\psi_n, \psi_n) = \Delta_{n-1}\Delta_n > 0, \tag{12}$$

d. h., Δ_{n-1} und Δ_n haben gleiches Vorzeichen.

Nunmehr kann der Satz leicht durch Induktion nach n bewiesen werden: Offenbar ist $\Delta_1 = (\varphi_1, \varphi_1) > 0$. Wird angenommen, daß $\Delta_{n-1} > 0$ gilt, so folgt aus der letzten Bemerkung auch $\Delta_n > 0$.

Im besonderen ergibt sich aus Satz 3 der

Satz 4. *Unter der Voraussetzung von Satz* 1 *ist* A^TA *positiv definit.*

Beweis. A^TA ist die Matrix der Gramschen Determinante der linear unabhängigen Spaltenvektoren von A.

6.1.2. Direkte (exakte) Verfahren

Prototyp eines solchen Verfahrens ist der in MfL Bd. 3, 5.4., behandelte Gaußsche Algorithmus, mit dessen Hilfe (2) in ein äquivalentes lineares Gleichungssystem von Trapezform übergeführt wird. Dessen Lösbarkeit und gegebenenfalls Lösungsgesamtheit ist wie bekannt leicht zu bestimmen. Das Verfahren beruht auf den in MfL Bd. 3, 5.4., Satz 1, genannten elementaren Umformungen

1. Vertauschung zweier Gleichungen,
2. Ersetzen einer Gleichung durch ein mit einem von Null verschiedenen Skalar gebildetes Vielfaches,
3. Ersetzen einer Gleichung durch die Summe dieser Gleichung und einer beliebigen anderen des Systems,

denen wir mit Rücksicht auf die numerischen Belange noch

4. Vertauschung von Spalten der Matrix A und entsprechende Umbenennung der Unbekannten

hinzufügen.

Die Umformung auf Trapezgestalt geschieht in Schritten, die wir mit der Variablen k zählen. Im k-ten Schritt wird in der Gesamtheit der Koeffizienten a_{ij} $(i, j \geqq k)$ des dann vorliegenden Systems ein von Null verschiedener gesucht; ist $a_{i_1 j_1}$ ein solcher, so vertauscht man in (1) die k-te Zeile und Spalte mit der i_1-ten Zeile bzw. j_1-ten Spalte. Wird die durch den Spaltenindex bestimmte Numerierung der Unbekannten beibehalten, so erfordert das — wenn $j_1 \neq k$ — eine Umbenennung derselben. Diese ist in geeigneter Weise zu protokollieren, wozu im folgenden ein Feld **integer array** $PER[1:n]$ vereinbart wird, das am Anfang mit den Indizes der Unbekannten in der natürlichen Reihenfolge zu belegen ist. Bei Vertauschung der j_1-ten und k-ten Spalte werden die Werte von $PER[j1]$ und $PER[k]$ ausgetauscht, so daß am Schluß auf dem betrachteten Feld eine Permutation der natürlichen Zahlen von 1 bis n steht. Zur Formulierung von ALGOL-Prozeduren für Probleme, die mit der Lösung von (1) zusammenhängen, wollen wir uns die a_{ij}, b_i als Elemente eines **array** $A[1:m, 1:n+1]$ vorstellen, das durch die vorzunehmenden Umformungen laufend verändert wird. Ursprünglich gilt

$$\begin{aligned} &A[i, j] := a_{ij} \qquad \text{für } 1 \leqq i \leqq m,\ 1 \leqq j \leqq n, \\ &A[i, n+1] := b_i \quad \text{für } 1 \leqq i \leqq m. \end{aligned} \tag{13}$$

Wenn im weiteren ein Element $A[i, j]$ angesprochen wird, so ist jeweils sein aktueller Wert gemeint. Der k-te Schritt wird mit der oben erwähnten Zeilen- und Spaltenvertauschung eingeleitet, wobei die Wertzuweisung $A[k, k] := A[i1, j1]$ erfolgt. Weiterhin subtrahiert man von der $(k+1)$-ten bis zur letzten Gleichung ein solches Multiplum der k-ten, daß (k-ter Eliminationsschritt)

$$A[i, k] = 0 \text{ für } i = k+1, \ldots, m$$

ist. Das hat im übrigen die Wertzuweisung

$$\begin{aligned} &A[i, j] := A[i, j] - A[k, j]/A[k, k] \times A[i, k] \\ &\text{für } k < i \leqq m,\ k < j \leqq n+1 \end{aligned} \tag{14}$$

zur Folge. Nach endlich vielen Schritten gelangt man auf diese Weise zu einem (bis auf die Bezeichnung der Unbekannten) äquivalenten System der Form

$$\begin{array}{llllllll} A[1,1]x_1 + & A[1,2]x_2 + & & & + \cdots + & A[1,n]x_n & = & A[1, n+1], \\ 0 \quad + & A[2,2]x_2 + & & & + \cdots + & A[2,n]x_n & = & A[2, n+1], \\ & \cdot \quad \cdot \quad \cdot & & & & & & \vdots \\ 0 \quad + & 0 \quad + & A[r,r]x_r & + \cdots + & A[r,n]x_n & = & A[r, n+1], \\ 0 \quad + & 0 \quad + & 0 \quad + & \cdots + & 0 & = & A[r+1, n+1], \\ & & & & & & & \vdots \\ 0 \quad + & 0 \quad + & 0 \quad + & \cdots + & 0 & = & A[m, n+1]. \end{array} \tag{15}$$

Bezüglich (15) lassen sich die Lösungsverhältnisse bei dem Gleichungssystem (1) mit dem folgenden Algorithmus entscheiden (Abb. 6.2): Im Lösbarkeitsfall $r = m = n$ hat die Koeffizientenmatrix von (15) obere Dreiecksgestalt, und die x_i sind eindeutig berechenbar. Wenn $r < m$ und die Bedingung

$$A[r+1, n+1] = A[r+2, n+1] = \cdots = A[m, n+1] = 0 \tag{16}$$

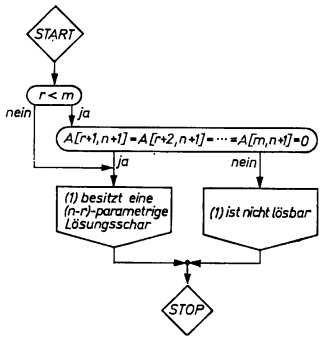


Abb. 6.2

erfüllt ist, erfordert die vollständige Lösung des Systems (1) die Bestimmung einer speziellen Lösung und eines Fundamentalsystems für die zugehörige homogene Gleichung (2). Eine spezielle Lösung ergibt sich aus (15), indem man die $(r+1)$-te bis m-te Gleichung beiseite läßt (reduziertes System),

$$x_{r+1} = x_{r+2} = \cdots = x_n = 0 \tag{17}$$

setzt und weiter wie im Falle $r = m = n$ verfährt. Ein Fundamentalsystem für die homogene Gleichung (2) gewinnt man aus dem reduzierten homogenen System (15), indem man nacheinander die Variablen $x_{r+1}, x_{r+2}, \ldots, x_n$ gemäß der Tabelle 6.1 belegt und dann sukzessive die Werte von $x_r, x_{r-1}, \ldots, x_1$ berechnet.

Die Durchführung des Gaußschen Algorithmus in dieser Form ist auf einem Rechner nicht sinnvoll. Zur Begründung wird ein numerisches Experiment betrachtet, welches das lineare Gleichungssystem

$$\begin{aligned} 5x_1 + 4x_2 + x_3 + x_4 + 3x_5 &= 0, \\ 3x_1 + 4x_2 + 3x_3 + 3x_4 + x_5 &= 0, \\ 2x_1 + 2x_2 + x_3 + x_4 + x_5 &= 0, \\ x_1 + 2x_2 + 2x_3 + 2x_4 &= 0 \end{aligned} \tag{18}$$

zum Gegenstand hat.

x_{r+1}	x_{r+2}	x_{r+3}	x_{r+4}	...	x_{n-1}	x_n
1	0	0	0	...	0	0
0	1	0	0	...	0	0
0	0	1	0	...	0	0
				... ...		
0	0	0	0	...	1	0
0	0	0	0	...	0	1

Tabelle 6.1

Zunächst lösen wir (18) von Hand nach dem Gaußschen Algorithmus, wobei rundungsfehlerfrei mit rationalen Zahlen gerechnet wird. Bei geeigneter Wahl der Eliminationsschritte gelangt man zu dem äquivalenten System

$$\begin{aligned} x_1 \quad &+ 2x_2 + 2x_3 + 2x_4 = 0, \\ x_5 &- 2x_2 - 3x_3 - 3x_4 = 0. \end{aligned}$$

Demnach ist $r = 2$, und man gewinnt durch Belegung der Variablen x_2, x_3 und x_4 gemäß Tabelle 6.1 zu (18) das Fundamentalsystem

$$\boldsymbol{x}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 3 \end{pmatrix} \quad \text{und} \quad \boldsymbol{x}_3 = \begin{pmatrix} -2 \\ 0 \\ 0 \\ 1 \\ 3 \end{pmatrix}. \tag{19}$$

Der von Hand abgearbeitete Gaußsche Algorithmus wurde in ALGOL programmiert und auf verschiedenen EDVA aktiviert. Bei einem R300-Lauf ergaben sich die in Tabelle 6.2 angegebenen Resultate. Der Rechner bestimmt also r zu 3 und gibt entsprechend nur zwei Fundamentallösungen aus, deren erste exakt ist. Mit dem gleichen Programm lieferte eine englische ICT 1900-Maschine nur eine Fundamentallösung, und zwar die erste der Tabelle 6.2. Ursache für diese Abweichungen sind die Rundungsfehler bei den Computerläufen. Da diese für das betrachtete Beispiel überschaubar klein sind, läßt sich der Sachverhalt auch so ausdrücken: Kleine Fehler in den Ausgangsdaten und Zwischenergebnissen bewirken eine sprunghafte Änderung des Resultates. Aufgaben, bei denen solche Effekte auftreten, sind offenbar für eine EDVA, wo mit maschinenspezifisch gerundeten Zahlen gerechnet wird, ungeeignet. Wir betrachten daher den Gaußschen Algorithmus weiterhin für ein quadratisches System (2) mit regulärer Koeffizientenmatrix $\boldsymbol{A}$.

Um den Rechenaufwand zu verringern, kann man in (14) für jedes $i = k + 1(1)n$ den Multiplikator $A[i, k]/A[k, k]$ berechnen und auf dem durch die Indizes i, k bestimmten Feldplatz speichern:

$$A[i, k] := A[i, k]/A[k, k]; \tag{14a}$$

danach nimmt (14) die Gestalt

$$A[i, j] := A[i, j] - A[k, j] \times A[i, k], \qquad k < i \leqq n, k < j \leqq n + 1, \tag{14b}$$

an. In 6.1.4. werden wir die Anzahl der insgesamt beim Gaußschen Algorithmus auszuführenden Operationen abschätzen.

	spezielle Lösung des inhomogenen Systems	1. Fundamentallösung	2. Fundamentallösung
x_1	.00000000 # 08	.00000000 # 01	−.66666666 # 00
x_2	.00000000 # 08	.00000000 # 08	.10000000 # 01
x_3	.00000000 # 15	−.10000000 # 01	−.66666667 # 00
x_4	.00000000 # 08	.10000000 # 01	.00000000 # 08
x_5	.00000000 # 15	.00000000 # 09	.00000000 # 09

Tabelle 6.2

Wenn mit Näherungswerten gerechnet wird, kann die Subtraktion in (14b) zu einem Verlust an gültigen Ziffern und — damit zusammenhängend — zu einer beträchtlichen Erhöhung des relativen Fehlers im Vergleich mit den Operanden (vgl. MfL Bd. 9, 2.5.3.) führen. Der absolute Fehler bei der Bestimmung von $A[i, j]$ wird wesentlich durch die Quotientenbildung (14a) beeinflußt; nach Formel (23) in MfL Bd. 9, 2.5.3., vergrößert sich dessen Betrag in dem Maße, wie $A[k, k]$ klein ist. Aus diesem Grunde ist es für die Gewinnung möglichst genauer Resultate geboten, den k-ten Schritt mit dem (bzw. einem) betragsgrößten der verfügbaren Koeffizienten einzuleiten. Man bezeichnet diesen als *Pivotelement* und den Vorgang selbst als *Pivotierung*.

Die folgende ALGOL-Prozedur *GAUSS(A,X,n)* berechnet nach dem Gaußschen Algorithmus für ein System von n linearen Gleichungen mit regulärer Koeffizientenmatrix den Lösungsvektor und übermittelt dessen Komponenten an das eindimensionale Feld X. Der formale Parameter A bezeichnet ein Feld, das der erweiterten Koeffizientenmatrix zuzuordnen ist. Die Pivotierung wird in *GAUSS* durch eine lokale Prozedur *PIV(k)* realisiert, welche im k-ten Schritt des Verfahrens die Indizes $i1$, $j1$ eines betragsgrößten unter den Elementen $A[i, j]$, $i = k(1)n$, $j = k(1)n$, ermittelt. $i1$, $j1$ sind in *PIV* global. Die möglicherweise vorgenommenen Spaltenvertauschungen bzw. Umbenennung von Unbekannten werden auf dem lokalen **integer array** *PER*[1 : n] protokolliert. Das lokale Feld $Y[1 : n]$ dient der Zwischenspeicherung bei der Berechnung der Lösungskomponenten nach Umformung der Koeffizientenmatrix auf Dreiecksgestalt.

```
procedure GAUSS(A,X,n);
integer n; array A,X;
begin
    integer i,j,k,i1,j1,hj,n1; real ha;
    integer array PER[1:n]; array Y[1:n];
    procedure PIV(k); integer k;
    begin
      integer i,j; real h,h1;
      h := 0;
      for i := k step 1 until n do
      for j := k step 1 until n do    begin h1 := abs(A[i,j]);
      if h < h1 then begin h := h1; i1 := i; j1 := j; end;
                                      end;
    end;
    n1 := n + 1; for k := 1 step 1 until n do PER[k] := k;
    for k := 1 step 1 until n-1 do begin PIV(k);
    if i1 ≠ k then for j := k step 1 until n1 do          begin
    ha := A[k,j]; A[k,j] := A[i1,j]; A[i1,j] := ha;       end;
    if j1 ≠ k then for i := 1 step 1 until n do          begin
    ha := A[i,k]; A[i,k] := A[i,j1]; A[i,j1] := ha;      end;
    hj := PER[k]; PER[k] := PER[j1]; PER[j1] := hj;
    for i := k + 1 step 1 until n do begin A[i,k] := A[i,k]/A(k,k];
    for j := k + 1 step 1 until n1 do A[i,j] := A[i,j] - A[k,j] × A[i,k];
                                      end;
                                      end;
    Y[n] := A[n,n1]/A[n,n];
    for i := n - 1 step -1 until 1 do begin Y[i] := A[i,n1];
    for j := i + 1 step 1 until n do Y[i] := Y[i] - A[i,j] × Y[j];
    Y[i] := Y[i]/A[i,i];                  end;
    for i := 1 step 1 until n do X[PER[i]] := Y[i];
end
```

Die Bestimmung eines betragsgrößten Elementes mit Hilfe der Prozedur *PIV* vor jedem Eliminationsschritt des Gaußschen Verfahrens zum Zwecke der Rundungsfehlerdämpfung erfordert natürlich zusätzliche Rechenzeit. Im folgenden konzipieren wir den sogenannten *verketteten Gaußschen Algorithmus* (auch als *Schema von Crout* bezeichnet) ohne Pivotsuche, d. h., wir arbeiten bei der Elimination in der k-ten Spalte unmittelbar mit dem dann gegebenen Element $A[k, k]$, dessen Nichtverschwinden wegen der Division in (14) gefordert werden, d. h. a priori bekannt sein muß.

Die hier zu erörternde Weiterentwicklung des Gaußschen Algorithmus vermeidet die Speicherung gewisser Zwischenresultate, vorausgesetzt, daß man *Produktsummen* (Skalarprodukte) durch Auflaufen der Produktsummanden in einem Register (möglichst doppelter Wortlänge) bilden kann. Da das bei den meisten Taschen-

rechnern der Fall ist, eignet sich der verkettete Gaußsche Algorithmus besonders für die Abarbeitung von Hand. Dafür spricht auch, daß man seine Anwendung wegen der nicht gedämpften Rundungsfehler i. a. auf kleine Systeme (1) beschränken muß.

Um die Zusammenhänge bei der Umformung der Koeffizientenmatrix des Systems (1) in eine obere Dreiecksmatrix nach dem Gaußschen Algorithmus zu erkennen, wird zunächst die Aktualisierung indizierter Variabler gemäß (14) durch die Einführung neuer Größen in jedem Umformungsschritt ersetzt. Wir verabreden, die Elemente a_{ij} und b_i, $i = 1(1)n$, $j = 1(1)n$, auch mit $a_{ij}^{(0)}$ bzw. $a_{i,n+1}^{(0)}$ zu bezeichnen, und definieren induktiv auf Grund von (14)

$$\begin{aligned} a_{ij}^{(k)} &:= a_{ij}^{(k-1)} - \frac{a_{kj}^{(k-1)} a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} && \text{für} && i = k + 1(1)n, \\ & && && j = k(1)n + 1, \\ a_{ij}^{(k)} &:= a_{ij}^{(k-1)} && \text{für} && i = 1(1)k,\ j = 1(1)n + 1 \\ & && \text{und} && i = k + 1(1)n,\ j = 1(1)k - 1. \end{aligned} \tag{20}$$

Dann ist die Matrix

$$A^{(k)} = (a_{ij}^{(k)}), \qquad i = 1(1)n,\ j = 1(1)n, \tag{21}$$

diejenige, welche aus der Koeffizientenmatrix des Systems (1) nach dem k-ten Umformungsschritt des Gaußschen Algorithmus erhalten wird.

Bestimmend für das weitere Vorgehen sind die Größen

$$b_{kj} := \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \qquad k < j \leqq n + 1, \qquad k = 1(1)n. \tag{22a}$$

$$c_{ik} := a_{ik}^{(k-1)}, \qquad k \leqq i \leqq n, \tag{22b}$$

Die c_{ik} werden für $k < i \leqq n$ im k-ten Umformungsschritt zum Verschwinden gebracht, während c_{kk} das k-te Diagonalelement der Matrix $A^{(n)}$ ist. Aus (20) gewinnt man damit schrittweise

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - c_{ik} b_{kj} = a_{ij}^{(k-2)} - c_{i,k-1} b_{k-1,j} - c_{ik} b_{kj} = \cdots \\ &= a_{ij}^{(0)} - c_{i1} b_{1j} - c_{i2} b_{2j} - \cdots - c_{ik} b_{kj} = a_{ij} - \sum_{l=1}^{k} c_{il} b_{lj}, \\ & i = k + 1(1)n, \quad j = k(1)n + 1, \end{aligned}$$

und speziell für $k = i - 1$ bzw. $k = j - 1$

$$b_{ij} = \frac{a_{ij}^{(i-1)}}{a_{ii}^{(i-1)}} = \frac{a_{ij} - \sum_{l=1}^{i-1} c_{il} b_{lj}}{c_{ii}}, \qquad i < j \leqq n + 1, \tag{23a}$$

$$c_{ij} = a_{ij}^{(j-1)} = a_{ij} - \sum_{l=1}^{j-1} c_{il} b_{lj}, \qquad j \leqq i \leqq n. \tag{23b}$$

Die Formeln (23) gestatten die rekursive Berechnung der b_{ij}, c_{ij} nach dem Schema

$$\begin{array}{llll} c_{11} & b_{12} & b_{13} & b_{14}\cdots \quad \text{1. Schritt} \\ c_{21} & c_{22} & b_{23} & b_{24}\cdots \quad \text{2. Schritt} \\ c_{31} & c_{32} & c_{33} & b_{34}\cdots \quad \text{3. Schritt} \\ c_{41} & c_{42} & c_{43} & \\ \vdots & \vdots & \vdots & \qquad \vdots \end{array} \tag{24}$$

Zunächst werden die c_{i1} und b_{1j} für $i = 1(1)n$, $j = 2(1)n+1$ gemäß (22) als die Elemente der ersten Spalte der ursprünglich gegebenen Koeffizientenmatrix bzw. als die durch den Diagonalkoeffizienten dividierten Elemente in deren erster Zeile bestimmt. Im folgenden Schritt gewinnt man nach (23b) mit Hilfe der c_{i1}, b_{1j} die c_{i2} für $i = 2(1)n$ und anschließend nach (23a) die b_{2j} für $j = 3(1)n+1$. In dieser Weise fortfahrend ergeben sich sämtliche c_{ij}, b_{ij}, wobei in jedem Schritt mit der Berechnung der c_{ij} zu beginnen ist.

Auf Grund von (22a) sind die b_{kj} die durch den Diagonalkoeffizienten c_{kk} dividierten Koeffizienten und Störglieder des nach dem Gaußschen Algorithmus umgeformten Gleichungssystems (1), so daß dessen Lösung durch

$$x_i = b_{i,n+1} - \sum_{j=i+1}^{n} b_{ij} x_j, \qquad i = n(-1)1, \tag{25}$$

gegeben ist.

Mit der folgenden ALGOL-Prozedur *CROUT* werden die b_{ij}, c_{ij} gemäß (23) bestimmt und auf einem Feld $BC[1:n, 1:n+1]$ gespeichert. Das noch als formaler Parameter auftretende Feld A ist der erweiterten Koeffizientenmatrix zuzuordnen.

```
procedure CROUT(A,BC,n); integer n; array A,BC;
begin
    integer i,j,k,n1; real h;
    n1 := n + 1;
    for i := 1 step 1 until n do BC[i,1] := A[i,1];
    for j := 2 step 1 until n1 do BC[1,j] := A[1,j]/A[1,1];
    for k := 2 step 1 until n do                    begin
    for i := k step 1 until n do  begin h := 0;
    for j := 1 step 1 until k - 1 do h := h + BC[i,j] × BC[j,k];
    BC[i,k] := A[i,k] - h          end;
    for j := k + 1 step 1 until n1 do  begin h := 0;
    for i := 1 step 1 until k - 1 do h := h + BC[k,i] × BC[i,j];
    BC[k,j] := (A[k,j] - h)/BC[k,k]  end;    end
end
```

Irgendeine Variante des Gaußschen Algorithmus, welche eine reguläre Matrix A in Dreiecksgestalt überführt und neben Zeilen- und Spaltenvertauschungen nur die Addition eines Vielfachen einer Zeile zu einer anderen benutzt (also keine Normierung der Diagonalelemente vornimmt), läßt bis auf das Vorzeichen die Determinante

von $\boldsymbol{A}$ invariant, d. h., es gilt

$$|\boldsymbol{A}| = \pm|\boldsymbol{B}|,^{1})$$

wenn $\boldsymbol{B}$ hier die durch den Eliminationsprozeß erzeugte obere Dreiecksmatrix bedeutet. Die Determinante von $\boldsymbol{B}$ ist aber gleich dem Produkt der Diagonalkoeffizienten, und man gewinnt so ein praktisches Verfahren zur Berechnung von Determinantenwerten. In 6.1.4. wird die Anzahl der dabei auszuführenden arithmetischen Operationen mit dem Aufwand bei der Auswertung der Leibnizschen Summenformel verglichen.

Läßt sich der Gaußsche Algorithmus ohne Zeilen- und Spaltenvertauschungen durchführen, so gilt

$$|\boldsymbol{A}| = |\boldsymbol{B}|$$

auch für die Abschnittsdeterminanten (9) von $\boldsymbol{A}$ und $\boldsymbol{B}$, d. h., man gewinnt mit den in (20) eingeführten Bezeichnungen

$$\Delta_1 = a_{11}^{(0)}, \quad \Delta_2 = a_{11}^{(0)} a_{22}^{(1)}, \quad \ldots, \quad \Delta_n = a_{11}^{(0)} a_{22}^{(1)} \cdots a_{nn}^{(n-1)}. \tag{26}$$

Im besonderen folgt aus (26), daß die Abschnittsdeterminanten von $\boldsymbol{A}$ nicht verschwinden. Offenbar gilt auch die Umkehrung dieses Sachverhalts.

Satz 5. *$\boldsymbol{A}$ sei eine Matrix vom Typ $n \times n$, und es sei $\Delta_j \neq 0$, $j = 1(1)n$. Dann ist $a_{jj}^{(j-1)} \neq 0$, und das Gaußsche Eliminationsverfahren kann ohne Zeilen- und Spaltenvertauschung durchgeführt werden.*

Damit haben wir ein hinreichendes Kriterium für die Durchführbarkeit des Croutschen Algorithmus gewonnen. Nach Definition 1 ist die Voraussetzung des Satzes 5 für eine positiv definite Matrix und nach Satz 3 speziell für die Koeffizientenmatrix der Normalgleichungen im Falle linear unabhängiger Vektoren erfüllt.

Wir gehen noch auf ein Faktorisierungsproblem für Matrizen ein, das eng mit dem Gaußschen Algorithmus verknüpft ist. Dabei wird nach der Darstellbarkeit einer quadratischen Matrix als Produkt einer unteren und oberen Dreiecksmatrix gefragt. Wir zeigen, daß eine solche existiert, wenn $\boldsymbol{A}$ die Voraussetzungen des Satzes 5 erfüllt. Es sei dann $\boldsymbol{B}$ diejenige obere Dreiecksmatrix, die man aus der in (21) definierten Matrix $\boldsymbol{A}^{(n)}$ durch Normierung der Hauptdiagonalelemente gewinnt, d. h., in

$$\boldsymbol{B} = \begin{pmatrix} 1 & b_{12} & b_{13} & \ldots & b_{1n} \\ & 1 & b_{23} & & b_{2n} \\ & & \cdot & & \\ & & & & \\ & 0 & & & 1 \end{pmatrix} \tag{27}$$

sind die

$$b_{ij} = \frac{a_{ij}^{(n)}}{a_{ii}^{(i-1)}} = \frac{a_{ij}^{(i-1)}}{a_{ii}^{(i-1)}}, \qquad i = 1(1)n-1, \quad i < j,$$

[1]) $|\boldsymbol{A}|$ bedeutet die Determinante von $\boldsymbol{A}$.

die mit dem Croutschen Algorithmus bestimmten Elemente. Die dabei auszuführenden Umformungen lassen sich durch linksseitige Multiplikation von $\boldsymbol{A}$ mit $n \times n$ Matrizen der folgenden Art realisieren:

$$\boldsymbol{D}_1 = (d_{ij}^{(1)}) = \begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & \ddots & & & \\ & & & \alpha & & \\ & & & & 1 & \\ & 0 & & & & \ddots \\ & & & & & & 1 \end{pmatrix}$$

$(d_{ij}^{(1)} = 0 \text{ für } i \neq j,\ d_{ii}^{(1)} = 1 \text{ für } i \neq k,\ d_{kk}^{(1)} = \alpha);$ (28)

$$\boldsymbol{D}_2 = (d_{ij}^{(2)}) = \begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ 0 & & \ddots & & & \\ \vdots & & \alpha & 1 & & \\ \vdots & & & & \ddots & \\ 0 & & \cdots & & 0 & 1 \end{pmatrix}$$

$(d_{ij}^{(2)} = 0 \text{ für } i \neq j \text{ und } i \neq l,\ j \neq k\ (l > k),\ d_{lk}^{(2)} = \alpha,\ d_{ii}^{(2)} = 1).$

In der Tat: Die linksseitige Multiplikation von $\boldsymbol{A}$ mit $\boldsymbol{D}_1$ hat die Ersetzung der k-ten Zeile von $\boldsymbol{A}$ durch ihr α-faches zur Folge, während alle übrigen Zeilen unverändert bleiben. Führt man diese Operation mit $\boldsymbol{D}_2$ aus, so wird der l-ten Zeile von $\boldsymbol{A}$ die mit α multiplizierte k-te Zeile hinzugefügt und sonst nichts verändert. Damit läßt sich die Matrix (27) durch linksseitige Multiplikation von $\boldsymbol{A}$ mit einem Produkt $\boldsymbol{D}$ von Matrizen (28) gewinnen:

$$\boldsymbol{DA} = \boldsymbol{B}. \tag{29}$$

$\boldsymbol{D}$ ist eine untere Dreiecksmatrix und regulär, da die α-Werte in den Matrixfaktoren vom Typ $\boldsymbol{D}_1$ beim verketteten Gaußschen Algorithmus nicht verschwinden. Offenbar ist aber die zu einer regulären unteren Dreiecksmatrix gebildete Inverse wieder eine untere Dreiecksmatrix, so daß sich aus (29) mit

$$\boldsymbol{A} = \boldsymbol{D}^{-1}\boldsymbol{B} \tag{30}$$

eine Lösung des oben formulierten Faktorisierungsproblems ergibt. Darüber hinaus läßt sich zeigen, daß in (30)

$$\boldsymbol{D}^{-1} = \boldsymbol{C}$$

gilt, wenn

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & & & 0 \\ c_{21} & c_{22} & & \\ & & \ddots & \\ c_{n1} & c_{n2} & & c_{nn} \end{pmatrix} \tag{31}$$

die mit den nach dem Croutschen Algorithmus zu bestimmenden Elementen (22b) gebildete untere Dreiecksmatrix ist. In der Tat ergibt sich, wenn $\boldsymbol{CB} = (u_{ij})$ gesetzt wird, nach (23a) für $j > i$

$$u_{ij} = \sum_{l=1}^{n} c_{il} b_{lj} = \sum_{l=1}^{i} c_{il} b_{lj} = \sum_{l=1}^{i-1} c_{il} b_{lj} + c_{ii} b_{ij} = a_{ij}$$

und nach (23b) für $i \geqq j$

$$u_{ij} = \sum_{l=1}^{n} c_{il} b_{lj} = \sum_{l=1}^{j} c_{il} b_{lj} = c_{ij} + \sum_{l=1}^{j-1} c_{il} b_{lj} = a_{ij},$$

also

$$\boldsymbol{CB} = \boldsymbol{A}.$$

Demnach gilt

Satz 6. *$\boldsymbol{A}$ sei eine Matrix vom Typ $n \times n$, und es sei $\Delta_j \neq 0$, $j = 1(1)n$. Dann läßt sich $\boldsymbol{A}$ als Produkt einer unteren und einer oberen Dreiecksmatrix darstellen.*

Eine spezielle Faktorisierung

$$\boldsymbol{A} = \boldsymbol{CB}$$

gewinnt man mit den Matrizen (27) und (31) nach dem Croutschen Algorithmus.

Der folgende Satz klärt die eindeutige Bestimmtheit der Produktdarstellung.

Satz 7. *Es sei $\boldsymbol{A}$ eine reguläre Matrix vom Typ $n \times n$, die sich als Produkt $\boldsymbol{A} = \boldsymbol{CB}$ einer unteren und einer oberen Dreiecksmatrix $\boldsymbol{C}$ bzw. $\boldsymbol{B}$ darstellen läßt. Dann ist diese Faktorisierung eindeutig durch die Diagonalelemente von $\boldsymbol{B}$ oder $\boldsymbol{C}$ bestimmt.*

Beweis. Mit $\boldsymbol{A}$ sind auch $\boldsymbol{B}$, $\boldsymbol{C}$ regulär, und das bedeutet $c_{ii} \neq 0$, $b_{ii} \neq 0$ für $i = 1(1)n$. Wir führen den Beweis durch Induktion nach n und nehmen etwa die Diagonalelemente von $\boldsymbol{B}$ als gegeben an. Für $n = 1$ folgt die Behauptung aus $a_{11} = c_{11} b_{11}$. Um von $n - 1$ auf n zu schließen, stellt man die Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ in der Form

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{n-1} & \boldsymbol{U} \\ \boldsymbol{V} & a_{nn} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{n-1} & \boldsymbol{O} \\ \boldsymbol{X} & c_{nn} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{n-1} & \boldsymbol{Y} \\ \boldsymbol{O} & b_{nn} \end{pmatrix}$$

dar; $\boldsymbol{U}$, $\boldsymbol{Y}$ sind Matrizen vom Typ $(n-1) \times 1$, $\boldsymbol{V}$, $\boldsymbol{X}$ solche vom Typ $1 \times (n-1)$ und $\boldsymbol{B}_{n-1}$, $\boldsymbol{C}_{n-1}$ obere bzw. untere Dreiecksmatrizen. Dann folgt aus

$$\boldsymbol{CB} = \begin{pmatrix} \boldsymbol{C}_{n-1}\boldsymbol{B}_{n-1} & \boldsymbol{C}_{n-1}\boldsymbol{Y} \\ \boldsymbol{X}\boldsymbol{B}_{n-1} & \boldsymbol{X}\boldsymbol{Y} + c_{nn} b_{nn} \end{pmatrix} = \boldsymbol{A},$$

daß

$$\boldsymbol{A}_{n-1} = \boldsymbol{C}_{n-1}\boldsymbol{B}_{n-1}, \tag{32a}$$

$$\boldsymbol{C}_{n-1}\boldsymbol{Y} = \boldsymbol{U}, \tag{32b}$$

$$\boldsymbol{X}\boldsymbol{B}_{n-1} = \boldsymbol{V}, \tag{32c}$$

$$\boldsymbol{X}\boldsymbol{Y} + c_{nn}\boldsymbol{b}_{nn} = \boldsymbol{a}_{nn} \tag{32d}$$

ist. Auf Grund der Induktionsannahme sind $\boldsymbol{C}_{n-1}$ und $\boldsymbol{B}_{n-1}$ eindeutig durch (32a) bestimmt. Die Komponenten von $\boldsymbol{Y}$ und $\boldsymbol{X}$ ergeben sich dann als wohlbestimmte Lösungen der linearen Gleichungssysteme (32b) bzw. (32c) in Dreiecksgestalt, deren Koeffizientenmatrizen wegen $b_{ii} \neq 0$ und $c_{ii} \neq 0$ regulär sind. Schließlich bestimmt man c_{nn} eindeutig aus (32d).

Auf Grund des Satzes 7 ist die mit dem Croutschen Algorithmus erzeugte Faktorisierung von $\boldsymbol{A}$ diejenige, bei der in der Hauptdiagonalen von $\boldsymbol{B}$ Einsen stehen. Speziell gilt

Satz 8. *Für eine symmetrische Matrix $\boldsymbol{A}$, deren Abschnittsdeterminanten ungleich Null sind, liefert der Croutsche Algorithmus eine Faktorisierung*

$$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{B} \quad \textit{mit} \quad b_{ij} = \frac{c_{ji}}{c_{ii}}, \quad i \geqq j. \tag{33}$$

Beweis. Wegen $\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}} = \boldsymbol{A} = \boldsymbol{C}\boldsymbol{B}$ und

$$\boldsymbol{C}^{\mathsf{T}} = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{n1} \\ & c_{22} & \cdots & c_{2n} \\ & 0 & \ddots & \\ & & & c_{nn} \end{pmatrix} = \begin{pmatrix} c_{11} & & & 0 \\ & c_{22} & & \\ & & \ddots & \\ 0 & & & c_{nn} \end{pmatrix} \begin{pmatrix} 1 & \frac{c_{21}}{c_{11}} & \cdots & \frac{c_{n1}}{c_{11}} \\ & 1 & & \frac{c_{n2}}{c_{22}} \\ & 0 & & \\ & & & 1 \end{pmatrix}$$

hat man

$$\boldsymbol{A} = \boldsymbol{B}^{\mathsf{T}} \begin{pmatrix} c_{11} & & & 0 \\ & c_{22} & & \\ & & \ddots & \\ 0 & & & c_{nn} \end{pmatrix} * \begin{pmatrix} 1 & \frac{c_{21}}{c_{11}} & \cdots & \frac{c_{n1}}{c_{11}} \\ & 1 & & \frac{c_{n2}}{c_{22}} \\ & & \ddots & \\ & & & 1 \end{pmatrix},$$

wobei $*$ eine Darstellung von $\boldsymbol{A}$ als Produkt einer unteren und einer oberen Dreiecksmatrix markiert. Da die Elemente in der Hauptdiagonalen des zweiten Faktors Einsen sind, muß dieser nach Satz 7 mit der durch den Croutschen Algorithmus bestimmten Matrix $\boldsymbol{B}$ übereinstimmen, d. h., es gilt (33).

Aus (33) können wir noch die Folgerung ziehen, daß sich eine symmetrische Matrix $\boldsymbol{A}$ unter der Voraussetzung des Satzes 7 in der Form

$$\boldsymbol{A} = \boldsymbol{S}^{\mathsf{T}}\boldsymbol{S} \tag{34}$$

mit einer oberen Dreiecksmatrix S darstellen läßt. (34) ergibt sich aus der Matrixgleichung

$$
\begin{pmatrix} c_{11} & & 0 \\ c_{21} & c_{22} & \\ \vdots & \ddots & \\ c_{n1} & c_{n2} & \cdots \; c_{nn} \end{pmatrix} * \begin{pmatrix} 1 & \frac{c_{21}}{c_{11}} & \cdots & \frac{c_{n1}}{c_{11}} \\ & 1 & & \frac{c_{n2}}{c_{22}} \\ 0 & & \ddots & \\ & & & 1 \end{pmatrix}
$$

$$
= \begin{pmatrix} \sqrt{c_{11}} & & & 0 \\ \frac{c_{21}}{\sqrt{c_{11}}} & \sqrt{c_{22}} & & \\ & & & \\ \frac{c_{n1}}{\sqrt{c_{11}}} & \frac{c_{n2}}{\sqrt{c_{22}}} & \cdots & \sqrt{c_{nn}} \end{pmatrix} \begin{pmatrix} \sqrt{c_{11}} & & 0 \\ & \sqrt{c_{22}} & \\ 0 & & \ddots \\ & & \sqrt{c_{nn}} \end{pmatrix}
$$

$$
* \begin{pmatrix} \frac{1}{\sqrt{c_{11}}} & & & 0 \\ & \frac{1}{\sqrt{c_{22}}} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{c_{nn}}} \end{pmatrix} \begin{pmatrix} \sqrt{c_{11}} & \frac{c_{21}}{\sqrt{c_{11}}} & \cdots & \frac{c_{n1}}{\sqrt{c_{11}}} \\ & \sqrt{c_{22}} & & \frac{c_{n2}}{\sqrt{c_{22}}} \\ & & \ddots & \\ 0 & & & \sqrt{c_{nn}} \end{pmatrix} \tag{35}
$$

$$
= \begin{pmatrix} \sqrt{c_{11}} & & & 0 \\ \frac{c_{21}}{\sqrt{c_{11}}} & \sqrt{c_{22}} & & \\ & & \ddots & \\ \frac{c_{n1}}{\sqrt{c_{11}}} & \frac{c_{n2}}{\sqrt{c_{22}}} & \cdots & \sqrt{c_{nn}} \end{pmatrix} \begin{pmatrix} \sqrt{c_{11}} & \frac{c_{21}}{\sqrt{c_{11}}} & \cdots & \frac{c_{n1}}{\sqrt{c_{11}}} \\ & \sqrt{c_{22}} & & \frac{c_{n2}}{\sqrt{c_{22}}} \\ & & \ddots & \\ 0 & & & \sqrt{c_{nn}} \end{pmatrix}.
$$

Die Elemente $\sqrt{c_{ii}}$ brauchen indessen nicht reell zu sein. Im Falle einer positiv definiten Matrix A gilt jedoch

Satz 9. *Es sei A eine positiv definite Matrix vom Typ $n \times n$. Dann existiert eine reelle obere Dreiecksmatrix S derart, daß*

$$A = S^T S$$

gilt.

Beweis. Wir wählen für $S = (s_{ij})$ die gemäß (35) mit Hilfe des Croutschen Algorithmus konstruierte obere Dreiecksmatrix. Dann gilt nach (21) $c_{kk} = a_{kk}^{(k-1)}$, und auf Grund von (26) ist $c_{kk} > 0$ wegen $\Delta_i > 0$, $i = 1(1)n$. Die in (35) auftretenden Wurzeln sind also sämtlich reell.

Satz 9 ist Grundlage für die Methode von CHOLESKY zur Lösung linearer Gleichungssysteme mit positiv definiter Koeffizientenmatrix A.

Es sei $S = (s_{ij})$ eine Matrix, die im Sinne von Satz 9 eine Faktorisierung von A bestimmt. Dafür gilt

$$a_{ij} = \sum_{l=1}^{n} s_{li} s_{lj},$$

wobei $i \leqq j$ wegen der Symmetrie von A angenommen wird. S ist eine obere Dreiecksmatrix und daher $s_{li} = 0$ für $l > i$, so daß

$$a_{ij} = \sum_{l=1}^{j} s_{li} s_{lj} \quad \text{für } i \leqq j \tag{36}$$

gilt. Mit Hilfe von (36) lassen sich die von Null verschiedenen s_{ij} zeilenweise mit $i = 1$ beginnend berechnen. Man findet zunächst $a_{11} = s_{11}^2$, also $s_{11} = \sqrt{a_{11}}$; hier wie im folgenden wählen wir für die Wurzel das positive Zeichen. Weiterhin ist

$$a_{j1} = s_{1j} s_{11}, \quad \text{d. h.} \quad s_{1j} = \frac{a_{j1}}{s_{11}} = \frac{a_{1j}}{s_{11}} \quad \text{für } j = 2(1)n.$$

Sind die Elemente der $(i-1)$-ten Zeile von S schon bestimmt, so folgt aus (36)

$$a_{ii} = \sum_{l=1}^{i} s_{li}^2,$$

also

$$s_{ii} = \sqrt{a_{ii} - \sum_{l=1}^{i-1} s_{li}^2} \tag{37}$$

und

$$a_{ji} = \sum_{l=1}^{i} s_{lj} s_{li}, \qquad j > i,$$

d. h.

$$s_{ij} = \frac{a_{ji} - \sum_{l=1}^{i-1} s_{lj} s_{li}}{s_{ii}}. \tag{38}$$

Wegen der Symmetrie von A kann in (38) a_{ji} durch a_{ij} ersetzt werden.

Hat man auf diese Weise für A eine Darstellung der Form $A = S^T S$ ermittelt, so kann die Lösung des inhomogenen Linearen Gleichungssystems

$$Ax = b, \quad \text{d. h. } S^T S x = b, \tag{39}$$

in folgenden Etappen erfolgen: Zunächst wird ein Vektor y als Lösung von

$$S^T y = b \tag{40}$$

und danach x als Lösung von

$$Sx = y \tag{41}$$

bestimmt. Da (40) und (41) Systeme in Dreiecksgestalt sind, bereitet deren Lösung keine Schwierigkeiten.

Die Faktorisierung von A gemäß Satz 9 und die Lösung von (39) in den Etappen (40) und (41) macht das Verfahren von CHOLESKY aus. Wir fassen dieses in einer ALGOL-Prozedur *CHOLESKY*(A,B,X,n) zusammen, wobei A ein Feld ist, das mit der positiv definiten Koeffizientenmatrix A vom Typ $n \times n$ in (39) korrespondiert und B, X eindimensionale Felder bedeuten, die sich auf die rechte Seite und den Lösungsvektor beziehen. S wird auf A oberhalb der Hauptdiagonalen gebildet.

```
procedure CHOLESKY(A,B,X,n); integer n; array A,B,X;
begin
    integer i,j,l; real h;
    A[1,1] := sqrt(A[1,1]);
    for j := 2 step 1 until n do A[1,j] := A[1,j]/A[1,1];
    for i := 2 step 1 until n do                    begin h := 0;
    for l := 1 step 1 until i − 1 do h := h + A[l,i] × A[l,i];
    A[i,i] := sqrt(A[i,i] − h);
    for j := i + 1 step 1 until n do begin h := 0;
    for l := 1 step 1 until i − 1 do h := h + A[l,j] × A[l,i];
    A[i,j] := (A[i,j] − h)/A[i,i]      end      end;
    for i := 1 step 1 until n do                    begin h := 0;
    for j := 1 step 1 until i − 1 do  h := h + A[j,i] × X[j];
    X[i] := (B[i] − h]/A[i,i]; B[i] := X[i] end;
    for i := n step −1 until 1 do     begin h := 0;
    for j := n step −1 until i + 1 do h := h + A[i,j] × X[j];
    X[i] := (B[i] −h)/A[i,i];         end
end
```

In 5.4.5. haben wir gesehen, daß die Interpolation mit kubischen Splines auf ein lineares Gleichungssystem mit einer tridiagonalen Koeffizientenmatrix führt, die wir allgemein in der Form

$$A = \begin{pmatrix} a_1 & b_1 & & & 0 \\ c_2 & a_2 & b_2 & & \\ \cdots & \cdots & \cdots & \cdots & \\ & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{pmatrix} \tag{42}$$

notieren. Speziell gilt für das mit den Knoten 5.4.(23) gebildete System, 5.4.(51) ($n-2$ Gleichungen)

$$\begin{aligned} & & a_1 &= 2(x_3 - x_1), & b_1 &= x_3 - x_2 \\ c_i &= (x_{i+1} - x_i), & a_i &= 2(x_{i+2} - x_i), & b_i &= x_{i+2} - x_{i+1} \\ c_{n-2} &= (x_{n-1} - x_{n-2}), & a_{n-2} &= 2(x_n - x_{n-2}) & & \\ & & & & i &= 2(1)n-3. \end{aligned} \tag{43}$$

Aus (43) folgt unter Beachtung von $x_1 < x_2 < \cdots < x_n$ nach Umbenennung von $n-2$ in n

$$\begin{aligned} & |a_1| > |b_1| > 0, \\ & |a_i| \geqq |b_i| + |c_i| \quad \text{für} \quad i = 2(1)n-1, \\ & |a_n| > |c_n| > 0. \end{aligned} \tag{44}$$

Diese Eigenschaften gestatten eine bemerkenswerte Folgerung:

Satz 10. *Unter der Voraussetzung* (44) *ist die Matrix* (42) *regulär und läßt sich eindeutig in der Form*

$$\begin{pmatrix} a_1 & b_1 & & & 0 \\ c_2 & a_2 & b_2 & & \\ & \cdots & \cdots & \cdots & \\ & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_1 & & & & 0 \\ c_2 & \alpha_2 & & & \\ & c_3 & \alpha_3 & & \\ & & \ddots & \ddots & \\ & 0 & & c_n & \alpha_n \end{pmatrix} \begin{pmatrix} 1 & \beta_1 & & & \\ & 1 & \beta_2 & & 0 \\ & & 1 & \beta_3 & \\ & & & \ddots & \ddots \\ & 0 & & & \beta_{n-1} \\ & & & & 1 \end{pmatrix} \tag{45}$$

faktorisieren. Dabei gelten die Abschätzungen

$$|\beta_i| < 1 \tag{46a}$$

und

$$|a_i| - |c_i| < |\alpha_i| < |a_i| + |c_i|. \tag{46b}$$

Beweis. Setzen wir zunächst voraus, daß eine Produktdarstellung der Form (45) mit von Null verschiedenen α_i, $i = 1(1)n$, existiert. Dann gilt

$$\alpha_1 = a_1, \qquad \beta_1 = b_1/\alpha_1, \tag{47a}$$

$$\alpha_i = a_i - c_i\beta_{i-1}, \qquad i = 2(1)n, \tag{47b}$$

$$\beta_i = b_i/\alpha_i, \qquad i = 2(1)n-1. \tag{47c}$$

Wir zeigen nun auf Grund von (44), daß sich die α_i, β_i rekursiv nach dem Schema (47) bestimmen lassen und für diese Größen die Ungleichungen (46) gelten. Wegen $|a_1| > |b_1| > 0$ ist zunächst klar, daß dies für $i = 1$ zutrifft. Nehmen wir nun im Sinne eines induktiven Beweises an, daß die Aussage für $i = 1, 2, \ldots, j-1$ wahr sei. Dann ist nach (47b) und (44)

$$|a_j| + |c_j| > |\alpha_j| > |a_j| - |c_j| \geqq |b_j|, \quad \text{also} \quad \alpha_j > 0.$$

β_j kann folglich nach (47c) bestimmt werden, und es gilt

$$|\beta_j| = \frac{|b_j|}{|\alpha_j|} < 1.$$

Offensichtlich realisieren die gemäß (47) konstruierten Bandmatrizen die Faktorisierung (45). Bezeichnen wir diese mit $\boldsymbol{L}$ bzw. $\boldsymbol{R}$, so kann ein mit (42) gebildetes lineares Gleichungssystem

$$\boldsymbol{Ax} = \boldsymbol{r} \tag{48}$$

analog zu (39) bis (41) in den Etappen

$$\boldsymbol{Ly} = \boldsymbol{r}, \qquad \boldsymbol{Rx} = \boldsymbol{y} \tag{49}$$

gelöst werden.

Wir schreiben dafür unter der Voraussetzung von (44) eine ALGOL-Prozedur *TRIDAG(A,B,C,R,X,n)*, in der *A, B, C, R, X* eindimensionale Felder sind, die mit den entsprechend bezeichneten Diagonalen der Koeffizientenmatrix (42) bzw. mit der rechten Seite und dem Lösungsvektor von (48) korrespondieren.

```
procedure TRIDAG(A,B,C,R,X,n); integer n; array A,B,C,R,X;
begin
    integer i;
    B[1] := B[1]/A[1];
    for i := 2 step 1 until n − 1 do                    begin
    A[i] := A[i] − C[i] × B[i − 1]; B[i] := B[i]/A[i]  end;
    A[n] := A[n] − C[n] × B[n − 1];
    X[1] := R[1]/A[1]; R[1] := X[1];
    for i := 2 step 1 until n do
    X[i] := (R[i] − C[i] × X[i − 1])/A[i];
    for i := n − 1 step −1 until 1 do X[i] := X[i] − B[i] × X[i + 1];
end
```

In einem Programm, das *TRIDAG* benutzt, müssen die Feldvereinbarungen **array** $A[1:n]$, **array** $B[1:n-1]$, **array** $C[2:n]$, **array** $R,X[1:n]$ erfolgen.

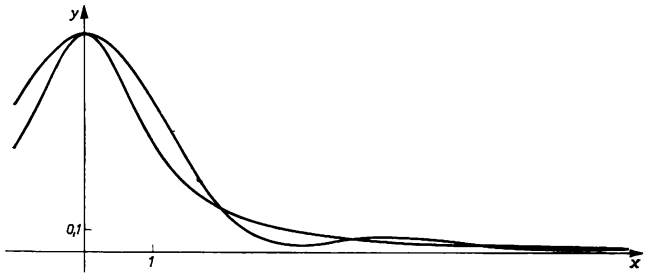
Mit Hilfe von *TRIDAG* lösen wir vergleichsweise zu dem in MfL Bd. 9, 4.2.4., betrachteten Beispiel folgende Interpolationsaufgabe: Es ist diejenige kubische Splinefunktion zu bestimmen, die in den Knoten

$$x_j = -8 + 2j, \qquad j = 0(1)8,$$

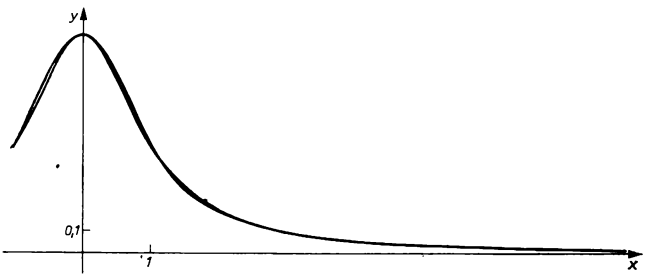
die Werte $\frac{1}{1+x_j^2}$ annimmt. Die Koeffizienten (43) sind dann

$$a_i = 8, \qquad b_i = c_i = 2.$$

Der Lösungsvektor $\mathbf{s}$ des linearen Gleichungssystems 5.4.(51) ist zusammen mit dessen rechter Seite in Tabelle 6.3 erfaßt. Dabei bedeutet s_k die zweite Ableitung des interpolierenden Splines an der Stelle $x_k = -8 + 2k$, $k = 1(1)7$. Abb. 6.3a zeigt den Graphen des mit diesen Daten nach 5.4.(42), (46) berechneten kubischen Splines und der zu interpolierenden Funktion $y = \frac{1}{1+x^2}$. In Abb. 6.3b ist zum Vergleich der mit den Knoten $x_j = -8 + j$, $j = 0(1)16$, konstruierte Interpolationsspline dargestellt.



a)



b)

Abb. 6.3

k	s_k	$6\{[x_k, x_{k+1}] - [x_{k-1}, x_k]\}$
1	0,029010	0,060462
2	−0,085810	0,328140
3	0,478299	1,976471
4	−0,839149	−4,800000
5	0,478299	1,976471
6	−0,085810	0,328140
7	0,029010	0,060462

Tabelle 6.3

6.1.3. Iterative Verfahren

Weiterhin wird die Gleichung (2) mit einer quadratischen Koeffizientenmatrix vom Typ $n \times n$ betrachtet. Iterative Verfahren bestimmen die Lösung als Grenzelement einer Folge von Vektoren; ihre Theorie beinhaltet wesentlich Kriterien für deren Konvergenz und Untersuchungen zur Konvergenzgeschwindigkeit bzw. Abschätzungen des Fehlers, der bei Abbruch einer solchen Folge entsteht.

Die im $\mathbf{R}^n$ durchzuführenden Konvergenzbetrachtungen werden übersichtlich, wenn man Normen verwendet. Da in linearen Gleichungssystemen Vektoren in Verbindung mit Matrizen auftreten, ist folgende Ergänzung der in 5.1.1. entwickelten Vorstellungen durch den Begriff der *Matrixnorm* zweckmäßig. Jede reelle Matrix $\boldsymbol{A}$ vom Typ $n \times n$ kann als ein Element des $\mathbf{R}^{n^2}$ aufgefaßt werden, und es ist dann $\|\boldsymbol{A}\|$ mit einer im $\mathbf{R}^{n^2}$ definierten Vektornorm $\|\cdot\|$ bildbar. Auf Grund von 5.1.(6) gilt für solche Matrizen $\boldsymbol{A}, \boldsymbol{B}$

$$\|\alpha \boldsymbol{A}\| = |\alpha| \cdot \|\boldsymbol{A}\|, \qquad \alpha \in \mathbf{R},$$

$$\|\boldsymbol{A} + \boldsymbol{B}\| \leqq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|,$$

im allgemeinen jedoch nicht

$$\|\boldsymbol{A}\boldsymbol{B}\| \leqq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|. \tag{50}$$

Da die letzte Ungleichung eine wesentliche Grundlage für Abschätzungen ist, wird folgende Definition eingeführt:

Definition 2. Eine (Vektor-)Norm des $\mathbf{R}^{n^2}$ heißt *Matrixnorm*, wenn für alle Matrizen des Typs $n \times n$ bei einer bestimmten Zuordnung ihrer Elemente zu den Vektorkoordinaten die Ungleichung (50) gilt.

Wir zeigen, daß es Matrixnormen gibt, und betrachten zu diesem Zweck eine beliebige Norm $\|\cdot\|$ des $\mathbf{R}^n$ und für eine Matrix $\boldsymbol{A}$ vom Typ $n \times n$ die Menge der Quotienten

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}, \qquad \boldsymbol{x} \in \mathbf{R}^n, \quad \boldsymbol{x} \neq \boldsymbol{0}. \tag{51}$$

Diese ist beschränkt: Zunächst gilt mit $\boldsymbol{y} := \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ $(\|\boldsymbol{y}\| = 1)$

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \left\| \frac{1}{\|\boldsymbol{x}\|} \boldsymbol{A}\boldsymbol{x} \right\| = \left\| \boldsymbol{A} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right\| = \|\boldsymbol{A}\boldsymbol{y}\|$$

und (52)

$$\sup_{\substack{\boldsymbol{x} \in \mathbf{R}^n \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sup_{\substack{\boldsymbol{y} \in \mathbf{R}^n \\ \|\boldsymbol{y}\| = 1}} \|\boldsymbol{A}\boldsymbol{y}\|.$$

$\|\boldsymbol{A}\boldsymbol{x}\|$ stellt eine im $\mathbf{R}^n$ stetige Funktion dar. Das läßt sich aus 5.1., Satz 1, folgern, wenn man $\boldsymbol{A}\boldsymbol{x}$ als Linearkombination $x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ der Spaltenvektoren von $\boldsymbol{A}$ auffaßt und folgende Zuordnung vornimmt:

Bezeichnung in 5.1., Satz 1	Interpretation
E	$\mathbf{R}^n$
f	$\boldsymbol{0}$
$\varphi_1, \varphi_2, \ldots, \varphi_n$	$\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$
$a_1, a_2, \ldots, a_n$	$x_1, x_2, \ldots, x_n$

Da die Menge $\boldsymbol{M} = \{\boldsymbol{y} : \boldsymbol{y} \in \mathbf{R}^n \wedge \|\boldsymbol{y}\| = 1\}$ abgeschlossen und beschränkt ist, nimmt $\|\boldsymbol{A}\boldsymbol{y}\|$ auf $\boldsymbol{M}$ nach MfL Bd. 4, 2.4.2., Satz 1, das Minimum an, d. h., die oberen Grenzen in (52) sind endlich.

Wir definieren nun für die Matrix $\boldsymbol{A}$ (und damit für ein beliebiges Element des $\mathbf{R}^{n^2}$)

$$|\boldsymbol{A}| := \sup_{\substack{\boldsymbol{x} \in \mathbf{R}^n \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sup_{\substack{\boldsymbol{y} \in \mathbf{R}^n \\ \|\boldsymbol{y}\| = 1}} \|\boldsymbol{A}\boldsymbol{y}\|. \tag{53}$$

Die durch (53) bestimmte Abbildung $|\cdot|$ des $\mathbf{R}^{n^2}$ in $\mathbf{R}_+$ hat die Eigenschaften einer Norm. Von den Normeigenschaften 5.1.(6) sind nur die Dreiecksungleichung und die Implikation $|\boldsymbol{A}| = 0 \Rightarrow \boldsymbol{A} = \boldsymbol{0}$ nicht unmittelbar ersichtlich. Gilt $|\boldsymbol{A}| = 0$, so folgt für alle $\boldsymbol{x} \in \mathbf{R}^n$, $\boldsymbol{x} \neq \boldsymbol{0}$,

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$$

und speziell für die Vektoren $\boldsymbol{x}_i = (\delta_{ij})$, $i = 1(1)\,n$, $j = 1(1)\,n$,

$$\boldsymbol{a}_i = \boldsymbol{0}, \quad \text{also} \quad \boldsymbol{A} = \boldsymbol{0}.$$

Für zwei quadratische Matrizen $\boldsymbol{A}, \boldsymbol{B}$ erhält man $\|(\boldsymbol{A} + \boldsymbol{B})\,\boldsymbol{y}\| \leqq \|\boldsymbol{A}\boldsymbol{y}\| + \|\boldsymbol{B}\boldsymbol{y}\|$ und, wenn $\|\boldsymbol{y}\| = 1$ ist, $\|(\boldsymbol{A} + \boldsymbol{B})\,\boldsymbol{y}\| \leqq |\boldsymbol{A}| + |\boldsymbol{B}|$. Auf Grund dessen ist

$$|\boldsymbol{A} + \boldsymbol{B}| = \sup_{\|\boldsymbol{y}\| = 1} \|(\boldsymbol{A} + \boldsymbol{B})\,\boldsymbol{y}\| \leqq |\boldsymbol{A}| + |\boldsymbol{B}|,$$

was zu beweisen war. Mit der Norm $|\cdot|$ gewinnt man aus (53) für eine beliebige Matrix $\boldsymbol{A}$ vom Typ $n \times n$ und $\boldsymbol{x} \in \mathbf{R}^n$, $\boldsymbol{x} \neq \boldsymbol{0}$,

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq |\boldsymbol{A}|,$$

also

$$\|\boldsymbol{A}\boldsymbol{x}\| \leqq |\boldsymbol{A}| \cdot \|\boldsymbol{x}\|; \tag{54}$$

die Abschätzung (54) gilt offenbar auch für den Nullvektor. Damit zeigt man leicht, daß $|\cdot|$ eine Matrixnorm ist. Für zwei Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ der betrachteten Art und einen Vektor $\boldsymbol{x} \in \mathbf{R}^n$ gilt nämlich

$$\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}(\boldsymbol{B}\boldsymbol{x})\| \leqq |\boldsymbol{A}| \cdot \|\boldsymbol{B}\boldsymbol{x}\| \leqq |\boldsymbol{A}| \cdot |\boldsymbol{B}| \cdot \|\boldsymbol{x}\| \tag{55}$$

und folglich

$$|\boldsymbol{A}\boldsymbol{B}| = \sup_{\substack{\boldsymbol{x} \in \mathbf{R}^n \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq |\boldsymbol{A}| \cdot |\boldsymbol{B}|. \tag{56}$$

Definition 3. $|\cdot|$ heißt die *durch die Vektornorm* $\|\cdot\|$ *induzierte Matrixnorm*. Irgendeine Matrixnorm, für die bezüglich einer Vektornorm $\|\cdot\|$ des $\mathbf{R}^n$ (54) gilt, wird als mit dieser *verträglich* bezeichnet.

Wir betrachten Beispiele. Für die in 5.1.(11) eingeführten Normen $\|\cdot\|_p$, $p \geqq 1$, einschließlich des durch 5.1.(14) definierten Falles $p = \infty$ sei die induzierte Matrixnorm mit $|\cdot|_p$ bezeichnet. Dann gilt für quadratische Matrizen $\boldsymbol{A}$ des Typs $n \times n$

$$|\boldsymbol{A}|_1 = \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}|, \tag{57a}$$

$$|\boldsymbol{A}|_\infty = \max_{i \in \{1,2,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|. \tag{57b}$$

Um $|\boldsymbol{A}|_1$ und $|\boldsymbol{A}|_\infty$ zu finden, muß man also bei den Komponenten von $\boldsymbol{A}$ zu den absoluten Beträgen übergehen und das Maximum der Spalten- bzw. Zeilensummen bestimmen. Beim Beweis von (57) gehen wir von (53) aus. Im Falle $p = 1$ ist gemäß 5.1.(11)

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{x}\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}| \cdot |x_j| \leqq \sum_{j=1}^{n} \left(\max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}| \right) |x_j| \\ &= \left(\max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| = \|\boldsymbol{x}\|_1 \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}| \end{aligned} \tag{58}$$

und speziell für einen normierten Vektor $\boldsymbol{y}$

$$\|\boldsymbol{A}\boldsymbol{y}\|_1 \leqq \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}|. \tag{59}$$

Damit folgt

$$|\boldsymbol{A}|_1 \leqq \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}|.$$

Andererseits läßt sich leicht ein normierter Vektor $\boldsymbol{y}$ angeben, so daß mit $\boldsymbol{x} = \boldsymbol{y}$ in (58) und (59) überall das Gleichheitszeichen auftritt: Wird das Maximum von $\sum_{i=1}^{n} |a_{ij}|$ für $j = k$ angenommen, so wähle man nur $y_j = 0$ für $j \neq k$ und $y_k = 1$. Es gilt also (57a). Im Falle (57b) ist

$$\|\boldsymbol{Ax}\|_\infty = \max_{i\in\{1,2,\ldots,n\}} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \leqq \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j|$$

$$\leqq \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} \left(\max_{k=\{1,2,\ldots,n\}} |x_k|\right) |a_{ij}| = \|\boldsymbol{x}\|_\infty \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|$$

und speziell für einen normierten Vektor $\boldsymbol{y}$

$$\|\boldsymbol{Ay}\|_\infty \leqq \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|.$$

Wird das Maximum von $\sum_{j=1}^{n} |a_{ij}|$ für $i = k$ angenommen und setzt man

$$y_j = \frac{|a_{kj}|}{a_{kj}} \quad \text{für} \quad a_{kj} \neq 0,$$

$$y_j = 1 \quad \text{für} \quad a_{kj} = 0,$$

so wird für den hierdurch bestimmten normierten Vektor $\boldsymbol{x} = \boldsymbol{y}$ in diesen Abschätzungen wieder überall das Gleichheitszeichen realisiert, und man gewinnt (57b).

Wir betrachten noch $\|\cdot\|_2$, begnügen uns aber mit einer Abschätzung der induzierten Matrixnorm. Die genaue Bestimmung von $\|\cdot\|_2$ würde ein Eingehen auf die Eigenwerttheorie bei Matrizen erfordern. Zunächst ist

$$\|\boldsymbol{Ax}\|_2^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}x_j\right)^2$$

und auf Grund der Schwarzschen Ungleichung (MfL Bd. 4, 1.5.2.)

$$\left(\sum_{j=1}^{n} a_{ij}x_j\right)^2 \leqq \left(\sum_{j=1}^{n} a_{ij}^2\right)\left(\sum_{j=1}^{n} x_j^2\right) = \|\boldsymbol{x}\|_2^2 \sum_{j=1}^{n} a_{ij}^2.$$

Damit folgt $\|\boldsymbol{Ax}\|_2^2 \leqq \|\boldsymbol{x}\|_2^2 \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$, also nach (53)

$$\|\boldsymbol{A}\|_2 \leqq \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2\right)^{1/2}. \tag{60}$$

Für ein umfassendes Verständnis der Iterationsmethoden sind — wie schon im Zusammenhang mit der letzten Betrachtung bemerkt — Kenntnisse aus der Eigenwerttheorie erforderlich. Da hierauf nicht zurückgegriffen werden kann, erörtern wir nur das einfachste dieser Verfahren, die Methode der sukzessiven Approximation(en). Dabei gehen wir von einer geeigneten Umformung von (2) in eine äquivalente Glei-

chung der Form

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c} \tag{61}$$

aus, in der $\boldsymbol{B}$ eine Matrix vom Typ $n \times n$ ist und $\boldsymbol{x}$, $\boldsymbol{c}$ Spaltenvektoren bedeuten.

In den Anwendungen tritt häufig der Fall auf, daß die Diagonalkoeffizienten der Matrix $\boldsymbol{A}$ in (2) von Null verschieden sind und ihre Beträge deutlich die der anderen Elemente übertreffen. Dann kann mit Rücksicht auf das Weitere folgende Umformung von (1) zweckmäßig sein:

$$\begin{aligned}
x_1 &= & -\frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n + \frac{b_1}{a_{11}},\\
x_2 &= -\frac{a_{21}}{a_{22}}x_1 & -\frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n + \frac{b_2}{a_{22}},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
x_n &= -\frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \frac{a_{n3}}{a_{nn}}x_3 - \cdots - & + \frac{b_n}{a_{nn}}.
\end{aligned} \tag{62}$$

In Matrixschreibweise stellt (62) offenbar eine Gleichung der Form (61) dar.

Die durch

$$f: \boldsymbol{x} \mapsto \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c} \tag{63}$$

gegebene Abbildung ist ausnahmslos im $\mathbf{R}^n$ definiert und bildet diesen Raum in sich ab. (61) erscheint damit als die Gleichung zweiter Art

$$\boldsymbol{x} = f(\boldsymbol{x}), \tag{64}$$

die wir weiterhin mit Bezug auf den Banachschen Fixpunktsatz diskutieren. Dieser wurde in MfL Bd. 4, 2.4.5., für einen Spezialfall bewiesen mit dem Kommentar, daß sich die Gültigkeit des Satzes, ohne am Beweis etwas zu ändern, auf Kontraktionsoperatoren in beliebigen vollständigen metrischen Räumen ausdehnen läßt. Der Begriff des metrischen Raumes und der Konvergenz von Punktfolgen in einem solchen ist in MfL Bd. 4, 1.5. und 2.1., erklärt. Der Übersichtlichkeit halber stellen wir die wichtigsten Fakten noch einmal zusammen:

1. Mit einer beliebigen Norm $\|\cdot\|$ des $\mathbf{R}^n$ ist durch

$$\varrho(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n, \tag{65}$$

eine Abstandsfunktion ϱ definiert, d. h., es gilt für beliebige $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbf{R}^n$

$$\begin{aligned}
&\varrho(\boldsymbol{x}, \boldsymbol{y}) \geqq 0,\\
&\varrho(\boldsymbol{x}, \boldsymbol{y}) = 0 \text{ genau dann, wenn } \boldsymbol{x} = \boldsymbol{y},\\
&\varrho(\boldsymbol{x}, \boldsymbol{y}) = \varrho(\boldsymbol{y}, \boldsymbol{x}),\\
&\varrho(\boldsymbol{x}, \boldsymbol{z}) \leqq \varrho(\boldsymbol{x}, \boldsymbol{y}) + \varrho(\boldsymbol{y}, \boldsymbol{z}).
\end{aligned} \tag{66}$$

Der Beweis folgt unmittelbar aus den Normeigenschaften 5.1.(6).

2. Eine Folge $(\boldsymbol{x}_k)$, $\boldsymbol{x}_k \in \mathbf{R}^n$, konvergiert bezüglich einer Abstandsfunktion ϱ gegen $\boldsymbol{x} \in \mathbf{R}^n$ genau dann, wenn

$$\lim_{k\to\infty} \varrho(\boldsymbol{x}_k, \boldsymbol{x}) = 0$$

ist. Definitionsgemäß heißt $(\boldsymbol{x}_k)$ *Fundamentalfolge* (*Cauchyfolge*), wenn $\varrho(\boldsymbol{x}_k, \boldsymbol{x}_l)$ gegen Null strebt mit $k, l \to \infty$ oder, genauer gesagt,

$$\bigwedge_{\varepsilon>0} \bigvee_{K\in\mathbf{R}} \bigwedge_{k\in\mathbf{N}} \bigwedge_{l\in\mathbf{N}} \left(k, l \geqq K \Rightarrow \varrho(\boldsymbol{x}_k, \boldsymbol{x}_l) < \varepsilon\right) \tag{67}$$

(vgl. MfL Bd. 4, 2.1.7.).

Der Raum $\mathbf{R}^n$ ist bezüglich jeder der Metriken (65) *vollständig*, d. h., es gilt das Cauchysche Konvergenzkriterium, nach dem eine Folge $(\boldsymbol{x}_k)$ dann und nur dann konvergiert, wenn sie Fundamentalfolge ist. Das wurde in MfL Bd. 4, 2.1.7., hinsichtlich der Norm $\|\cdot\|_2$ festgestellt. Um diesen Sachverhalt in der ausgesprochenen Allgemeinheit zu begründen, führen wir in der Menge aller Normen eines linearen normierten Raumes E folgende Äquivalenzrelation ein.

Definition 4. Zwei Normen $\|\cdot\|$ und $\|\cdot\|_*$ des Raumes E heißen *äquivalent* genau dann, wenn

$$\bigvee_{K,L\in\mathbf{R}_+} \bigwedge_{\boldsymbol{x}\in E} (\|\boldsymbol{x}\| \leqq K \|\boldsymbol{x}\|_* \wedge \|\boldsymbol{x}\|_* \leqq L \|\boldsymbol{x}\|). \tag{68}$$

Offensichtlich ist die durch (68) erklärte Relation zwischen Normen reflexiv, symmetrisch und transitiv.

Für $E = \mathbf{R}^n$ gilt

Satz 11. *Zwei beliebige Normen des* $\mathbf{R}^n$ *sind äquivalent.*

Beweis. Es sei $\boldsymbol{e}_j$, $j = 1(1)n$, der Spaltenvektor mit den Koordinaten δ_{ij}, $i = 1(1)n$. Dann gilt für jeden Vektor $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\mathrm{T}$ die Darstellung

$$\boldsymbol{x} = \sum_{j=1}^{n} x_j \boldsymbol{e}_j. \tag{69}$$

Wegen der Transitivität und Symmetrie der Normäquivalenz genügt es zu zeigen, daß jede Norm $\|\cdot\|$ des $\mathbf{R}^n$ der in 5.1.(11) definierten Norm $\|\cdot\|_1$ äquivalent ist. Auf Grund von (69) ist

$$\|\boldsymbol{x}\| \leqq \sum_{j=1}^{n} |x_j| \, \|\boldsymbol{e}_j\|$$

und mit $K := \max\limits_{j\in\{1,2,\ldots,n\}} \|\boldsymbol{e}_j\|$

$$\|\boldsymbol{x}\| \leqq K \|\boldsymbol{x}\|_1.$$

Um die zweite Abschätzung in (68) zu beweisen, betrachten wir die Funktion

$$F(\boldsymbol{x}) = F(x_1, x_2, \ldots, x_n) = \left\| \sum_{j=1}^{n} x_j \boldsymbol{e}_j \right\|.$$

Diese ist für alle $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\mathrm{T}$ stetig, was genauso zu begründen ist wie die Stetigkeit der Größe $\|\boldsymbol{A}\boldsymbol{x}\|$ bei der Erörterung von (52). F ist für $\boldsymbol{x} \neq \boldsymbol{0}$ positiv und nimmt daher nach dem

Satz von BOLZANO-WEIERSTRASS auf der abgeschlossenen beschränkten Menge

$$M = \{y : y \in \mathbf{R}^n \wedge \|y\|_1 = 1\}$$

ein positives Minimum an, das in der Form $\frac{1}{L}$ ausgedrückt sei. Für einen beliebigen Vektor $x \neq 0$ des $\mathbf{R}^n$ gilt dann mit $y_j := \frac{x_j}{\|x\|_1}$, $j = 1(1)n$,

$$y = (y_1, y_2, \ldots, y_n)^T \in M$$

und

$$F(y) = \frac{\|x\|}{\|x\|_1} \geqq \frac{1}{L},$$

also

$$\|x\|_1 \leqq L\,\|x\|.$$

Diese Abschätzung gilt offenbar auch für den Nullvektor.

Wir heben zwei Konsequenzen des Satzes 11 hervor: a) Eine Folge (x_k) des $\mathbf{R}^n$, die in einer Norm $\|\cdot\|$ gegen ein Element des Raumes konvergiert, hat dieses auch bezüglich einer anderen Norm zum Grenzelement. b) Ist (x_k) Cauchyfolge in einer Norm, so auch hinsichtlich jeder anderen.

3. Der in MfL Bd. 4, 2.4.5., bewiesene Banachsche Fixpunktsatz wäre für Funktionen (Operatoren) f, die auf dem $\mathbf{R}^n$ definiert sind, bezüglich einer bestimmten Norm $\|\cdot\|$ dieses Raumes so zu formulieren:

Satz 12. *Ist f eine kontrahierende Abbildung des $\mathbf{R}^n$ in sich, d. h. gilt mit einem gewissen q, $0 \leqq q < 1$, für beliebige $x_1, x_2 \in \mathbf{R}^n$*

$$\|f(x_1) - f(x_2)\| \leqq q\,\|x_1 - x_2\|, \tag{70}$$

so besitzt f genau einen Fixpunkt $\xi \in \mathbf{R}^n$, und die mit einem beliebigen Startpunkt $x_0 \in \mathbf{R}^n$ gebildete Iterationsfolge

$$x_{j+1} := f(x_j) \qquad (j \in \mathbf{N}) \tag{71}$$

konvergiert gegen diesen:

$$\xi = \lim_{j \to \infty} x_j.$$

Satz 12 ergibt sich aus dem in MfL Bd. 4, 2.4.5., erörterten Spezialfall des Banachschen Fixpunktsatzes, wenn man dort $\mathbf{R}$ durch $\mathbf{R}^n$ und in den Abschätzungen Beträge durch Normen ersetzt. Mit dieser Modifikation kann der Beweis wortwörtlich übernommen werden. Das gilt in entsprechender Weise für die in MfL Bd. 9, 4.1.1., behandelte Variante des Prinzips der kontrahierenden Abbildung, speziell für die Herleitung der Fehlerschranken (8) und (9):

Satz 13. *Unter der Voraussetzung des Satzes 12 gelten für jede Iterationsfolge (71) bei der näherungsweisen Bestimmung des Fixpunktes ξ die Fehlerabschätzungen a priori*

bzw. a posteriori

$$\|\xi - x_j\| \leqq \frac{q^j}{1-q} \|x_1 - x_0\| \tag{72}$$

und

$$\|\xi - x_j\| \leqq \frac{q}{1-q} \|x_j - x_{j-1}\|. \tag{73}$$

Der Algorithmus zur näherungsweisen Berechnung des Fixpunktes von f bzw. der Lösung der mit f gebildeten Gleichung zweiter Art (64) gemäß (71) heißt das Verfahren der sukzessiven Approximation(en); ein PAP mit einer nach (73) gebildeten Abbruchbedingung ist in MfL Bd. 9, Abb. 4.1, dargestellt. Man hat nur wieder den Betrag durch die Norm des Kontraktionsoperators und die Wertzuweisungen durch Punktzuweisungen zu ersetzen, die über Feldbelegungen zu realisieren sind.

Wir untersuchen, unter welchen Voraussetzungen die Operatoren (63) kontrahierend sind:

Satz 14. *Die mit dem linearen Gleichungssystem* (61) *gegebenen Operatoren* (63) *sind bezüglich einer Vektornorm* $\|\cdot\|$ *kontrahierend, wenn für eine damit verträgliche Matrixnorm* $|\boldsymbol{B}| < 1$ *gilt;* $q := |\boldsymbol{B}|$ *ist ein Kontraktionsfaktor.*

Beweis. Nach Definition 3 und (54) ist

$$\|f(x_1) - f(x_2)\| = \|\boldsymbol{B}(x_1 - x_2)\| \leqq |\boldsymbol{B}| \cdot \|x_1 - x_2\| = q \|x_1 - x_2\|. \tag{74}$$

Speziell kann man in Satz 14 die durch $\|\cdot\|$ induzierte Matrixnorm betrachten. In Verbindung mit (57) und (60) folgt nun

Satz 15. *Die Operatoren* (63) *sind kontrahierend, wenn für* $\boldsymbol{B} = (b_{ij})$

$$\max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |b_{ij}| < 1 \qquad (\textit{Spaltensummenkriterium}), \tag{75a}$$

$$\max_{i \in \{1,2,\ldots,n\}} \sum_{j=1}^{n} |b_{ij}| < 1 \qquad (\textit{Zeilensummenkriterium}), \tag{75b}$$

$$\left(\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}^2 \right)^{1/2} < 1 \qquad (\textit{Quadratsummenkriterium}) \tag{75c}$$

gilt. Die links vom $<$*-Zeichen stehenden Größen stellen Kontraktionsfaktoren für die Vektornormen*

a) $\|\cdot\|_1$, b) $\|\cdot\|_\infty$ bzw. c) $\|\cdot\|_2$

dar.

Da die Menge der Fixpunkte von f die Lösungsmenge der entsprechenden Gleichung zweiter Art ist, folgt nun auf Grund von Satz 12:

Satz 16. *Die Gleichung* (61) *hat genau eine Lösung* ξ*, wenn für eine mit einer Vektornorm* $\|\cdot\|$ *verträgliche Matrixnorm* $|\cdot|$

$$|\boldsymbol{B}| < 1$$

gilt. Für die mit (63) *gebildete Iterationsfolge* (71) *ist*

$$\|\xi - x_{j+1}\| \leqq |B| \cdot \|\xi - x_j\|.$$

Die Folge (x_j) *konvergiert in jeder Vektornorm gegen* ξ.

Beweis. Nach Satz 14 ist der Operator (63) bezüglich der Norm $\|\cdot\|$ kontrahierend mit dem Kontraktionsfaktor $q := |B|$ und besitzt daher nach Satz 12 genau einen Fixpunkt ξ, welcher zugleich einzige Lösung der Gleichung (61) ist. Subtrahiert man $x_{j+1} = Bx_j + c$ von $\xi = B\xi + c$, so folgt

$$\|\xi - x_{j+1}\| = \|B(\xi - x_j)\| \leqq |B| \cdot \|\xi - x_j\|.$$

(x_j) konvergiert in der Norm $\|\cdot\|$ und dann nach Satz 11 in jeder Norm des $\mathbf{R}^n$ gegen ξ. Die Voraussetzung des Satzes 16 ist speziell erfüllt, wenn eine der Abschätzungen (75) gilt.

Wir wollen noch das Verfahren der sukzessiven Approximation zur Lösung linearer Gleichungssysteme nach dem PAP der Abb. 4.1 in MfL Bd. 9, 4.1.1., in einer ALGOL-Prozedur *SAP*0 darstellen. Als formale Parameter treten auf:

B	zweidimensionales Feld, das der Koeffizientenmatrix (61) entspricht;
c, *x*0, *y*	eindimensionale Felder zur Aufnahme der Komponenten des Vektors *c* in (61), eines Startvektors bzw. der sukzessive zu berechnenden Iterationen;
eps	reelle Variable, die der Fehlerschranke, und
n	ganzzahlige Variable, welche der Anzahl der Gleichungen oder Unbekannten in (61) zuzuordnen ist.

Der Algorithmus wird mit der Fehlerschranke (73) als Abbruchbedingung programmiert, und zwar bezüglich der durch $\|\cdot\|_\infty$ definierten Metrik. Die Abstandsbestimmung erfolgt mit Hilfe einer lokal vereinbarten Prozedur *DISTANZ*, dabei wird die Abweichung aufeinanderfolgender Näherungen der Variablen *m* zugewiesen. Im übrigen ist *SAP*0 analog der Prozedur *SAP* in MfL Bd. 9, 4.1.1., aufgebaut, wobei jedoch der Kontraktionsfaktor nicht über einen formalen Parameter vermittelt, sondern gemäß (75 b) intern mit Hilfe der lokalen Größen *m*, *p* berechnet wird. Das lokal vereinbarte eindimensionale Feld *x* entspricht der Variablen *x* in *SAP* und dient zum Zwischenspeichern von Vektoren der Iterationsfolge.

```
procedure SAP0(B,x0,c,y,eps,n);
value eps; integer n; real eps; array B,x0,c,y;
begin
    integer i,j; real m,p; array x[1:n];
    procedure DISTANZ;
    begin integer i; real h;
      m := abs(y[1] - x[1]);
      for i := 2 step 1 until n do          begin
      h := abs(y[i] - x[i]);
```

```
        if m < h then m := h                end
    end;
    p := 0;
    for i := 1 step 1 until n do    begin m := 0;
    for j := 1 step 1 until n do  m := m + abs(B[i,j]);
    if m > p then p := m            end;
    p := p/(1 - p);
    for i := 1 step 1 until n do x[i] := x0[i];
L:  for i := 1 step 1 until n do    begin
    y[i] := c[i];
    for j := 1 step 1 until n do y[i] := y[i] + B[i,j] × x[j]
                                        end;
    DISTANZ;
    if p × m ≧ eps then  begin
    for i := 1 step 1 until n do x[i] := y[i];
    goto L                      end
end
```

Beispiel. Wir betrachten das lineare Gleichungssystem 5.4.(51), das bei Interpolation mit kubischen Splines zu lösen ist. Dieses umfaßt bei n Knoten $n-2$ Gleichungen. Nimmt man die Umformung (62) vor, so ergibt sich eine Gleichung zweiter Art mit der Koeffizientenmatrix

$$B = \begin{pmatrix} 0 & -\frac{1}{2}\frac{x_3 - x_2}{x_3 - x_1} & & & \\ -\frac{1}{2}\frac{x_3 - x_2}{x_4 - x_2} & 0 & -\frac{1}{2}\frac{x_4 - x_3}{x_4 - x_2} & & 0 \\ & & & & \\ & & & & \\ 0 & & -\frac{1}{2}\frac{x_{n-2} - x_{n-3}}{x_{n-1} - x_{n-3}} & 0 & -\frac{1}{2}\frac{x_{n-1} - x_{n-2}}{x_{n-1} - x_{n-3}} \\ & & & -\frac{1}{2}\frac{x_{n-1} - x_{n-2}}{x_n - x_{n-2}} & 0 \end{pmatrix} \tag{76}$$

In (76) bedeuten die x_i, $i = 1(1)n$, gemäß früherer Bezeichnung die Knoten 5.4.(23) und dürfen nicht mit den Unbekannten des linearen Gleichungssystems verwechselt werden. Offenbar sind das Spalten- und Zeilensummenkriterium erfüllt. Der durch die linken Seiten von (75a), (75b) bestimmte Kontraktionsfaktor ist $q = \frac{1}{2}$.

Bei äquidistanten Knoten haben die von Null verschiedenen Elemente in (76) den Wert $-\frac{1}{4}$. Für die am Schluß von 6.1.2. formulierte Interpolationsaufgabe findet man mit *SAP0* die in Tabelle 6.4 erfaßte Näherungslösung des entsprechenden linearen Systems, wenn das Verfahren mit dem Nullvektor gestartet und $eps = 10^{-5}$ gesetzt wird. Wie in Tabelle 6.3 sind die Vektorkoordinaten mit k, $k = 1(1)7$, indiziert; j zählt die Iterationen. Der Abbruch erfolgt bei $j = 15$.

k $j\rightarrow$ ↓	1	2	...	13	14	15
1	0,007558	−0,002697	...	0,029002	0,029007	0,029008
2	0,041017	−0,022637	...	−0,085796	−0,085803	−0,085807
3	0,247059	0,386804	...	0,478279	0,478290	0,478294
4	−0,600000	−0,723529	...	−0,839130	−0,839140	−0,839145
5	0,247059	0,386804	...	0,478279	0,478290	0,478294
6	0,041017	−0,022637	...	−0,085796	−0,085803	−0,085807
7	0,007558	−0,002697	...	0,029002	0,029007	0,029008

Tabelle 6.4

Bei der Bildung der sukzessiven Approximationen gemäß (71) berechnet man die Komponenten von $\boldsymbol{x}_{j+1}$ durch Einsetzen der Komponenten von $\boldsymbol{x}_j$ in den Matrizenausdruck (63). Dieses Vorgehen wird als *Gesamtschrittverfahren* bezeichnet. Statt dessen ließe sich eine Vektorfolge konstruieren, bei der man in die Berechnung der i-ten Komponente von $\boldsymbol{x}_{j+1}$ die jeweils schon vorliegenden Komponenten von $\boldsymbol{x}_{j+1}$ einbezieht. Bezeichnen wir die i-ten Komponenten des Vektors $\boldsymbol{x}_j$ mit $x_i^{(j)}$, so werden bei diesem sogenannten *Einzelschrittverfahren* die Iterationen nach der Vorschrift

$$x_i^{(j+1)} = \sum_{k=1}^{i-1} b_{ik} x_k^{(j+1)} + \sum_{k=i}^{n} b_{ik} x_k^{(j)} + c_i \tag{77}$$

an Stelle von

$$x_i^{(j+1)} = \sum_{k=1}^{n} b_{ik} x_k^{(j)} + c_i \tag{78}$$

beim Gesamtschrittverfahren gebildet. Gemäß (77) läßt sich die Bildung von $\boldsymbol{x}_{j+1}$ aus $\boldsymbol{x}_j$ in die Zwischenstufen

$$\begin{pmatrix} x_1^{(j+1)} \\ x_2^{(j)} \\ x_3^{(j)} \\ \vdots \\ x_n^{(j)} \end{pmatrix}, \quad \begin{pmatrix} x_1^{(j+1)} \\ x_2^{(j+1)} \\ x_3^{(j)} \\ \vdots \\ x_n^{(j)} \end{pmatrix}, \quad \ldots, \quad \begin{pmatrix} x_1^{(j+1)} \\ x_2^{(j+1)} \\ x_3^{(j+1)} \\ \vdots \\ x_n^{(j+1)} \end{pmatrix} \tag{79}$$

zerlegen. Weiterhin wird jedoch der Übergang von $\boldsymbol{x}_j$ zu $\boldsymbol{x}_{j+1}$ als ein Schritt des Verfahrens betrachtet.

Das Einzelschrittverfahren kann als sukzessive Approximation gemäß (71) mit einem geeignet gewählten Operator (63) aufgefaßt werden. Stellt man (61) in der Form

$$\boldsymbol{x} = (\boldsymbol{L} + \boldsymbol{U})\,\boldsymbol{x} + \boldsymbol{c} \tag{80}$$

mit

$$\boldsymbol{L} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ b_{21} & 0 & \dots & 0 \\ b_{31} & b_{32} & 0 \dots & \\ \dots & \dots & \dots & \\ b_{n1} & b_{n2} & \dots\ b_{n,n-1} & 0 \end{pmatrix} \quad \text{und} \quad \boldsymbol{U} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ 0 & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \\ 0 & 0 & \dots & b_{nn} \end{pmatrix} \tag{81}$$

dar, so läßt sich die Iterationsvorschrift (77) in der Matrizenform

$$\boldsymbol{x}_{j+1} = \boldsymbol{L}\boldsymbol{x}_{j+1} + \boldsymbol{U}\boldsymbol{x}_j + \boldsymbol{c}$$

ausdrücken. Mit $\boldsymbol{I}$ als Einheitsmatrix gewinnt man nach Zusammenfassung der Glieder, die $\boldsymbol{x}_{j+1}$ enthalten,

$$\boldsymbol{x}_{j+1} = (\boldsymbol{I} - \boldsymbol{L})^{-1}\,\boldsymbol{U}\boldsymbol{x}_j + (\boldsymbol{I} - \boldsymbol{L})^{-1}\,\boldsymbol{c}; \tag{82}$$

die Matrix $\boldsymbol{I} - \boldsymbol{L}$ ist wegen $|\boldsymbol{I} - \boldsymbol{L}| = 1$ regulär. (82) stellt die zur Gleichung

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{L})^{-1}\,\boldsymbol{U}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{L})^{-1}\,\boldsymbol{c} \tag{83}$$

gebildeten sukzessiven Approximationen dar. Offenbar ist (83) mit

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c}$$

äquivalent.

Der folgende Satz beinhaltet ein hinreichendes Kriterium für die Konvergenz des Einzelschrittverfahrens ([12], § 32).

Satz 17. *Ist in* (61)

$$|\boldsymbol{B}|_\infty = \max_{i \in \{1,2,\dots,n\}} \sum_{j=1}^{n} |b_{ij}| \leqq \mu < 1,$$

so konvergiert das Einzelschrittverfahren für jeden Startvektor gegen die wohlbestimmte Lösung dieser Gleichung.

Beweis. Satz 16 besagt, daß (61) unter dieser Voraussetzung genau eine Lösung ξ besitzt. Für diese gilt

$$\xi_i = \sum_{k=1}^{n} b_{ik}\xi_k + c_i, \qquad i = 1(1)n.$$

Subtrahiert man davon (77), so folgt

$$\xi_i - x_i^{(j+1)} = \sum_{k=1}^{i-1} b_{ik}(\xi_k - x_k^{(j+1)}) + \sum_{k=i}^{n} b_{ik}(\xi_k - x_k^{(j)})$$

und

$$|\xi_i - x_i^{(j+1)}| \leqq \sum_{k=1}^{i-1} |b_{ik}| \cdot |\xi_k - x_k^{(j+1)}| + \sum_{k=i}^{n} |b_{ik}| \cdot |\xi_k - x_k^{(j)}|. \tag{84}$$

Mit

$$\beta_1 := 0, \quad \beta_i := \sum_{k=1}^{i-1} |b_{ik}|, \quad i = 2(1)n, \quad \gamma_i := \sum_{k=i}^{n} |b_{ik}|, \quad i = 1(1)n,$$

gewinnt man aus (84)

$$|\xi_i - x_i^{(j+1)}| \leqq \beta_i \, \|\xi - x_{j+1}\|_\infty + \gamma_i \, \|\xi - x_j\|_\infty. \tag{85}$$

Das Maximum der Beträge $|\xi_i - x_i^{(j+1)}|$ möge für $i = i_0$ angenommen werden, d. h., es sei

$$|\xi_{i_0} - x_{i_0}^{(j+1)}| = \|\xi - x_{j+1}\|_\infty.$$

Dann liefert (85) für $i = i_0$

$$\|\xi - x_{j+1}\|_\infty \leqq \beta_{i_0} \, \|\xi - x_{j+1}\|_\infty + \gamma_{i_0} \, \|\xi - x_j\|_\infty,$$

also

$$\|\xi - x_{j+1}\|_\infty \leqq \frac{\gamma_{i_0}}{1 - \beta_{i_0}} \, \|\xi - x_j\|_\infty. \tag{86}$$

Man beachte, daß β_i, $i = 1(1)n$, auf Grund der Voraussetzung kleiner als Eins ist. Mit

$$\rho := \max_{i \in \{1,2,\ldots,n\}} \frac{\gamma_i}{1 - \beta_i} \tag{87}$$

resultiert aus (86)

$$\|\xi - x_{j+1}\|_\infty \leqq \rho \, \|\xi - x_j\|_\infty. \tag{88}$$

Die Konvergenz der mit dem Einzelschrittverfahren erzeugten Iterationsfolge beruht auf der Beziehung

$$\rho \leqq \mu < 1. \tag{89}$$

In der Tat folgt aus den Abschätzungen

$$\beta_i + \gamma_i = \sum_{k=1}^{n} |b_{ik}| \leqq \mu$$

und

$$\beta_i + \gamma_i - \frac{\gamma_i}{1 - \beta_i} = \frac{\beta_i(1 - \beta_i - \gamma_i)}{1 - \beta_i} \geqq 0,$$

daß

$$\mu \geqq \max_{i \in \{1,2,\ldots,n\}} (\beta_i + \gamma_i) \geqq \max_{i \in \{1,2,\ldots,n\}} \frac{\gamma_i}{1 - \beta_i} = \rho$$

ist. Damit gilt

$$\|\xi - x_{j+1}\|_\infty \leqq \rho^{j+1} \, \|\xi - x_0\|_\infty$$

und wegen (89)

$$\lim_{j \to \infty} \|\xi - x_{j+1}\|_\infty = 0.$$

Aus (88) läßt sich eine Fehlerabschätzung herleiten, die nur Elemente der Iterationsfolge benutzt und sich somit zur Formulierung einer Abbruchbedingung eignet. Zunächst ist

$$\|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|_\infty = \|\boldsymbol{\xi} - \boldsymbol{x}_j - (\boldsymbol{\xi} - \boldsymbol{x}_{j+1})\|_\infty \geqq | \, \|\boldsymbol{\xi} - \boldsymbol{x}_j\|_\infty - \|\boldsymbol{\xi} - \boldsymbol{x}_{j+1}\|_\infty |.$$

Mit (88) folgt daraus unter Beachtung von (89)

$$\|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|_\infty \geqq \|\boldsymbol{\xi} - \boldsymbol{x}_j\|_\infty (1 - \mu)$$

und nach Multiplikation mit μ die Fehlerabschätzung a posteriori

$$\|\boldsymbol{\xi} - \boldsymbol{x}_{j+1}\|_\infty \leqq \frac{\mu}{1 - \mu} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|_\infty. \tag{90}$$

Die folgende aus *SAP*0 entwickelte Prozedur *SAP*1 realisiert das Einzelschrittverfahren. Die Einführung der Hilfsvariablen s ist durch den Unterschied in den Iterationsvorschriften (77) und (78) bedingt. Die Abbruchbedingung wurde gemäß (90) gebildet, wobei μ mit Benutzung der lokalen Variablen p, s, m innerhalb der Prozedur berechnet wird. Danach ist p mit der Größe $\mu/(1 - \mu)$ belegt. Wie in *SAP*0 bestimmt die lokale Prozedur *DISTANZ* die Abweichungen aufeinanderfolgender Vektoren einer Iterationsfolge und aktualisiert damit die Variable m.

```
procedure SAP1(B,c,x0,y,eps,n);
value eps; integer n; real eps; array B,c,x0,y;
begin
      integer i,j; real p,s,m; array x[1:n];
      procedure DISTANZ;
      begin
      ⋮
      end;
      p := 0;
      for j := 1 step 1 until n do p := p + abs(B[1,j]);
      for i := 2 step 1 until n do  begin m := s := 0;
      for j := 1 step 1 until i - 1 do  m := m + abs(B[i,j]);
      for j := i step 1 until n do      s := s + abs(B[i,j]);
      m := s/(1 - m);
      if m > p then p := m          end;
      p := p/(1 - p);
      for i := 1 step 1 until n do y[i] := x[i] := x0[i];
L:    for i := 1 step 1 until n do  begin s := c[i];
      for j := 1 step 1 until n do  s := s + B[i,j] × y[j];
      y[i] := s                     end;
      DISTANZ;
      if p × m ≧ eps then begin
      for i := 1 step 1 until n do x[i] := y[i]; goto L
                                 end
end
```

Zum Vergleich mit den Ergebnissen der Tabelle 6.4 wurde das Beispiel noch einmal mit *SAP*1 gerechnet. Gemäß (87) hat μ hier den Wert $\frac{1}{3}$ und ist damit kleiner als $\|B\|_\infty = \frac{1}{2}$. Die Werte μ und $\|B\|_\infty$ bestimmen nach (88) bzw. Satz 16 die Konvergenzgeschwindigkeit in *SAP*1 und *SAP*0. Einige der nach dem Einzelschrittverfahren berechneten Näherungen sind in Tabelle 6.5 erfaßt. Der Algorithmus wurde mit dem Nullvektor gestartet; für $eps := 10^{-5}$ erfolgte der Abbruch bei $j = 9$.

k ↓ $j\rightarrow$	1	2	...	7	8	9
1	0,007558	−0,002224	...	0,028957	0,028998	0,029008
2	0,039128	−0,017746	...	−0,085764	−0,085800	−0,085808
3	0,237277	0,416325	...	0,478271	0,478293	0,478297
4	−0,659319	0,807023	...	−0,839135	−0,839146	−0,839149
5	0,411889	0,464311	...	0,478293	0,478297	0,478298
6	−0,061955	−0,080822	...	−0,085808	−0,085809	−0,085810
7	0,023046	0,027736	...	0,029010	0,029010	0,029010

Tabelle 6.5

6.1.4. Untersuchung des Rechenaufwandes und Fehlerbetrachtungen

Für die Bewertung eines Verfahrens (Algorithmus) der numerischen Mathematik sind wesentlich zwei Gesichtspunkte maßgebend: die *Genauigkeit der Ergebnisse* und der *Rechenaufwand*. Letzterer drückt sich in dem Speicherplatzbedarf und der erforderlichen Rechenzeit aus, die wiederum von der Anzahl der auszuführenden arithmetischen Operationen abhängt. Da bei jeder dieser Operationen Rundungsfehler auftreten können, welche die Genauigkeit des Resultates beeinflussen, müssen die genannten beiden Aspekte im Zusammenhang gesehen werden.

Wir beginnen damit, die Verfahren zur Lösung linearer Gleichungssysteme hinsichtlich der erforderlichen arithmetischen Operationen zu vergleichen. Deren Anzahl ist bei den iterativen Verfahren das Produkt aus der Anzahl der Schritte und der in einem solchen auszuführenden Operationen. Betrachten wir wie bisher n lineare Gleichungen mit n Unbekannten, so ergeben sich zum Beispiel in einem Iterationsschritt im allgemeinen n^2 Multiplikationen. Wir werden uns darauf beschränken, die Multiplikationen und Divisionen (weiterhin „Operationen" genannt) zu zählen. Damit gewinnt man ein Maß für die Rechenzeit, da in den meisten Verfahren zur Lösung linearer Gleichungssysteme diese Operationen etwa gleich oft wie Additionen und Subtraktionen auftreten und moderne Rechner — grob geschätzt — dafür das Doppelte der für die letzteren zu veranschlagenden Zeit benötigen. Die Anzahl der Operationen wird mit α_n bezeichnet.

Bei der Erörterung der direkten Verfahren wollen wir zunächst α_n für den Fall bestimmen, daß man zu der als regulär angenommenen Koeffizientenmatrix A die

Inverse bildet und (2) gemäß der Formel

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{91}$$

löst. Legen wir der Bildung der Adjunkten von $\boldsymbol{A}$ die Leibnizsche Berechnung von Determinanten zugrunde, so ergeben sich insgesamt $n^2(n-2)\,(n-1)!$ Multiplikationen. Denen sind noch n Multiplikationen bei der Bestimmung von $|\boldsymbol{A}|$ nach dem Entwicklungssatz und weitere n^2 bei der Bildung von $\boldsymbol{A}^{-1}\boldsymbol{b}$ hinzuzufügen. Damit ergibt sich

$$\lim_{n\to\infty} \frac{\alpha_n}{(n-1)!\,n^3} = 1,$$

wofür man auch

$$\alpha_n \cong (n-1)!\,n^3 \tag{92}$$

schreibt. Für $n = 10$ hat die auf der rechten Seite von (92) stehende Größe den Wert $3{,}6288 \cdot 10^8$. Offensichtlich ist das betrachtete Verfahren schon für relativ kleine Systeme numerisch unbrauchbar wegen der Akkumulation von Rundungsfehlern und unrealistischen Rechenzeiten.

Wir vergleichen damit den Rechenaufwand beim Gaußschen Verfahren, wobei der Einfachheit halber angenommen sei, daß keine Zeilenvertauschungen und keine Spaltenvertauschungen vorgenommen werden. Im k-ten Eliminationsschritt hat man bei der Umformung der Koeffizientenmatrix auf Dreiecksgestalt gemäß (20a), (20b) $n-k$ Quotienten $A[i,k]/A[k,k]$, $i = k+1(1)n$, zu bilden und damit die Elemente $A[k,j]$ der k-ten Zeile für $j = k+1(1)n$ zu multiplizieren. Wie in der Prozedur *GAUSS* mag man sich diese Quotienten in der k-ten Spalte unterhalb der Hauptdiagonalen gespeichert denken. Dabei läuft k von 1 bis $n-1$, so daß sich insgesamt

$$\beta_n := \sum_{k=1}^{n-1}(n-k) + \sum_{k=1}^{n-1}(n-k)^2 = \sum_{\nu=1}^{n-1}\nu + \sum_{\nu=1}^{n-1}\nu^2 \qquad (\nu = n-k) \tag{93}$$

Operationen ergeben. Mit Hilfe der durch vollständige Induktion leicht beweisbaren Formeln

$$\sum_{\nu=1}^{n}\nu = \frac{n(n+1)}{2}, \qquad \sum_{\nu=1}^{n}\nu^2 = \frac{n(n+1)\,(2n+1)}{6} \tag{94}$$

gewinnt man

$$\beta_n = \frac{n(n^2-1)}{3}. \tag{95}$$

Dem sind im k-ten Schritt $n-k$ Multiplikationen zur Umformung der rechten Seite des linearen Gleichungssystems, insgesamt also

$$\gamma_n = \frac{(n-1)\,n}{2} \tag{96}$$

Operationen hinzuzufügen. Schließlich erfordert die Lösung des trianguliertен Systems weitere

$$\delta_n = \frac{n(n+1)}{2} \tag{97}$$

Operationen. In der Tat fallen $n - k$ Multiplikationen und eine Division bei der Berechnung von x_k an, woraus sich durch Summation über k von 1 bis n (97) ergibt. Betrachten wir weiterhin den Fall, daß m Gleichungssysteme (2) mit derselben Koeffizientenmatrix für verschiedene rechte Seiten zu lösen sind. Dann hat man die Größen γ_n und δ_n noch mit dem Faktor m zu versehen, so daß sich insgesamt

$$\alpha_n = \frac{n^3}{3} + mn^2 - \frac{n}{3} \tag{98}$$

Operationen ergeben. Zum Vergleich mit (92): Für $n = 10$ und $m = 1$ ist $\alpha_n = 430$.

Nach MfL Bd. 3, 6.5., läßt sich die zu einer regulären Matrix $\boldsymbol{A}$ inverse Matrix $\boldsymbol{A}^{-1}$ aus den Lösungen der linearen Gleichungssysteme $\boldsymbol{Ax} = \boldsymbol{e}_j$, $j = 1(1)n$, bilden, wobei $\boldsymbol{e}_j$ den Spaltenvektor mit den Koordinaten δ_{ij} bedeutet. Dazu sind gemäß (98)

$$\alpha_n = \frac{4}{3} n^3 - \frac{n}{3} \tag{99}$$

Operationen erforderlich. Dieser Aufwand läßt sich reduzieren, wenn man zur Lösung der betrachteten Gleichungssysteme eine Variante des Gaußschen Algorithmus benutzt, welche die spezielle Form der rechten Seiten berücksichtigt.

Wir wollen noch die Auswirkungen von Ungenauigkeiten in den Koeffizienten und der rechten Seite des linearen Gleichungssystems (2) untersuchen. Es sei $\boldsymbol{x}$ die Lösung von (2) und $\boldsymbol{x} + \delta\boldsymbol{x}$ die eines gestörten Systems

$$(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \boldsymbol{b}\delta; \tag{100}$$

$\delta\boldsymbol{A}$ und $\delta\boldsymbol{b}$ sind Matrizen vom Typ $n \times n$ bzw. $n \times 1$, deren Komponenten als Inkremente der exakt bestimmten Koeffizienten und rechten Seiten aufzufassen sind. Das Ergebnis der folgenden Erörterung wird eine Abschätzung der relativen Fehlergröße

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \tag{101}$$

sein, wobei $\|\cdot\|$ eine Norm des $\mathbf{R}^n$ bedeutet. Dazu beweisen wir den

Hilfssatz 1. *Für eine Matrix $\boldsymbol{B}$ des Typs $n \times n$ sei bezüglich der durch $\|\cdot\|$ induzierten Matrixnorm $|\boldsymbol{B}| < 1$. Dann ist $\boldsymbol{I} - \boldsymbol{B}$ regulär, und es gilt*

$$\frac{1}{1 + |\boldsymbol{B}|} \leqq |(\boldsymbol{I} - \boldsymbol{B})^{-1}| \leqq \frac{1}{1 - |\boldsymbol{B}|}. \tag{102}$$

Beweis. Nach Satz 16 hat die Gleichung $(\boldsymbol{I} - \boldsymbol{B})\,\boldsymbol{x} = \boldsymbol{c}$ für beliebiges $\boldsymbol{c} \in \mathbb{R}^n$ genau eine Lösung ξ. Daraus folgt die Regularität von $\boldsymbol{I} - \boldsymbol{B}$. Nach (53) gewinnt man

$$|\boldsymbol{I}| = 1 \tag{103}$$

und wegen $\boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{B})\,(\boldsymbol{I} - \boldsymbol{B})^{-1}$ auf Grund von (50)

$$1 \leqq |\boldsymbol{I} - \boldsymbol{B}|\,|(\boldsymbol{I} - \boldsymbol{B})^{-1}| \leqq (1 + |\boldsymbol{B}|)\,|(\boldsymbol{I} - \boldsymbol{B})^{-1}|.$$

Damit ist der linke Teil von (102) bewiesen. Die andere Abschätzung erhält man auf Grund von (103) gemäß

$$(\boldsymbol{I} - \boldsymbol{B})^{-1} = \boldsymbol{I} + \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{B})^{-1},$$

$$|(\boldsymbol{I} - \boldsymbol{B})^{-1}| \leqq 1 + |\boldsymbol{B}|\,|(\boldsymbol{I} - \boldsymbol{B})^{-1}|,$$

$$|(\boldsymbol{I} - \boldsymbol{B})^{-1}| \leqq \frac{1}{1 - |\boldsymbol{B}|}.$$

Im Hinblick auf das oben formulierte Problem nehmen wir nun an, daß die Störung der Matrix $\boldsymbol{A}$ so klein sei, daß

$$|\delta\boldsymbol{A}| < 1/|\boldsymbol{A}^{-1}|. \tag{104}$$

Dann ist $|\boldsymbol{A}^{-1}\delta\boldsymbol{A}| < 1$ und der Hilfssatz 1 für $\boldsymbol{B} := \boldsymbol{A}^{-1}\delta\boldsymbol{A}$ anwendbar, d. h., $\boldsymbol{I} - \boldsymbol{A}^{-1}\delta\boldsymbol{A}$ ist regulär, und es gilt

$$|(\boldsymbol{I} - \boldsymbol{A}^{-1}\delta\boldsymbol{A})^{-1}| \leqq \frac{1}{1 - |\boldsymbol{A}^{-1}\delta\boldsymbol{A}|} \leqq \frac{1}{1 - |\boldsymbol{A}^{-1}| \cdot |\delta\boldsymbol{A}|}. \tag{105}$$

Durch linksseitige Multiplikation der Gleichung (100) mit $\boldsymbol{A}^{-1}$ gewinnt man

$$(\boldsymbol{I} - \boldsymbol{A}^{-1}\delta\boldsymbol{A})\,(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{A}^{-1}\delta\boldsymbol{b}$$

und unter Beachtung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

$$\delta\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A}^{-1}\delta\boldsymbol{A})^{-1}\,\boldsymbol{A}^{-1}(\delta\boldsymbol{A}\boldsymbol{x} + \delta\boldsymbol{b}).$$

Mit (105) folgt daraus die Abschätzung

$$\|\delta\boldsymbol{x}\| \leqq \frac{|\boldsymbol{A}^{-1}| \cdot \|\delta\boldsymbol{A}\boldsymbol{x} + \delta\boldsymbol{b}\|}{1 - |\boldsymbol{A}^{-1}| \cdot |\delta\boldsymbol{A}|}$$

oder

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq \frac{|\boldsymbol{A}^{-1}|\,(|\delta\boldsymbol{A}| + \|\delta\boldsymbol{b}\|/\|\boldsymbol{x}\|)}{1 - |\boldsymbol{A}^{-1}| \cdot |\delta\boldsymbol{A}|}.$$

Wegen $|\boldsymbol{A}| \cdot \|\boldsymbol{x}\| \geqq \|\boldsymbol{b}\|$ kann auf der rechten Seite der Term $\dfrac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|}$ majorisierend durch $\dfrac{\|\delta\boldsymbol{b}\| \cdot |\boldsymbol{A}|}{\|\boldsymbol{b}\|}$ ersetzt werden. Damit haben wir folgenden Satz [23] gewonnen, in dem wie bisher $|\cdot|$ die durch eine Vektornorm $\|\cdot\|$ induzierte Matrixnorm bedeutet:

Satz 18. *A sei eine reguläre Matrix, und für die Störmatrix δA gleichen Typs gelte* (104). *Genügen dann x und δx den Gleichungen* (2) *bzw.* (100), *so ist*

$$\frac{\|\delta x\|}{\|x\|} \leqq \frac{\mu}{1-\mu\,|\delta A|/|A|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{|\delta A|}{|A|}\right) \tag{106}$$

mit

$$\mu = \mu(A) := |A| \cdot |A^{-1}|. \tag{107}$$

Die Quotienten $\frac{\|\delta x\|}{\|x\|}$, $\frac{\|\delta b\|}{\|b\|}$ und $\frac{|\delta A|}{|A|}$ sind analog dem relativen Fehler eines Näherungswertes gebildet. Bezüglich dieser relativen Änderungen läßt sich die Sachlage so einschätzen:

> Kleinen relativen Änderungen in den Daten von (2) (Koeffizientenmatrix und rechte Seite) entspricht eine kleine relative Änderung der Lösung, sofern die sogenannte *Konditionszahl* μ (107) der Matrix A klein ist. Ist letzteres nicht der Fall, so nennt man *A schlecht konditioniert.*

Wir entnehmen [12] folgendes Beispiel für eine solche Matrix:

$$A = \begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix}.$$

Dafür ist $|A| = 1$ und

$$A^{-1} = \begin{pmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{pmatrix}.$$

Bezüglich $\|\cdot\|_\infty$ findet man nach (57b) $|A|_\infty = 33$, $|A^{-1}|_\infty = 136$ und $\mu(A) = 4488$.

Wir schließen mit einer Fehlerbetrachtung zu den Iterationsverfahren. Diese konvergieren unter geeigneten Voraussetzungen bei exakter Berechnung der sukzessiven Approximationen für jeden Startvektor gegen die wohlbestimmte Lösung ξ des linearen Gleichungssystems. Sie sind dann insofern selbstkorrigierend, als ein fehlerbehafteter Vektor in der Iterationsfolge (x_j) stets wieder als Startvektor einer neuen Iterationsfolge aufgefaßt werden kann und so die Konvergenz von (x_j) gegen ξ zwar verlangsamen, aber nicht verhindern kann. Die Dinge liegen anders, wenn mit Hilfe des Operators (63) nur Näherungen x_j^* der sukzessiven Approximationen x_j berechnet werden. In diesem Falle gilt ein MfL, Band 9, 4.1.3.(42), entsprechendes Resultat.

Bezüglich der Gleichung (64) wird vorausgesetzt, daß für eine mit einer Vektornorm $\|\cdot\|$ verträglichen Matrixnorm $|B| < 1$ gilt. Wie in MfL Bd. 9, 4.1.3.(40),

stellen wir die Näherungen x_j^* in der Form

$$x_{j+1}^* = f(x_j^*) + h_j \qquad (j = 0, 1, 2, \ldots) \tag{108}$$

dar und nehmen an, daß $\|h_j\| < \delta$ mit einer festen positiven Konstanten δ gilt; $f(x)$ bedeutet stets den mit dem Operator (63) zu x exakt gebildeten Vektor. Nach Satz 16 gilt für genau ein $\xi \in \mathbf{R}^n$

$$\xi = f(\xi).$$

Durch Subtraktion dieser Gleichung von (108) gewinnt man

$$\begin{aligned} \|x_{j+1}^* - \xi\| &= \|B(x_j^* - \xi) + h_j\| \leqq |B| \cdot \|x_j^* - \xi\| + \|h_j\| \\ &\leqq |B| \cdot \|x_j^* - \xi\| + \delta \end{aligned}$$

und nach wiederholter Anwendung dieser Abschätzung mit $q := |B|$

$$\begin{aligned} \|x_{j+1}^* - \xi\| &\leqq q\|x_j^* - \xi\| + \delta \leqq q^2 \|x_{j-1}^* - \xi\| + q\delta + \delta \leqq \cdots \\ &\leqq q^{j+1} \|x_0^* - \xi\| + (q^j + q^{j-1} + \cdots + 1)\,\delta \\ &\leqq q^{j+1} \|x_0^* - \xi\| + \frac{\delta}{1-q}. \end{aligned} \tag{109}$$

Aus (109) kann nicht mehr auf die Konvergenz der Folge (x_j^*) gegen ξ geschlossen werden, wohl aber gilt

$$\limsup_{j \to \infty} \|x_j^* - \xi\| \leqq \frac{\delta}{1-q},$$

und dies besagt, daß man ξ mit der Folge (x_j^*) angenähert bestimmen kann, sofern keine höhere Genauigkeit als $\dfrac{\delta}{1-q}$ gefordert ist.

6.2. Nichtlineare Gleichungen

6.2.1. Iterative Lösung nichtlinearer Gleichungssysteme

Es seien g_i, $i = 1(1)n$, reellwertige Funktionen von n reellen Veränderlichen, die in einem Gebiet G des $\mathbf{R}^n$ definiert sind. Entsprechend MfL Bd. 9, 4.1.(1), und in Verallgemeinerung von 6.1.(2) ($m = n$) betrachten wir das Gleichungssystem

$$\begin{aligned} &g_1(x_1, x_2, \ldots, x_n) = g_1(x) = 0, \\ &g_2(x_1, x_2, \ldots, x_n) = g_2(x) = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &g_n(x_1, x_2, \ldots, x_n) = g_n(x) = 0 \end{aligned} \tag{1}$$

oder in Vektorform die Gleichung

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0} \tag{2}$$

mit

$$\boldsymbol{g}(\boldsymbol{x}) = \left(g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x})\right)^{\mathsf{T}}. \tag{3}$$

Ohne eine weitergehende Spezifizierung der Funktion g_i ist die Erörterung exakter Verfahren wie bei den linearen Systemen offenbar gegenstandslos. Wir wenden uns daher von vornherein iterativen Methoden zu, die sich auf Gleichungen zweiter Art beziehen. Diese haben die Gestalt

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}, \tag{4}$$

wobei jetzt

$$\boldsymbol{f}(\boldsymbol{x}) = \left(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})\right)^{\mathsf{T}} \tag{5}$$

und die

$$f_i(\boldsymbol{x}): G \to \mathbf{R}, \qquad i = 1(1)n, \tag{6}$$

in einem Gebiet G des $\mathbf{R}^n$ definierte reellwertige Funktionen sind.

Grundlage für die iterative Lösung von (4) ist eine Verallgemeinerung des Satzes 1 aus MfL Bd. 9, 4.1. Dabei fassen wir den $\mathbf{R}^n$ als vollständigen metrischen Raum (vgl. MfL Bd. 4, 2.1.7.) bezüglich einer mit einer Norm $\|\cdot\|$ gebildeten Distanzfunktion ϱ auf, d. h., wir definieren

$$\varrho(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{7}$$

Dann gilt

Satz 1. *Für $r > 0$ und $\boldsymbol{z} \in \mathbf{R}^n$ sei*

$$\overline{U}_r(\boldsymbol{z}) = \{\boldsymbol{x}: \boldsymbol{x} \in \mathbf{R}^n \wedge \varrho(\boldsymbol{x}, \boldsymbol{z}) \leqq r\} \tag{8}$$

die abgeschlossene r-Umgebung des Punktes $\boldsymbol{z}$ und $\boldsymbol{f}: \overline{U}_r(\boldsymbol{z}) \to \mathbf{R}^n$ kontrahierende Abbildung zum Kontraktionsfaktor q $(0 \leqq q < 1)$, d. h., es ist für alle $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \overline{U}_r(\boldsymbol{z})$

$$\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\| \leqq q\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|; \tag{9}$$

ferner gelte

$$\|\boldsymbol{z} - \boldsymbol{f}(\boldsymbol{z})\| \leqq (1 - q)\,r. \tag{10}$$

Dann besitzt $\boldsymbol{f}$ genau einen Fixpunkt $\boldsymbol{\xi} \in \overline{U}_r(\boldsymbol{z})$, und die mit einem beliebigen Startpunkt $\boldsymbol{x}_0 \in \overline{U}_r(\boldsymbol{z})$ gebildete Iterationsfolge

$$\boldsymbol{x}_{j+1} := \boldsymbol{f}(\boldsymbol{x}_j) \qquad (j \in \mathbf{N}) \tag{11}$$

konvergiert gegen diesen:

$$\boldsymbol{\xi} = \lim_{j \to \infty} \boldsymbol{x}_j. \tag{12}$$

Es gelten die Fehlerabschätzungen

$$\|\xi - x_j\| \leqq \frac{q^j}{1-q} \|x_1 - x_0\| \tag{13}$$

und

$$\|\xi - x_j\| \leqq \frac{q}{1-q} \|x_j - x_{j-1}\|. \tag{14}$$

Der Beweis entspricht vollständig dem des Satzes 1 in MfL Bd. 9, 4.1.1., wenn man dort die Beträge durch Normwerte ersetzt.

Bemerkung 1. Im Fall, daß f kontrahierende Abbildung des $\mathbf{R}^n$ in sich ist, gilt Satz 1 ohne die Bedingung (10) für den $\mathbf{R}^n$ (an Stelle von $\overline{U}_r(z)$). (10) wird nur benötigt, um die Bildbarkeit der Iterationsfolge zu sichern, die auf dem Nachweis beruht, daß $x_j \in \overline{U}_r(z)$ für $j \in \mathbf{N}$ ist.

Wir befassen uns mit Kriterien, die für die Kontraktivität von f hinreichend sind, und führen dazu den Begriff der *Lipschitzbedingung* ein:

Definition 1. Eine Funktion $f: G \to \mathbf{R}$, $G \subseteq \mathbf{R}^n$, *erfüllt* in G *eine Lipschitzbedingung* mit nichtnegativen Konstanten b_j, $j = 1(1)n$, genau dann, wenn für alle $x, y \in G$

$$|f(y) - f(x)| \leqq \sum_{j=1}^{n} b_j |y_j - x_j| \tag{15}$$

gilt.

Im folgenden Satz wird angenommen, daß die Funktionen (6) Lipschitzbedingungen in G mit den Konstanten $b_{ij} \geqq 0$ erfüllen, die wir in einer $n \times n$-Matrix $\boldsymbol{B}$ zusammenfassen. Wir sagen dann: Der Operator f erfüllt in G eine Lipschitzbedingung mit der Matrix $\boldsymbol{B}$.

Satz 2. *Der Operator $f: G \to \mathbf{R}^n$ möge in G einer Lipschitzbedingung mit der Matrix $\boldsymbol{B}$ genügen. Dann ist f kontrahierend in G, wenn $\boldsymbol{B}$ eine der Bedingungen* 6.1(75) *erfüllt. Die linken Seiten von* (75a)—(75c) *stellen Kontraktionsfaktoren für die Vektornormen* a) $\|\cdot\|_1$, b) $\|\cdot\|_\infty$ *bzw.* c) $\|\cdot\|_2$ *dar.*

Beweis. Für zwei Punkte $x_1, x_2 \in G$ mit den Koordinaten $x_j^{(1)}$ bzw. $x_j^{(2)}$, $j = 1(1)n$, ist gemäß (15)

$$|f_i(x_1) - f_i(x_2)| \leqq \sum_{j=1}^{n} b_{ij} |x_j^{(1)} - x_j^{(2)}|. \tag{15a}$$

Ist nun eine der Bedingungen 6.1.(75) erfüllt, so gilt im

Fall a)

$$\|f(x_1) - f(x_2)\|_1 = \sum_{i=1}^{n} |f_i(x_1) - f_i(x_2)|$$

$$\leqq \sum_{i,j=1}^{n} b_{ij} |x_j^{(1)} - x_j^{(2)}| = \sum_{j=1}^{n} \left(|x_j^{(1)} - x_j^{(2)}| \sum_{i=1}^{n} b_{ij} \right)$$

$$\leqq \left(\max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} b_{ij} \right) \|x_1 - x_2\|_1, \tag{16}$$

d. h., f ist bezüglich der Norm $\|\cdot\|_1$ kontrahierend mit dem Kontraktionsfaktor

$$q := \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} b_{ij}; \tag{16a}$$

Fall b)

$$\|f(x_1) - f(x_2)\|_\infty = \max_{i \in \{1,2,\ldots,n\}} |f_i(x_1) - f_i(x_2)|$$

$$\leqq \max_{i \in \{1,2,\ldots,n\}} \left(\sum_{j=1}^{n} b_{ij} |x_j^{(1)} - x_j^{(2)}| \right)$$

$$\leqq \max_{i \in \{1,2,\ldots,n\}} \sum_{j=1}^{n} b_{ij} \|x_1 - x_2\|_\infty, \tag{17}$$

d. h., f ist bezüglich der Norm $\|\cdot\|_\infty$ kontrahierend mit dem Kontraktionsfaktor

$$q := \max_{i \in \{1,2,\ldots,n\}} \sum_{j=1}^{n} b_{ij}; \tag{17a}$$

Fall c)

$$\|f(x_1) - f(x_2)\|_2^2 = \sum_{i=1}^{n} |f_i(x_1) - f_i(x_2)|^2,$$

und durch Anwendung der Schwarzschen Ungleichung auf die rechte Seite von (15a) gewinnt man

$$|f_i(x_1) - f_i(x_2)|^2 \leqq \sum_{j=1}^{n} b_{ij}^2 \|x_1 - x_2\|_2^2$$

und folglich

$$\|f(x_1) - f(x_2)\|_2^2 \leqq \sum_{i,j=1}^{n} b_{ij}^2 \|x_1 - x_2\|_2^2, \tag{18}$$

d. h., f ist bezüglich der Norm $\|\cdot\|_2$ kontrahierend mit dem Kontraktionsfaktor

$$q := \left(\sum_{i,j=1}^{n} b_{ij}^2 \right)^{1/2}. \tag{18a}$$

Bedeutet $\|\cdot\|$ eine der in Satz 2 betrachteten Normen, so gilt auf Grund von (16), (17) bzw. (18)

$$\|f(x_1) - f(x_2)\| \leqq Q \|x_1 - x_2\| \tag{19}$$

mit einer positiven Konstanten Q für beliebige $x_1, x_2 \in \mathbf{R}^n$. Da nach 6.1., Satz 11, alle Normen des $\mathbf{R}^n$ äquivalent sind, folgt (nach 6.1., Definition 4) für eine Norm $\|\cdot\|_*$ des $\mathbf{R}^n$ mit positiven Konstanten K, L

$$\|f(x_1) - f(x_2)\| \geqq \frac{1}{L} \|f(x_1) - f(x_2)\|_*$$

und

$$\|x_1 - x_2\| \leqq K \|x_1 - x_2\|_*,$$

nach (19) also

$$\|f(x_1) - f(x_2)\|_* \leqq KLQ \|x_1 - x_2\|_*.$$

(19) gilt demnach bei Erfülltsein einer Lipschitzbedingung für f mit einer beliebigen Norm des $\mathbf{R}^n$, wobei die Konstante Q von dieser abhängt. Ist $Q < 1$, so ist f Kontraktionsoperator. Wir fassen das Ergebnis in dem folgenden Satz zusammen:

Satz 3. *Genügt f einer Lipschitzbedingung in G, so gibt es für jede Norm $\|\cdot\|$ des $\mathbf{R}^n$ eine positive Konstante Q derart, daß* (19) *für alle $x_1, x_2 \in G$ gilt. f ist Kontraktionsoperator im Fall $Q < 1$.*

Der konstruktive Teil des Satzes 1 beinhaltet den Algorithmus der sukzessiven Approximation, der für den eindimensionalen Fall im PAP der Abb. 4.1 aus MfL Bd. 9 dargestellt ist. Allgemein hat man dort den Betrag durch eine Norm zu ersetzen und jede Wertzuweisung auf alle Komponenten entsprechender Vektoren auszudehnen.

Die folgende ALGOL-Prozedur *SAP2(f1,f2,x0,q,eps)* beschreibt den Algorithmus für $n = 2$ und die mit $\|\cdot\|_1$ gebildete Abbruchbedingung. Darin bedeuten die formalen Parameter:

f1,f2 linke Seiten der Gleichung (4);

x0,y eindimensionale Felder zur Speicherung der Komponenten des Startvektors bzw. der sukzessive zu berechnenden Iterationen;

q,eps reelle Variable, die dem Kontraktionsfaktor bzw. der Fehlerschranke zuzuordnen sind.

```
procedure SAP2(f1,f2,x0,y,q,eps); value q,eps;
real q,eps; array x0,y; real procedure f1,f2;
begin
      real p; array x[1:2];
      p := q/(1 - q);
      x[1] := x0[1]; x[2] := x0[2];
L:    y[1] := f1(x); y[2] := f2(x);
      if p × (abs(y[1] - x[1]) + abs(y[2] - x[2])) ≧ eps then
      begin x[1] := y[1]; x[2] := y[2]; goto L end
end
```

Wir betrachten ein Beispiel:[1])

$$\begin{aligned} f_1(x) &= \frac{1}{4}\cos x_1 - \frac{1}{5}\sin x_2 + \frac{1}{5}x_2 = x_1, \\ f_2(x) &= \frac{1}{8}\sin x_1 + \frac{1}{6}\cos x_2 - \frac{1}{4}x_1 = x_2. \end{aligned} \tag{20}$$

[1]) Ein Beispiel dieser Art wird in [11] behandelt.

Für $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^2$ ist

$$f_1(y_1, y_2) - f_1(x_1, x_2) = \frac{1}{4}(\cos y_1 - \cos x_1) - \frac{1}{5}(\sin y_2 - \sin x_2) + \frac{1}{5}(y_2 - x_2)$$

$$= \frac{1}{2}\sin\frac{x_1 + y_1}{2}\sin\frac{x_1 - y_1}{2}$$

$$+ \frac{2}{5}\cos\frac{x_2 + y_2}{2}\sin\frac{x_2 - y_2}{2} + \frac{1}{5}(y_2 - x_2)$$

und wegen $|\sin u| \leqq |u|$ für alle $u \in \mathbf{R}$

$$|f_1(y_1, y_2) - f_1(x_1, x_2)| \leqq \frac{1}{4}|y_1 - x_1| + \frac{1}{5}|y_2 - x_2| + \frac{1}{5}|y_2 - x_2|$$

$$= \frac{1}{4}|y_1 - x_1| + \frac{2}{5}|y_2 - x_2|.$$

Entsprechend findet man

$$|f_2(y_1, y_2) - f_2(x_1, x_2)| \leqq \frac{3}{8}|y_1 - x_1| + \frac{1}{6}|y_2 - x_2|.$$

f_1, f_2 genügen also Lipschitzbedingungen zur Matrix

$$\boldsymbol{B} = \begin{pmatrix} \frac{1}{4} & \frac{2}{5} \\ \frac{3}{8} & \frac{1}{6} \end{pmatrix}.$$

Auf Grund von (16a), (17a) und (18a) folgt, daß $\boldsymbol{f}$ kontrahierende Abbildung des $\mathbf{R}^2$ in sich bezüglich der Normen $\|\cdot\|_1$, $\|\cdot\|_\infty$ und $\|\cdot\|_2$ ist. Das System (20) besitzt also nach der Bemerkung 1 genau eine Lösung $\boldsymbol{\xi} \in \mathbf{R}^2$. Für $\|\cdot\|_1$ erhält man nach (16a) den Kontraktionsfaktor $q = \max\left\{\frac{5}{8}, \frac{17}{30}\right\} = 0{,}625$. Mit diesem, $eps = 10^{-6}$ und dem Nullvektor als Startelement liefert die Prozedur *SAP*2 nach 8 Schritten die Näherungslösung

$$\xi_1^* = 0{,}2431779, \qquad \xi_2^* = 0{,}2522838. \tag{21}$$

Im Beispiel ergaben sich die Lipschitzkonstanten auf Grund bekannter trigonometrischer Formeln. Meist bestimmt man diese mit Hilfe des Mittelwertsatzes der Differentialrechnung, so daß die Bemerkung 1 aus MfL Bd. 9, 4.1.1., auch auf den allgemeinen Fall zutrifft. Dementsprechend werden die Funktionen f_i, $i = 1(1)n$, in $\overline{U}_r(\boldsymbol{z})$ ($\boldsymbol{x} \in U_r(\boldsymbol{z}) \Leftrightarrow \|\boldsymbol{x} - \boldsymbol{z}\| < r$) als differenzierbar und auf $\overline{U}_r(\boldsymbol{z})$ als stetig angenommen. Dann gilt für zwei Punkte $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \overline{U}_r(\boldsymbol{z})$ mit den Koordinaten $x_j^{(1)}, x_j^{(2)}$

$$f_i(\boldsymbol{x}_1) - f_i(\boldsymbol{x}_2) = \sum_{j=1}^{n} \frac{\partial f_i(\boldsymbol{\xi}_i)}{\partial x_j}(x_j^{(1)} - x_j^{(2)}), \tag{22}$$

wobei die partiellen Ableitungen an einer gewissen Stelle ξ_i im Inneren der geradlinigen Verbindungsstrecke von x_1 und x_2 zu nehmen sind. Setzen wir noch voraus, daß diese in $U_r(x)$ beschränkt sind und etwa

$$\left|\frac{\partial f_i(x)}{\partial x_j}\right| \leqq b_{ij} \tag{23}$$

für $x \in U_r(x)$ gilt, so folgt aus (22) und (23)

$$|f_i(x_1) - f_i(x_2)| \leqq \sum_{j=1}^{n} b_{ij}\,|x_j^{(1)} - x_j^{(2)}|, \tag{24}$$

d. h., f erfüllt auf $\overline{U}_r(x)$ eine Lipschitzbedingung mit der Matrix $B = (b_{ij})$.

Die in MfL Bd. 9, 4.1.2., durchgeführten Überlegungen zum *Newtonschen Verfahren*, die auf einer Linearisierung der linken Seite der Gleichung $g(x) = 0$ beruhen, lassen sich auf (2) übertragen. Dabei setzen wir folgendes voraus:

I. Die Komponenten des Vektors (3) besitzen in einem Gebiet $G \subseteq \mathbf{R}^n$ stetige partielle Ableitungen bis zur zweiten Ordnung.

II. Für ein Element $x^{(0)} \in G$ ist die mit der Tschebyscheffnorm 5.1.(15) gebildete Umgebung $\overline{U}_r(x^{(0)}) = \{x : \|x - x^{(0)}\| \leqq r\}$ in G enthalten.

III. Die Funktionalmatrix (*Jacobische Matrix*)

$$J(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \frac{\partial g_1}{\partial x_3} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \frac{\partial g_2}{\partial x_3} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \frac{\partial g_n}{\partial x_3} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}$$

besitzt an der Stelle $x^{(0)}$ eine Inverse K_0, und bezüglich der durch $\|\cdot\|$ induzierten Matrixnorm $|\cdot|$ gilt mit einer positiven Konstanten A_0

$$|K_0| \leqq A_0.$$

IV. Für ein gewisses $B_0 > 0$ ist mit Bezug auf Voraussetzung II

$$\|K_0 g(x^{(0)})\| \leqq B_0 \leqq \frac{r}{2}.$$

V. Es sei

$$\sum_{k=1}^{n} \left|\frac{\partial^2 g_i(x)}{\partial x_j\,\partial x_k}\right| \leqq C$$

für $i, j = 1(1)n$ und $x \in \overline{U}_r(x^{(0)})$.

VI. Die Konstanten A_0, B_0 und C genügen der Bedingung

$$\mu_0 := 2nA_0B_0C \leqq 1.$$

Dem Vorgehen im eindimensionalen Fall entsprechend, ersetzen wir in (1) die g_j durch die Tangentialebenenanteile der Taylorentwicklung bei $\boldsymbol{x}^{(0)}$ (vgl. MfL Bd. 5, 3.2.2.(7)) und erhalten auf diese Weise das lineare Gleichungssystem

$$g_j(\boldsymbol{x}^{(0)}) + \sum_{i=1}^{n} \frac{\partial g_j(\boldsymbol{x}^{(0)})}{\partial x_i} (x_i - x_i^{(0)}) = 0. \tag{25}$$

Die auf Grund von Voraussetzung III wohlbestimmte Lösung sei mit $\boldsymbol{x}^{(1)}$ bezeichnet:

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \boldsymbol{K}_0\boldsymbol{g}(\boldsymbol{x}^{(0)}). \tag{26}$$

$\boldsymbol{x}^{(1)}$ ist das mit dem Operator

$$\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{x} - \boldsymbol{J}(\boldsymbol{x})^{-1}\,\boldsymbol{g}(\boldsymbol{x}) \tag{27}$$

nach der Methode der sukzessiven Approximation aus $\boldsymbol{x}^{(0)}$ erzeugte Element. Der folgende Satz bringt zum Ausdruck, daß man die so begonnene Iterationsfolge fortsetzen kann und diese gegen eine Lösung $\boldsymbol{\xi} \in \overline{U}_r(\boldsymbol{x}^{(0)})$ von (2) konvergiert. Für $n = 1$ ($g_1 = g$) ist $\boldsymbol{J}(\boldsymbol{x})^{-1} = \left(\frac{1}{g'(x)}\right)$, und (27) ist der in MfL Bd. 9, 4.1.(13), eingeführte Operator des Newtonschen Verfahrens.

Satz 4. *Unter den Voraussetzungen* I—VI *ist die Iterationsfolge des Newtonschen Verfahrens*

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \boldsymbol{K}(\boldsymbol{x}^{(p)})\,\boldsymbol{g}(\boldsymbol{x}^{(p)}), \qquad p = 0, 1, \ldots, \tag{28}$$

bildbar und konvergiert gegen eine Lösung $\boldsymbol{\xi} \in \overline{U}_r(\boldsymbol{x}^{(0)})$ *von* (2). *Dafür gilt die Abschätzung*

$$\|\boldsymbol{\xi} - \boldsymbol{x}^{(p)}\| \leqq \frac{1}{2^{p-1}}\,\mu_0^{2^p-1}B_0. \tag{29}$$

$\boldsymbol{K}(\boldsymbol{x})$ *bedeutet die Inverse der Funktionalmatrix* $\boldsymbol{J}(\boldsymbol{x}) = \left(\frac{\partial g_i(\boldsymbol{x})}{\partial x_j}\right)$.

Beweis. Der Darstellung in [9] folgend zeigen wir, daß die Voraussetzungen I—VI auch für $\boldsymbol{x}^{(1)}$ mit geeignet gewählten Konstanten A_1, B_1 und bezüglich der Umgebung $\overline{U}_{r/2}(\boldsymbol{x}^{(1)})$ erfüllt sind. Zunächst ergibt sich aus (26) mit Voraussetzung IV

$$\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| = \|\boldsymbol{K}_0\boldsymbol{g}(\boldsymbol{x}^{(0)})\| \leqq B_0 \leqq \frac{r}{2}, \tag{30}$$

also

$$\boldsymbol{x}^{(1)} \in \overline{U}_{r/2}(\boldsymbol{x}^{(0)})$$

und daher

$$\overline{U}_{r/2}(\boldsymbol{x}^{(1)}) \subseteq \overline{U}_r(\boldsymbol{x}^{(0)}) \subseteq G.$$

Nun wird die Invertierbarkeit von $\boldsymbol{J}(\boldsymbol{x}^{(1)})$ gezeigt. Auf Grund von 6.1.(57b) erhält man nach Anwendung des Mittelwertssatzes der Differentialrechnung unter Beachtung von Voraussetzung V und (30)

$$\begin{aligned}
\|\boldsymbol{J}(\boldsymbol{x}^{(1)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})\| &= \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} \left| \frac{\partial g_i(\boldsymbol{x}^{(1)})}{\partial x_j} - \frac{\partial g_i(\boldsymbol{x}^{(0)})}{\partial x_j} \right| \\
&= \max_{i\in\{1,2,\ldots,n\}} \sum_{j=1}^{n} \left| \sum_{k=1}^{n} \frac{\partial^2 g_i(\eta_{ij})}{\partial x_j\, \partial x_k} \left(x_k^{(1)} - x_k^{(0)}\right) \right| \\
&\leqq \max_{i\in\{1,2,\ldots,n\}} \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^2 g_i(\eta_{ij})}{\partial x_j\, \partial x_k} \right| \\
&\leqq nC\, \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \leqq nCB_0; \qquad (31)
\end{aligned}$$

η_{ij} bedeutet einen inneren Punkt der Verbindungsstrecke von $\boldsymbol{x}^{(0)}$ und $\boldsymbol{x}^{(1)}$. Mit (31) und den Voraussetzungen III und VI folgt ($\boldsymbol{I}$ Einheitsmatrix)

$$\begin{aligned}
\|\boldsymbol{I} - \boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)})\| &= \|\boldsymbol{K}_0(\boldsymbol{J}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(1)}))\| \\
&\leqq \|\boldsymbol{K}_0\| \cdot \|\boldsymbol{J}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(1)})\| \leqq nA_0B_0C = \frac{\mu_0}{2} \leqq \frac{1}{2}.
\end{aligned}$$

Diese Abschätzung hat auf Grund von 6.1., Satz 16, zur Konsequenz, daß die Matrix

$$\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}) = \boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))$$

und somit auch $\boldsymbol{J}(\boldsymbol{x}^{(1)})$ regulär ist. (Bekanntlich folgt aus der eindeutigen Lösbarkeit eines linearen Gleichungssystems 6.1.(2) ($m = n$) für beliebige rechte Seite das Nichtverschwinden der Koeffizientendeterminante.) Wir setzen zur Abkürzung

$$\boldsymbol{B} := \boldsymbol{I} - \boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}) \qquad \left(\|\boldsymbol{B}\| \leqq \frac{1}{2}\right)$$

und erhalten aus $(\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))^{-1} = (\boldsymbol{I} - \boldsymbol{B})^{-1} = \boldsymbol{I} + \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{B})^{-1}$

$$\|(\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))^{-1}\| \leqq \|\boldsymbol{I}\| + \|\boldsymbol{B}\| \cdot \|(\boldsymbol{I} - \boldsymbol{B})^{-1}\| = 1 + \|\boldsymbol{B}\| \cdot \|(\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))^{-1}\|,$$

also

$$\|(\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))^{-1}\| \leqq \frac{1}{1 - \|\boldsymbol{B}\|} \leqq 2. \qquad (32)$$

Für

$$\boldsymbol{K}_1 := \boldsymbol{K}(\boldsymbol{x}^{(1)}) = \boldsymbol{J}(\boldsymbol{x}^{(1)})^{-1}$$

gewinnt man über die Darstellung

$$\boldsymbol{K}_1 = [\boldsymbol{J}(\boldsymbol{x}^{(0)})\, \boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)})]^{-1} = (\boldsymbol{K}_0\boldsymbol{J}(\boldsymbol{x}^{(1)}))^{-1}\, \boldsymbol{K}_0$$

mit Hilfe von (32) die Normabschätzung

$$\|\boldsymbol{K}_1\| \leqq 2\, \|\boldsymbol{K}_0\| \leqq 2A_0. \qquad (33)$$

Im Hinblick auf Voraussetzung III wird daher

$$A_1 := 2A_0 \qquad (33a)$$

gesetzt.

Wir wenden uns der Normabschätzung von $\boldsymbol{K}_1\boldsymbol{g}(\boldsymbol{x}^{(1)})$ zu. Aus (26) folgt

$$\boldsymbol{g}(\boldsymbol{x}^{(0)}) + \boldsymbol{J}(\boldsymbol{x}^{(0)})\,(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) = \boldsymbol{0},$$

so daß

$$\boldsymbol{g}(\boldsymbol{x}^{(1)}) = \boldsymbol{g}(\boldsymbol{x}^{(1)}) - \boldsymbol{g}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})\,(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)})$$

gilt. Die Komponenten der rechten Seite können durch die mit den zweiten Ableitungen gebildeten Restglieder der Taylorentwicklung von g_i bei $\boldsymbol{x}^{(0)}$ zum Inkrementvektor $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}$ ausgedrückt werden. Man hat also

$$\boldsymbol{g}(\boldsymbol{x}^{(1)}) = \frac{1}{2}\begin{pmatrix} \left[\sum_{j=1}^{n} (x_j^{(1)} - x_j^{(0)})\,\frac{\partial}{\partial x_j}\right]^2 g_1(\eta_1) \\ \left[\sum_{j=1}^{n} (x_j^{(1)} - x_j^{(0)})\,\frac{\partial}{\partial x_j}\right]^2 g_2(\eta_2) \\ \vdots \\ \left[\sum_{j=1}^{n} (x_j^{(1)} - x_j^{(0)})\,\frac{\partial}{\partial x_j}\right]^2 g_n(\eta_n) \end{pmatrix}; \tag{34}$$

bei der polynomischen Entwicklung der Klammern sind Produkte von Differentiationsoperatoren als entsprechende partielle Ableitungen zweiter Ordnung zu deuten; η_i, $i = 1(1)n$, bezeichnet einen inneren Punkt der Verbindungsstrecke von $\boldsymbol{x}^{(0)}$ und $\boldsymbol{x}^{(1)}$. Nun ist aber

$$\left[\sum_{j=1}^{n} (x_j^{(1)} - x_j^{(0)})\,\frac{\partial}{\partial x_j}\right]^2 g_i(\eta_i) = \sum_{j=1}^{n}\left[(x_j^{(1)} - x_j^{(0)})\sum_{k=1}^{n}(x_k^{(1)} - x_k^{(0)})\right]\frac{\partial^2 g_i(\eta_i)}{\partial x_j\,\partial x_k}$$

und folglich unter Beachtung von Voraussetzung V

$$\left|\left[\sum_{j=1}^{n} (x_j^{(1)} - x_j^{(0)})\,\frac{\partial}{\partial x_j}\right]^2 g_i(\eta_i)\right| \leqq n\,\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|^2\,C,$$

also wegen (30)

$$\|\boldsymbol{g}(\boldsymbol{x}^{(1)})\| \leqq \frac{1}{2}\,nC\,\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|^2 \leqq \frac{1}{2}\,nCB_0^2. \tag{35}$$

Unter Berücksichtigung von (33) ergibt sich damit

$$\|\boldsymbol{K}_1\boldsymbol{g}(\boldsymbol{x}^{(1)})\| \leqq |\boldsymbol{K}_1| \cdot \|\boldsymbol{g}(\boldsymbol{x}^{(1)})\| \leqq nA_0B_0^2C = \frac{1}{2}\,\mu_0 B_0 \tag{36}$$

und — wenn

$$B_1 := \frac{1}{2}\,\mu_0 B_0 \tag{36a}$$

gesetzt wird —

$$\mu_1 := 2nA_1B_1C = 2nA_0B_0C\mu_0 = \mu_0^2 \leqq 1.$$

Zusammengefaßt gilt:

Wählt man A_1, B_1 gemäß (33a) bzw. (36a), so sind die Bedingungen I—VI für $\boldsymbol{x}^{(1)}$ mit diesen Konstanten an Stelle von A_0, B_0 bezüglich der Umgebung $\overline{U}_{r/2}(\boldsymbol{x}^{(1)})$ erfüllt.

Wendet man diese Überlegung wiederholt an, so ergibt sich die Existenz der Punktfolge (28) mit der Einschließungseigenschaft

$$G \supseteq \overline{U}_r(\boldsymbol{x}^{(0)}) \supseteq \overline{U}_{r/2}(\boldsymbol{x}^{(1)}) \supseteq \cdots \supseteq \overline{U}_{r/2^p}(\boldsymbol{x}^{(p)}) \supseteq \cdots. \tag{37}$$

Für jedes p gewinnt man Konstanten A_p, B_p, μ_p, die gemäß (33a) und (36a) den Rekursionsgleichungen

$$A_p = 2A_{p-1}, \qquad B_p = \frac{1}{2}\,\mu_{p-1}B_{p-1} \tag{38}$$

genügen; dabei ist

$$\mu_p := 2nA_pB_pC, \qquad p = 1, 2, \ldots . \tag{39}$$

Da nach (37) für beliebige $p, q \in \mathbb{N}$

$$\|x^{(p+q)} - x^{(p)}\| \leqq \frac{r}{2^p} \tag{40}$$

gilt, ist $(x^{(p)})$ eine Fundamentalfolge im $\mathbb{R}^n$ und konvergiert wegen der Vollständigkeit dieses Raumes gegen ein Element $\xi \in \overline{U}_r(x^{(0)})$:

$$\lim_{p\to\infty} x^{(p)} = \xi. \tag{41}$$

ξ ist Lösung der Gleichung (2): Aus (28) folgt

$$g(x^{(p)}) = J(x^{(p)})\,(x^{(p)}) - x^{(p+1)} \tag{42}$$

und wegen der Stetigkeit der Komponenten von g

$$\lim_{p\to\infty} g(x^{(p)}) = g(\xi).$$

Bezüglich der rechten Seiten von (42) ist zu beachten, daß die Komponenten der Funktionalmatrix auf der abgeschlossenen Menge $\overline{U}_r(x^{(0)})$ ebenfalls stetig und daher beschränkt sind. Daraus folgt die Beschränktheit von $|J(x)|$ auf $\overline{U}_r(x^{(0)})$, und man gewinnt aus (42) mit (40) für $p \to \infty$

$$\|g(\xi)\| = 0,$$

also

$$g(\xi) = 0.$$

Auf Grund von (38) und (39) ist $\mu_p = \mu_0^{2^p}$ und

$$B_p = \frac{1}{2^p}\,\mu_{p-1}\mu_{p-2}\cdots\mu_0 B_0 = \frac{1}{2^p}\,\mu_0^{2^p-1}B_0.$$

Damit erhält man für $q > p$ unter Beachtung von $0 \leqq \mu_0 \leqq 1$

$$\begin{aligned}\|x^{(p)} - x^{(q)}\| &\leqq \|x^{(p)} - x^{(p+1)}\| + \|x^{(p+1)} - x^{(p+2)}\| + \cdots + \|x^{(q-1)} + x^{(q)}\| \\ &\leqq B_p + B_{p+1} + \cdots + B_{q-1} \\ &\leqq \frac{1}{2^p}\,\mu_0^{2^p-1}B_0\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{q-p-1}}\right) \\ &< \frac{1}{2^{p-1}}\,\mu_0^{2^p-1}B_0\end{aligned}$$

und für $q \to \infty$

$$\|\xi - x^{(p)}\| \leqq \frac{1}{2^{p-1}}\,\mu_0^{2^p-1}B_0.$$

Bemerkung 2. Unter den Voraussetzungen I—VI besitzt die Gleichung (2) auf der abgeschlossenen Umgebung

$$\|x - x^{(0)}\| \leqq 2B_0 \tag{43}$$

genau eine Lösung. Ferner gilt für die Konvergenzgeschwindigkeit des Newtonschen Verfahrens eine dem eindimensionalen Fall entsprechende Aussage. Man zeigt leicht, daß die mit $d_j := \|x^{(j)} - \xi\|$ gebildete Größe $\frac{d_{j+1}}{d_j^2}$ für $j \to \infty$ beschränkt bleibt.

Bemerkung 3. Auch für $\|\cdot\|_1$, $\|\cdot\|_2$ und die davon induzierten Matrixnormen (vgl. 6.1.(57), (60)) läßt sich ein Konvergenzsatz für das Newtonsche Verfahren formulieren. Dieses wurde von L. W. KANTOROWITSCH auf die Lösung von Operatorgleichungen in Banachräumen ausgedehnt (vgl. [24], Kap. XVIII; [27], Kap. III). Grundlage für diese wichtige funktionalanalytische Methode ist eine Verallgemeinerung des Ableitungsbegriffs.

Bemerkung 4. Der Rechenaufwand bei der Bildung der Iterationsfolge (28) rührt wesentlich von der Invertierung der Jacobischen Matrix an den Stellen $x^{(p)}$ her. Man kann ein *modifiziertes Newtonsches Verfahren* betrachten, wo in jedem Schritt in (28) mit derselben Matrix $K_0 = K(x^{(0)})$ multipliziert wird. Dieses konvergiert, wenn die Voraussetzungen I—VI mit der Verschärfung

$$\mu_0 := 2nA_0B_0C < 1$$

erfüllt sind.

Beispiel. Gesucht sind Lösungen des Gleichungssystems

$$g_1(x_1, x_2) = x_1^4 + 3x_1^2x_2 + x_2^2 - x_1^2 - 2x_1 - x_2 - 2 = 0,$$

$$g_2(x_1, x_2) = x_2x_1^3 + 2x_1x_2^2 - x_1x_2 - 2x_2 - 1 = 0.$$

Die mit einer BESM6 berechnete Tabelle 6.6 enthält für $p = 0(1)10$ die Komponenten der mit dem Nullvektor gestarteten Iterationsfolge (28) und die Werte von $|K(x^{(p)})|$ sowie $\|K(x^{(p)})\, g(x^{(p)})\|$. Man erkennt, daß die Folge der $x^{(p)}$ auf der EDVA konvergiert, da sie für $p \geqq 9$ stationär ist. Auf Grund des Satzes 4 ergibt sich die Konvergenz, indem man etwa für $x^{(5)}$ die Umgebung $\overline{U}_{0,2}$ betrachtet, wo Voraus-

p	$x_1^{(p)}$	$x_2^{(p)}$	$\|K(x^{(p)})\|$	$\|K(x^{(p)})\, g(x^{(p)})\|$
0	0.00000000000 +00	0.00000000000 +00	7.50000000000 −01	7.50000000000 −01
1	−7.50000000000 −01	−5.00000000000 −01	1.27179487179 +01	2.56089743589 +00
2	−1.16987179485 −01	−3.06089743589 +00	1.43838304528 −01	1.48172606701 +00
3	−2.04944765880 −01	−1.57917136888 +00	2.62001535508 −01	5.49655529604 −01
4	−2.90735405978 −01	−1.02951583928 +00	4.20215562949 −01	1.14809442388 −01
5	−3.27819774629 −01	−9.14706396893 −01	5.39392286721 −01	5.91156473677 −03
6	−3.32272097464 −01	−9.08794832156 −01	5.46866923727 −01	4.72098233555 −05
7	−3.32319307288 −01	−9.08790735955 −01	5.46868264754 −01	2.83413528483 −09
8	−3.32319303168 −01	−0.08790738790 −01	5.46868260767 −01	1.50715467993 −12
9	−3.32319309168 −01	−9.08790738788 −01	5.46868260768 −01	0.00000000000 +00
10	−3.32319309168 −01	−9.08790738788 −01		

Tabelle 6.6

setzung V mit $C := 12$ erfüllt ist. Nach den Angaben in den letzten beiden Spalten der Tabelle 6.6 gelten die Voraussetzungen III und IV mit

$$A_0 = 5{,}4 \cdot 10^{-1} \qquad \text{bzw.} \qquad B_0 = 6{,}0 \cdot 10^{-3},$$

so daß

$$\mu_0 = 2nA_0B_0C < 0{,}2$$

ist. Nach Bemerkung 2 zu Satz 4 besitzt das Gleichungssystem auf $\overline{U}_{0,012}(x^{(5)})$ genau eine Lösung, und zwar die mit der betrachteten Iterationsfolge näherungsweise bestimmte. In 6.2.2. werden wir sehen, daß außerhalb dieser Umgebung noch eine weitere Lösung existiert.

6.2.2. Lösung von Polynomgleichungen

Die Bestimmung der Nullstellen eines Polynoms

$$P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0, \qquad a_n \neq 0, \tag{45}$$

ist eine Standardaufgabe der Numerischen Mathematik, nicht zuletzt, weil Polynome häufig zur Approximation anderer Funktionen benutzt werden. Außerdem ergibt sie sich im Zusammenhang mit zahlreichen physikalischen und technischen Fragestellungen. Auch für diesen Spezialfall der in MfL Bd. 9, 4.1., behandelten Gleichungsprobleme ist die Lokalisierung der Nullstellen in gewissen Einschließungsintervallen wesentliche Voraussetzung für die Anwendung schrittweise vorgehender Verfahren zu ihrer beliebig genauen Bestimmung. Wir beginnen daher mit der Erörterung einiger Lokalisierungssätze. Grundlage für die meisten Betrachtungen ist der Fundamentalsatz der Algebra, nach dem ein Polynom (45) mit komplexen Koeffizienten $a_0, a_1, \ldots, a_n$ in der Gaußschen Zahlenebene genau n Nullstellen besitzt, vorausgesetzt, daß diese entsprechend ihrer Vielfachheit gezählt werden.

Eine mit einem Polynom gebildete Gleichung $P(x) = 0$ wird *algebraisch* genannt.

Satz 5. *Bezüglich der Koeffizienten in* (45) *sei*

$$A := \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}. \tag{46}$$

Dann sind die Beträge sämtlicher Nullstellen dieses Polynoms kleiner als

$$K := 1 + \frac{A}{|a_n|}. \tag{47}$$

Beweis. Für $|x| > 1$ folgt aus (45)

$$\begin{aligned} |P(x)| &\geqq |a_nx^n| - |a_{n-1}x^{n-1} + \cdots + a_1x + a_0| \\ &\geqq |a_n| \cdot |x|^n - A(|x|^{n-1} + |x|^{n-2} + \cdots + |x| + 1) \\ &= |a_n| \cdot |x|^n - A\,\frac{|x|^n - 1}{|x| - 1} \\ &> \left(|a_n| - \frac{A}{|x| - 1}\right)|x|^n. \end{aligned}$$

$|P(x)|$ ist also positiv, wenn

$$|a_n| - \frac{A}{|x|-1} \geqq 0, \quad \text{d. h.} \quad |x| \geqq 1 + \frac{A}{|a_n|},$$

und die durch (47) definierte Größe K gewiß eine obere Schranke[1]) für die Beträge sämtlicher Nullstellen.

Weiterhin betrachten wir nur Polynome (45) mit reellen Koeffizienten. Für ein eingehendes Studium der Verteilung der Nullstellen solcher Polynome sei der Leser auf die Monographie [34] von N. OBRESCHKOFF verwiesen.

Satz 6 (*Regel von Lagrange und Maclaurin*). *In dem normierten Polynom* (45) ($a_n = 1$) *sei* α *der absolute Betrag des betragsgrößten negativen Koeffizienten und* m *die Differenz zwischen dem Grad und dem Exponenten des höchsten Gliedes mit einem negativen Koeffizienten. Dann ist*

$$L := 1 + \sqrt[m]{\alpha} \tag{48}$$

eine obere Schranke für die reellen Nullstellen von (45).

Beweis. Wir zeigen, daß $P(x) > 0$ für $x \geqq L$. Offenbar ist für positive x

$$P(x) \geqq x^n - \alpha(x^{n-m} + x^{n-m-1} + \cdots + 1) = x^n - \alpha \frac{x^{n-m+1} - 1}{x - 1},$$

also $P(x) > 0$, sofern

$$x^n > \alpha \frac{x^{n-m+1} - 1}{x - 1}$$

gilt. Wird sogar $x > 1$ angenommen, so ist dafür hinreichend, daß

$$x^n \geqq \alpha \frac{x^{n-m+1}}{x - 1} \qquad \text{oder} \qquad x^{m-1}(x-1) \geqq \alpha$$

oder a fortiori

$$(x-1)^m \geqq \alpha$$

ist. Also gilt $P(x) > 0$ für $x \geqq 1 + \sqrt[m]{\alpha} = L$ und $\xi < L$ für jede Nullstelle ξ von P.

Bemerkung 5. Wenn in dem normierten Polynom (45) keine negativen Koeffizienten auftreten, hat dieses keine positiven Nullstellen, und $L = 0$ ist eine obere Schranke für die reellen Wurzeln von $P(x) = 0$.

Satz 7 (*Regel von Newton*). *Wenn für* $L \in \mathbf{R}$ *sämtliche Ableitungen* $P^{(k)}(L)$, $k = 0(1)n$, *des Polynoms* (45) *positiv sind, ist* L *eine obere Schranke für die reellen Wurzeln von* $P(x) = 0$.

[1]) Auch im folgenden wird der Begriff „Schranke" bei der Einschließung von Polynomnullstellen stets in Verbindung mit dem echten Kleiner- bzw. Größerzeichen benutzt.

Beweis. Durch Umordnung des Polynoms P nach Potenzen von $x - L$ gewinnt man

$$P(x) = \frac{P^{(n)}(L)}{n!}(x-L)^n + \frac{P^{(n-1)}(L)}{(n-1)!}(x-L)^{n-1} + \cdots + \frac{P'(L)}{1!}(x-L) + P(L)$$

und auf Grund der Voraussetzung

$$P(x) > 0 \quad \text{für} \quad x \geqq L.$$

Es ist naheliegend, die Regel von NEWTON in Verbindung mit dem erweiterten Hornerschen Schema (MfL Bd. 9, 4.1.2.) anzuwenden. Schneller kommt man nach der folgenden *Regel von Laguerre* ([54, 29]) zum Ziel.

Satz 8. *Nach Division des Polynoms P durch $x - L$ $(L > 0)$ mit Rest,*

$$P(x) = (x - L)\,Q_{n-1}(x) + r, \qquad r \in \mathbf{R}, \tag{49}$$

mögen r positiv und die Koeffizienten von Q_{n-1} nicht negativ sein. Dann ist L eine obere Schranke für die reellen Wurzeln von $P(x) = 0$.

Beweis. Offenbar ist nach (49) unter den Voraussetzungen des Satzes $P(x) > 0$ für $x \geqq L$. Die Koeffizienten von Q_{n-1} und r gewinnt man nach MfL Bd. 9, 4.1. (15)–(17), indem man mit dem Polynom P einen Hornerschritt für $x_0 = L$ rechnet.

Bemerkung 6. Ist ξ eine Nullstelle von $P(-x)$, so ist $-\xi$ eine solche von $P(x)$ und umgekehrt. Sind die reellen Nullstellen von $P(-x)$ also kleiner als $L \in \mathbf{R}$, so ist $-L$ eine untere Schranke für die reellen Wurzeln von $P(x) = 0$ und kann demnach durch Anwendung der Sätze 6, 7 und 8 auf $P(-x)$ oder $(-1)^n P(-x)$ bestimmt werden. Auch positive untere und negative obere Schranken für die positiven bzw. negativen Nullstellen eines Polynoms lassen sich damit in Verbindung mit der linearen Transformation $x' = \frac{1}{x}$ finden. Zum Beispiel ist ξ eine positive Nullstelle des Polynoms (45) genau dann, wenn $\xi' = \frac{1}{\xi} > 0$ der Gleichung $P\left(\frac{1}{x}\right) = 0$, also auch $x^n P\left(\frac{1}{x}\right) = 0$, d. h.

$$Q(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n = 0 \tag{50}$$

genügt. Ist daher L eine obere Schranke für die positiven Nullstellen von Q, so folgt aus $\xi' < L$

$$\xi > \frac{1}{L},$$

d. h., $l := \frac{1}{L}$ ist positive untere Schranke für die positiven Nullstellen von $P(x)$. Entsprechend begründet man: Ist L eine obere Schranke für die positiven Nullstellen von $Q(-x)$, so ist $l := -\frac{1}{L}$ eine negative obere Schranke für die negativen Nullstellen von $P(x)$.

Wir betrachten ein

Beispiel. Es sind die Wurzeln der algebraischen Gleichung

$$P(x) = x^5 - x^3 - 2x^2 - 2x - 1 = 0 \tag{51}$$

zu lokalisieren. Gemäß (46) ist $A = 2$ und nach Satz 5 für jede Nullstelle x von P

$$|x| < 3. \tag{52}$$

Nach Satz 6, (48) findet man

$$L = 1 + \sqrt{2} < 2{,}4142 \tag{53}$$

als eine obere Schranke für die reellen Wurzeln von (51).

Um Satz 8 anzuwenden, rechnen wir mit dem Polynom in (51) einen Hornerschritt für $x_0 = 2$:

	1	0	−1	−2	−2	−1
		2	4	6	8	12
2	1	2	3	4	6	11

Da unter dem Strich nur positive Werte erscheinen, kann L durch den kleineren Wert $L_1 = 2$ ersetzt werden.

Für die mit $P(-x)$ gebildete Gleichung

$$x^5 - x^3 + 2x^2 - 2x + 1 = 0 \tag{54}$$

findet man nach Satz 6 wieder (53) als eine obere Schranke für die reellen Wurzeln von (54). Diese läßt sich nach der Regel von Laguerre zu $L_2 = 1$ verschärfen. In der Tat ist

	1	0	−1	2	−2	1
		1	1	0	2	0
1	1	1	0	2	0	1

Auf Grund der Bemerkung 6 liegen damit sämtliche reellen Wurzeln von (51) im Intervall

$$-L_2 = -1 < x < L_1 = 2. \tag{55}$$

Die (50) entsprechende Gleichung lautet

$$-Q(x) = x^5 + 2x^4 + 2x^3 + x^2 - 1 = 0. \tag{56}$$

Nach der Regel von Laguerre findet man $L_3 = 1$ als eine obere Schranke für die positiven Wurzeln von (56), d. h., die positiven Wurzeln von (51) liegen nach Bemerkung 6 im Intervall

$$\frac{1}{L_3} = 1 < x < L_1 = 2. \tag{57}$$

Für die reellen Wurzeln der Gleichung

$$Q(-x) = x^5 - 2x^4 + 2x^3 - x^2 + 1 = 0 \tag{58}$$

liefert Satz 8 die obere Schranke $L_4 = 2$, so daß nach Bemerkung 6 sämtliche negativen Wurzeln der Gleichung (51) dem Intervall

$$-L_2 = -1 < x < -\frac{1}{L_4} = -\frac{1}{2} \tag{59}$$

angehören.

Wir wenden uns nun Methoden zu, die es gestatten, die Anzahl der reellen Wurzeln eines Polynoms P in einem Intervall zu bestimmen. Mit dem folgenden Satz läßt sich entscheiden, ob diese gerade oder ungerade ist; der Beweis ergibt sich durch Anwendung des Satzes 2 aus MfL Bd. 4, 2.4., auf P.

Satz 9. *Für ein Intervall* $[\![a, b]\!]$ *sei* a) $P(a)\,P(b) < 0$ *bzw.* b) $P(a)\,P(b) > 0$. *Dann besitzt* P *in* $]\!]a, b[\!]$ *im Fall* a) *eine ungerade, im Fall* b) *keine oder eine gerade Anzahl von Nullstellen. Dabei sind mehrfache Nullstellen entsprechend oft zu zählen.*

In dem zuvor betrachteten Beispiel liegt für das Intervall (57) der Fall a), für (59) der Fall b) vor.

Die Bestimmung der Anzahl reeller Nullstellen eines Polynoms in einem Intervall beruht auf der Feststellung von Zeichenwechseln über gewissen endlichen Zahlenfolgen. Wir präzisieren diesen Begriff nach [9] durch die folgende

Definition 2. Es sei

$$c_1, c_2, \ldots, c_n \qquad (n \geqq 2) \tag{60}$$

eine endliche Folge von Null verschiedener reeller Zahlen. Für das Paar c_k, c_{k+1} liegt kein Zeichenwechsel vor, wenn $c_k c_{k+1} > 0$ ist.
ein $c_k c_{k+1} < 0$
Die Anzahl sämtlicher Zeichenwechsel bei aufeinanderfolgenden Elementen c_k, c_{k+1}, $k = 1(1)n - 1$, heißt *Anzahl der Zeichenwechsel in* (60) und wird mit W bezeichnet. Werden in (60) auch verschwindende Elemente c_k, $k = 2(1)n - 1$, $c_1 \neq 0$, $c_n \neq 0$, zugelassen, so sei $\underline{W}$ die Anzahl der Zeichenwechsel, die in (60) nach Weglassen derselben auftreten, $\overline{W}$ diejenige, welche sich für eine nach folgender Vorschrift transformierte Folge (60) ergibt: Man betrachte sämtliche Null-Sequenzen

$$c_k = c_{k+1} = \cdots = c_{k+l-1} = 0,$$

wobei $c_{k-1} \neq 0$ und $c_{k+l} \neq 0$ ist, und ersetze darin c_{k+i} durch eine reelle Zahl $\tilde{c}_{k+i}$, $i = 0(1)l - 1$, mit

$$\operatorname{sgn} \tilde{c}_{k+i} = (-1)^{l-i} \operatorname{sgn} c_{k+l}. \tag{61}$$

Für eine Folge nichtverschwindender Zahlen (60) sei definitionsgemäß

$$W = \underline{W} = \overline{W}.$$

Beispielsweise ist für die Folge

$$1, 0, 0, -2, 0, 1$$

$\underline{W} = 2$. Zur Bestimmung von $\overline{W}$ wird mit einem positiven ε die transformierte Folge

$$1, -\varepsilon, \varepsilon, -2, -\varepsilon, 1$$

gebildet, wonach sich $\overline{W} = 4$ ergibt. Man beachte, daß gemäß (61) am Ende einer Null-Sequenz gegenüber dem folgenden Element stets ein Vorzeichenwechsel auftritt.

Satz 10 (*Regel von Budan-Fourier*). *Es sei*

$$P(x) = a_n x^{n-1} + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

ein Polynom vom Grade n und $a, b \in \mathbf{R}$, $a < b$, $P(a) \neq 0$, $P(b) \neq 0$. $W(x)$, $\underline{W}(x)$, $\overline{W}(x)$ *bedeuten die in Definition 2 eingeführten Größen für die Folge der Ableitungen*

$$P(x), P'(x), \ldots, P^{(n-1)}(x), P^{(n)}(x). \tag{63}$$

Dann gilt für die Anzahl $N(a, b)$ *der Nullstellen von P im Intervall* $[\![a, b]\!]$ *mit Berücksichtigung ihrer Vielfachheit*

$$N(a, b) = \underline{W}(a) - \overline{W}(b) - g, \tag{64}$$

wobei g eine nichtnegative gerade Zahl bedeutet. Man nennt $\underline{W}(a) - \overline{W}(b)$ *die Anzahl der verlorenen Zeichenwechsel in der Folge* (63) *beim Durchlaufen des Intervalls* $[\![a, b]\!]$.

Beweis. Man gewinnt (64) aus der Einsicht, daß die Größe $W(x)$ eine monoton fallende Treppenfunktion mit einem charakteristischen Sprungverhalten an den Unstetigkeitsstellen ist. Um das zu zeigen, beweisen wir zunächst den folgenden

Hilfssatz 1. *Für ein beliebiges Argument* $\xi \in \mathbf{R}$ *gilt*

$$\lim_{x \to \xi - 0} W(x) = \overline{W}(\xi) \geqq \underline{W}(\xi) = \lim_{x \to \xi + 0} W(x) \tag{65a}$$

und mit einer nichtnegativen geraden Zahl g

$$\overline{W}(\xi) - \underline{W}(\xi) = \lim_{x \to \xi - 0} W(x) - \lim_{x \to \xi + 0} W(x) = g + m, \tag{65b}$$

wenn ξ *eine m-fache Nullstelle von P ist* ($m = 0$ *für* $P(\xi) \neq 0$).

Beweis. Wir beginnen mit dem *Fall* $P(\xi) \neq 0$ und stellen in Tabelle 6.7 schematisch die Vorzeichenverhältnisse in der Folge (63) für $x = \xi$ dar. Das Zeichen $\circ$ bzw. $\blacktriangle$ wird gesetzt, je nachdem, ob die entsprechende Größe verschwindet oder nicht. Am Anfang und am Ende der Zeile für $P^{(j)}(\xi)$ steht $\blacktriangle$, da $a_n = \dfrac{P^{(n)}(\xi)}{n!}$ und $P(\xi)$ nach Voraussetzung von Null verschieden sind. Offensichtlich gilt (65a) mit dem Gleichheitszeichen an Stelle von $\geqq$ und (65b) mit $g = m = 0$, wenn $P^{(j)}(\xi) \neq 0$ für $j = 0(1)n$ ist. Dann ist nämlich für alle x einer gewissen Umgebung von ξ

$$W(x) = W(\xi) = \overline{W}(\xi) = \underline{W}(\xi).$$

j	0	1	2	...	$(r-k)$	$(r-k+1)$	...	$(r-1)$	r	$(r+1)$	...	$(n-1)$	n
$P^{(j)}(\xi - \lvert h\rvert)$				...	▲	▲		... ▲	▲				
$P^{(j)}(\xi)$	▲	...	...	...	▲	○		... ○	○	▲	...	...	▲
$P^{(j)}(\xi + \lvert h\rvert)$				...	▲	▲		... ▲	▲				

Tabelle 6.7

Weiterhin wird daher angenommen, daß in (63) für $x = \xi$ mindestens eine Nullsequenz

$$P^{(r-k+1)}(\xi) = P^{(r-k+2)}(\xi) = \cdots = P^{(r)}(\xi) = 0,$$
$$P^{(r-k)}(\xi) \neq 0, \qquad P^{(r+1)}(\xi) \neq 0 \qquad (r, k \in \mathbf{N}^*) \tag{66}$$

auftritt.[1]) Wir diskutieren mit einem Inkrement h in einer hinreichend kleinen Umgebung

$$I :=]\!]\xi - \varepsilon, \xi + \varepsilon[\![\qquad (\varepsilon > 0)$$

von ξ die Taylorentwicklungen der Funktionen $P^{(j)}$ bei ξ. Die Größe ε sei so klein gewählt, daß für

$$P^{(j)}(\xi + h) = P^{(j)}(\xi) + hP^{(j+1)}(\xi) + \frac{h^2}{2!} P^{(j+2)}(\xi) + \cdots, \qquad j = 0(1)n,$$

das erste Glied der rechten Seite mit einer von Null verschiedenen Ableitung in I vorzeichenbestimmend ist. Tragen wir dann für $|h| < \varepsilon$ in Tabelle 6.7 die Werte für $P^{(j)}(\xi - |h|)$ und $P^{(j)}(\xi + |h|)$ ein, so stimmen diese im Vorzeichen mit denen von $P^{(j)}(\xi)$ überein, sofern $P^{(j)}(\xi) \neq 0$ ist, in der Zeile dieser Größen also ▲ steht. Wir untersuchen die Vorzeichenverhältnisse über einer Nullsequenz (66) und betrachten zu diesem Zweck für $x = \xi + h \in I$ speziell die Taylorentwicklungen

$$\begin{aligned}
P^{(r+1)}(\xi + h) &= P^{(r+1)}(\xi) + hP^{(r+2)}(\xi) + \cdots, \\
P^{(r)}(\xi + h) &= hP^{(r+1)}(\xi) + \frac{h^2}{2!} P^{(r+2)}(\xi) + \cdots, \\
P^{(r-1)}(\xi + h) &= \frac{h^2}{2!} P^{(r+1)}(\xi) + \frac{h^3}{3!} P^{(r+2)}(\xi) + \cdots, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
P^{(r-k+1)}(\xi + h) &= \frac{h^k}{k!} P^{(r+1)}(\xi) + \frac{h^{k+1}}{(k+1)!} P^{(r+2)}(\xi) + \cdots, \\
P^{(r-k)}(\xi + h) &= P^{(r-k)}(\xi) + \frac{h^{k+1}}{(k+1)!} P^{(r+1)}(\xi) + \cdots.
\end{aligned} \tag{67}$$

Für $x \neq \xi$ sind die Elemente der Folge

$$P^{(r-k)}(x), P^{(r-k+1)}(x), \ldots, P^{(r)}(x), P^{(r+1)}(x) \qquad (x \in I) \tag{68}$$

[1]) Offenbar genügt es, eine Nullsequenz zu betrachten.

ungleich Null. Aus (67) ist zu erkennen, daß die Werte (68) für $x > \xi$ das Vorzeichen von $P^{(r+1)}(\xi)$ besitzen und für $x < \xi$ alternieren, und zwar so, daß beim Übergang von $P^{(r)}(x)$ zu $P^{(r+1)}(x)$ ein Vorzeichenwechsel auftritt[1]). Damit stellt die Zeile der Werte von $P^{(j)}(\xi - |h|)$ in Tabelle 6.7 eine Zahlenfolge dar, wie man sie zur Bestimmung von $\overline{W}(\xi)$ gemäß (61) konstruieren müßte. D. h., $W(x)$ existiert für $x \in I$, $x < \xi$, und es gilt

$$\lim_{x \to \xi - 0} W(x) = \overline{W}(\xi).$$

(68) liefert für $x > \xi$ einen oder keinen Vorzeichenwechsel, je nachdem, ob

$$P^{(r+1)}(\xi) \neq P^{(r-k)}(\xi) \tag{69a}$$

oder

$$P^{(r+1)}(\xi) = P^{(r-k)}(\xi) \tag{69b}$$

gilt. Aus diesem Grunde enthält die Folge der Werte $P^{(j)}(\xi + |h|)$ eben so viele Vorzeichenwechsel wie die um die Nullsequenzen reduzierte Folge der Werte $P^{(j)}(\xi)$, d. h., es gilt

$$\lim_{x \to \xi + 0} W(x) = \underline{W}(\xi).$$

Damit ist (65a) bewiesen.

Wir betrachten nun die Differenz $\overline{W}(\xi) - \underline{W}(\xi)$, d. h. die Abnahme der für $x < \xi$ bzw. $x > \xi$ in I konstanten Größe $W(x)$ beim Überschreiten von ξ. Offenbar treten keine Vorzeichenwechselverluste beim Vergleich der ersten und dritten Folge in Tabelle 6.7 über Abschnitten auf, die auch in der mittleren Zeile mit ▲ belegt sind. Bezüglich einer Nullsequenz führen wir die Betrachtungen auf Grund von (67) getrennt für gerades bzw. ungerades k durch. Die Anzahl der Zeichenwechsel in (68) für $x \in I$ und $x < \xi$ bzw. $x > \xi$ sind in Tabelle 6.8 mit den daraus resultierenden Zeichenwechselverlusten beim Überschreiten von ξ vermerkt. Dabei spielt eine Rolle, ob (69a) oder (69b) gilt. In jedem Falle ist die Anzahl der verlorenen Zeichenwechsel beim Überschreiten von ξ gerade, und man findet insgesamt (65b) mit $m = 0$ bestätigt.

	k gerade			k ungerade		
	$x < \xi$	$x > \xi$	Zeichenwechselverluste	$x < \xi$	$x > \xi$	Zeichenwechselverluste
(69a)	$k+1$	1	k	k	1	$k-1$
(69b)	k	0	k	$k+1$	0	$k+1$

Tabelle 6.8

[1]) In Tabelle 6.7 durch $\overline{\blacktriangle}$ markiert.

Beweisen wir nun (65) unter der Annahme, daß ξ *eine m-fache Nullstelle von* P ist, also

$$P(\xi) = P'(\xi) = \cdots = P^{(m-1)}(\xi) = 0, \tag{70}$$

aber

$$P^{(m)}(\xi) \neq 0$$

gilt. Der Unterschied zu den Betrachtungen im Fall $m = 0$ besteht lediglich darin, daß die Tabelle 6.7 gemäß (70) mit einer Nullsequenz beginnt. Um über diesem Abschnitt die Zeichenwechselverluste beim Überschreiten von ξ zu bestimmen, betrachten wir analog zu (67) die Taylorentwicklungen

$$\begin{aligned}
P^{(m)}(\xi + h) &= P^{(m)}(\xi) + hP^{(m+1)}(\xi) + \cdots, \\
P^{(m-1)}(\xi + h) &= hP^{(m)}(\xi) + \frac{h^2}{2!} P^{(m+1)}(\xi) + \cdots, \\
P^{(m-2)}(\xi + h) &= \frac{h^2}{2!} P^{(m)}(\xi) + \frac{h^3}{3!} P^{(m+1)}(\xi) + \cdots, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
P(\xi + h) &= \frac{h^m}{m!} P^{(m)}(\xi) + \frac{h^{(m+1)}}{(m+1)!} P^{(m+1)}(\xi) + \cdots
\end{aligned} \tag{71}$$

und entnehmen daraus, daß in der Folge

$$P(x), P'(x), \ldots, P^{(m-1)}(x), P^{(m)}(x) \tag{72}$$

für $x \in I$, $x < \xi$, m Zeichenwechsel auftreten, die beim Überschreiten von ξ sämtlich verlorengehen. Die Zeichenwechselverluste über eventuell sonst noch vorhandenen Nullsequenzen sind gerade, und man gewinnt insgesamt (65b). Der Beweis von (65a) entspricht vollkommen dem im Falle $m = 0$.

Nunmehr ergibt sich die Regel von BUDAN-FOURIER sehr einfach: Wir betrachten neben den Intervallendpunkten $\xi_0 = a$ und $\xi_{s+1} = b$ die endlich vielen Stellen

$$\xi_0 = a < \xi_1 < \xi_2 < \cdots < \xi_s < b = \xi_{s+1},$$

an denen irgendein Polynom der Folge (63) verschwindet. Nach (65a) ist $W(x)$ eine auf $[\![a, b]\!] \setminus \{\xi_0, \xi_1, \ldots, \xi_s, \xi_{s+1}\}$ definierte monoton fallende stückweise konstante Funktion. Bezeichnet man die in (65b) auftretende gerade Zahl für ξ_j mit g_j, so gilt auf Grund von (65)

$$\begin{aligned}
\underline{W}(a) - \overline{W}(b) &= \lim_{x \to a+0} W(x) - \lim_{x \to b-0} W(x) \\
&= \sum_{j=1}^{s} \left[\lim_{x \to \xi_j - 0} W(x) - \lim_{x \to \xi_j + 0} W(x) \right] = \sum_{j=1}^{s} g_j + N(a, b),
\end{aligned}$$

also (64) mit $g := \sum_{j=1}^{s} g_j$.

Wir wenden Satz 10 auf die Gleichung (51) an und betrachten zunächst das Intervall (57). Um die Zeichenwechsel in der Polynomfolge (63) bei $x = 1$ und $x = 2$ zu bestimmen, bedient man sich des erweiterten Hornerschen Schemas. Für $x = 1$ resultiert

	1	0	−1	−2	−2	−1
		1	1	0	−2	−4
1	1	1	0	−2	−4	−5
		1	2	2	0	
	1	2	2	0	−4	
		1	3	5		
	1	3	5	5		
		1	4			
	1	4	9			
	1	1				
	1	5				

also $W(1) = 1$

und entsprechend $W(2) = 0$.

Nach (64) ist $N(1, 2) = 1$, da g hier notwendigerweise verschwinden muß. Für das Intervall (59) ergibt sich auf die gleiche Weise

$$W(-1) = 5 \quad \text{und} \quad W\left(-\frac{1}{2}\right) = 3,$$

d. h., im Intervall (59) liegen zwei oder keine Wurzeln der Gleichung (51).

Wie die letzte Betrachtung zeigt, gestattet die Regel von Budan-Fourier im allgemeinen keine endgültige Bestimmung der Anzahl der Nullstellen eines Polynoms in einem Intervall. Das wird jedoch durch eine aufwendigere Methode möglich, welche den Zeichenwechselverlust an Stelle von (63) bezüglich einer sogenannten *Sturmschen Kette* untersucht. Ohne diesen Begriff allgemein zu definieren, sei nur mitgeteilt, wie eine solche, ausgehend von $P(x)$ und $P_1(x) = P'(x)$, durch einen modifizierten Euklidischen Algorithmus konstruiert werden kann. Zunächst dividiert man P durch P_1 mit Rest,

$$P(x) = Q_1(x)\, P_1(x) + R_2(x),$$

und definiert

$$P_2(x) := -R_2(x).$$

Anschließend wird

$$P_1(x) = Q_2(x)\, P_2(x) + R_3(x)$$

gebildet und

$$P_3(x) := -R_3(x)$$

gesetzt und so fort. Da bei diesem Vorgehen die Gradzahlen von $P, P_1, P_2, \ldots$ ständig abnehmen, gelangt man nach endlich vielen Schritten erstmalig zu einem Polynom

$$P_m \equiv \text{const},$$

mit dem die Sturmsche Kette

$$P, P_1, P_2, \ldots, P_m \tag{73}$$

abbricht. Diesbezüglich gilt

Satz 11. *P sei ein Polynom ohne mehrfache Nullstellen*[1]) *und $[\![a, b]\!]$ ein Intervall, für welches $P(a) \neq 0$, $P(b) \neq 0$ und die mit (73) zu bildenden Größen $W(a)$, $W(b)$ existieren. Dann ist die Anzahl der im Intervall $[\![a, b]\!]$ gelegenen Nullstellen von P*

$$N(a, b) = W(a) - W(b). \tag{74}$$

Auf den Beweis des Satzes 11 kann hier nicht eingegangen werden. Wir erläutern seine Anwendung bei der Bestimmung der Anzahl der Wurzeln von (51) im Intervall (59). Für die Sturmsche Kette (73) erhält man

$$\begin{aligned} P(x) &= x^5 - x^3 - 2x^2 - 2x - 1, \\ P_1(x) &= 5x^4 - 3x^2 - 4x - 2, \\ P_2(x) &= 0{,}4x^3 + 1{,}2x^2 + 1{,}6x + 1, \\ P_3(x) &= -22x^2 - 43{,}5x - 35{,}5, \\ P_4(x) &= -0{,}1456635x - 0{,}3398775, \\ P_5(x) &= 53{,}77615. \end{aligned} \tag{75}$$

Wegen $P_5 \neq 0$ treten keine mehrfachen Wurzeln auf. Für die Sturmsche Kette (75) gewinnt man leicht durch eine Überschlagsrechnung

$$W(-1) = W\left(-\frac{1}{2}\right) = 3,$$

d. h., im Intervall (59) liegen auf Grund von (74) keine Lösungen der Gleichung (51). Diese besitzt also neben der im Intervall (57) bestimmten noch zwei Paare konjugiert komplexer Wurzeln.

Wir beschließen die Betrachtungen über die Lokalisierung von Polynomnullstellen mit dem Beweis eines Satzes von Descartes, welcher eine Aussage über die Anzahl der positiven Nullstellen eines Polynoms macht.

Satz 12 (*Cartesische Zeichenregel*). *Die Anzahl der positiven Wurzeln einer algebraischen Gleichung*

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0, \qquad a_n \neq 0,\ a_0 \neq 0,$$

[1]) Nach der in MfL Bd. 3, Kap. 13 und 14, entwickelten Teilbarkeitslehre für Polynome ist das genau dann der Fall, wenn $P_m \neq 0$ ist.

ist (mit Berücksichtigung ihrer Vielfachheit) gleich der für die Koeffizientenfolge $a_n, a_{n-1}, \ldots, a_1, a_0$ *gebildeten Größe* $\underline{W}$ *oder um eine gerade Zahl kleiner als diese.*

Beweis. Die Aussage kann aus dem Satz von BUDAN-FOURIER gefolgert werden. Dazu sei $G > 0$ so gewählt, daß in der Folge (63)

$$\begin{aligned} P(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \\ P'(x) &= n a_n x^{n-1} + (n-1)\, a_{n-1} x^{n-2} + \cdots + a_1, \\ P''(x) &= n(n-1)\, a_n x^{n-2} + (n-1)\,(n-2)\, a_{n-1} x^{n-3} + \cdots + 2a_2, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P^{(n)}(x) &= n!\, a_n \end{aligned}$$

sämtliche Glieder für $x \geqq G$ das Vorzeichen von a_n besitzen. Wir wenden Satz 10 auf das Intervall $[\![0, G]\!]$ an. Dann ist $\underline{W}(0)$ an der Folge

$$a_0,\ 1!a_1,\ 2!a_2,\ 3!a_3,\ \ldots,\ n!a_n$$

oder — was auf dasselbe hinauskommt — an der Koeffizientenfolge zu bestimmen. $\overline{W}(G) = W(G)$ verschwindet, da in der entsprechenden Folge keine Zeichenwechsel auftreten. Damit folgt die Behauptung aus (64).

In Anwendung auf die Gleichung (51) hat man $\underline{W}$ bezüglich der Folge

$$-1,\ -2,\ -2,\ -1,\ 0,\ 1$$

zu bestimmen und findet gemäß Definition 2

$$\underline{W} = 1,$$

was zugleich die Anzahl der positiven Wurzeln von (51) ist.

Die Lokalisierungsmethoden gestatten in Verbindung mit den in MfL Bd. 9, 4.1., erörterten Iterationsverfahren, die reellen Nullstellen eines Polynoms mit beliebiger Genauigkeit zu berechnen. Beispielsweise findet man für die im Intervall $[\![1, 2]\!]$ gelegene Wurzel ξ der Gleichung (51) den Näherungswert

$$\xi^* = 1{,}7346913457, \tag{76}$$

dessen sämtliche Ziffern gültig sind. Die Bestimmung von ξ^* kann man etwa mit der Methode der Bisektion anlaufen lassen und danach mit dem Newtonschen Verfahren fortsetzen.

Nach Berechnung einer reellen Wurzel ξ der Polynomgleichung $P(x) = 0$ liegt es nahe, den Linearfaktor $x - \xi$ gemäß

$$P(x) = (x - \xi)\, \bar{P}(x)$$

abzuspalten und — wenn alle Nullstellen von P bestimmt werden sollen — weiterhin diejenigen von $\bar{P}$ zu ermitteln. Dabei treten i. a. Genauigkeitsverluste auf, die davon herrühren, daß ξ meist nur näherungsweise bekannt ist und auch beim „Abdividieren" des Linearfaktors nach dem Hornerschema Fehler entstehen. Diese Effekte

machen sich natürlich nach mehreren solchen Schritten, d. h. bei den zuletzt berechneten Nullstellen verstärkt bemerkbar. In diesem Zusammenhang ist auf ein Verfahren von MAEHLY hinzuweisen, bei welchem diese Schwierigkeiten nicht auftreten (vgl. [48], 5.5).

Wir wollen uns noch mit der näherungsweisen Bestimmung auch der komplexen Wurzeln einer algebraischen Gleichung

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$

mit reellen Koeffizienten befassen. Ist

$$\zeta_1 = \xi + i\eta$$

eine solche, so auch die dazu konjugiert komplexe Größe

$$\zeta_2 = \bar{\zeta}_1 = \xi - i\eta,$$

und ζ_1, ζ_2 sind Nullstellen des reellen quadratischen Polynoms

$$\begin{aligned} R(x) &= (x - \zeta_1)(x - \zeta_2) = x^2 - (\zeta_1 + \zeta_2)\, x + \zeta_1 \zeta_2 \\ &= x^2 - rx - q. \end{aligned}$$

Weiterhin bedeutet R dieses Polynom zweiten Grades mit zunächst unbestimmten reellen Koeffizienten r, q. Nach Division von P durch R ergibt sich

$$P(x) = Q(x)\,(x^2 - rx - q) + Ax + B, \tag{77}$$

wobei die Koeffizienten A, B des Restpolynoms natürlich von r und q abhängen:

$$A = A(r, q), \qquad B = B(r, q).$$

Gelingt es, r und q so zu bestimmen, daß

$$A(r, q) = 0, \qquad B(r, q) = 0 \tag{78}$$

gilt, so sind die Wurzeln der quadratischen Gleichung

$$x^2 - rx - q = 0 \tag{79}$$

zugleich auch Nullstellen von P und als solche reell oder konjugiert komplex. Beispielsweise findet man für das Polynom der Gleichung (51)

$$\begin{aligned} P(x) = [x^3 + rx^2 + (r^2 + q - 1)\, x + (r^3 + 2rq - r - 2)]\,(x^2 - rx - q) \\ + Ax + B \end{aligned}$$

mit

$$A(r, q) = r^4 + 3r^2 q + q^2 - r^2 - 2r - 2,$$

$$B(r, q) = qr^3 + 2rq^2 - qr - 2q - 1.$$

Das entsprechende Gleichungssystem (78) haben wir in 6.2.1. als Beispiel zum Newtonschen Verfahren betrachtet. Die mit der dort bestimmten Lösung gebildete

Gleichung (79) hat die Wurzeln

$$\zeta_1 = -0{,}166159654 + i \cdot 0{,}938712792, \qquad \zeta_2 = \zeta_1.$$

Spaltet man von dem Polynom

$$Q = x^3 + rx^2 + (r^2 + q - 1)\,x + (r^3 + 2rq - r - 2)$$

den Linearfaktor der reellen Nullstelle ab, so ergibt sich die quadratische Gleichung

$$x^2 + 1{,}402372037x + 0{,}634328020 = 0$$

zur Bestimmung des zweiten Paares konjugiert komplexer Wurzeln der Gleichung (51):

$$x_1 = -0{,}701186018 + i \cdot 0{,}377711778, \qquad x_2 = \bar{x}_1.$$

Das im folgenden beschriebene Verfahren von Bairstow und Hitchcock beruht auf der Lösung des Systems (78) nach dem Newtonschen Verfahren, wobei jedoch die dazu erforderlichen partiellen Ableitungen von A, B und diese Größen selbst mit Hilfe des doppelzeiligen Hornerschemas 5.2.(112) aus r, q und den Koeffizienten von P berechnet werden. Der Darstellung in [48] folgend, gewinnt man durch Differentiation von (77) nach r und q

$$\begin{aligned} \frac{\partial P}{\partial r} \equiv 0 &= \frac{\partial Q}{\partial r} R - xQ + \frac{\partial A}{\partial r} x + \frac{\partial B}{\partial r}, \\ \frac{\partial P}{\partial q} \equiv 0 &= \frac{\partial Q}{\partial q} R - Q + \frac{\partial A}{\partial q} x + \frac{\partial B}{\partial q} \end{aligned} \tag{80}$$

und nach Division von Q durch R

$$Q(x) = Q_1(x)\,R(x) + A_1 x + B_1. \tag{81}$$

Nach (81) ist, wenn ζ_1, ζ_2 die als verschieden angenommenen Nullstellen von R bedeuten,

$$Q(\zeta_i) = A_1\zeta_i + B_1 \qquad (i = 1, 2)$$

und wegen (80)

$$\begin{aligned} -\zeta_i(A_1\zeta_i + B_1) + \frac{\partial A}{\partial r}\zeta_i + \frac{\partial B}{\partial r} &= 0, \\ -(A_1\zeta_i + B_1) + \frac{\partial A}{\partial q}\zeta_i + \frac{\partial B}{\partial q} &= 0. \end{aligned} \tag{82}$$

Die zweite dieser Gleichungen liefert für $i = 1, 2$ ein lineares System zur Bestimmung von $\dfrac{\partial A}{\partial q}$ und $\dfrac{\partial B}{\partial q}$ mit nichtverschwindender Koeffizientendeterminante. Als Lösung liest man unmittelbar ab:

$$\frac{\partial A}{\partial q} = A_1, \qquad \frac{\partial B}{\partial q} = B_1. \tag{83}$$

Auf entsprechende Weise ergibt sich aus der ersten Gleichung (82) wegen $\zeta_i^2 = r\zeta_i + q$

$$B_1\zeta_i + (r\zeta_i + q)\,A_1 = (B_1 + rA_1)\,\zeta_i + qA_1 = \frac{\partial A}{\partial r}\,\zeta_i + \frac{\partial B}{\partial r}$$

und somit

$$\frac{\partial A}{\partial r} = B_1 + rA_1, \qquad \frac{\partial B}{\partial r} = qA_1. \tag{84}$$

Setzt man $Q(x) = b_{n-2}x^{n-2} + b_{n-3}x^{n-3} + \cdots + b_1x + b_0$, so folgt durch einen Koeffizientenvergleich aus (77)

$$\begin{aligned}
a_n &= b_{n-2},\\
a_{n-1} &= b_{n-3} - b_{n-2}r,\\
a_i &= b_{i-2} - b_{i-1}r - b_iq \quad \text{für } i = n-2(-1)2,\\
a_1 &= -b_0r - b_1q + A,\\
a_0 &= -b_0q + B
\end{aligned}$$

oder

$$\begin{aligned}
b_{n-2} &= a_n,\\
b_{n-3} &= a_{n-1} + b_{n-2}r,\\
b_{i-2} &= a_i + b_{i-1}r + b_iq \quad \text{für } i = n-2(-1)2,\\
A &= a_1 + b_0r + b_1q,\\
B &= a_0 + b_0q.
\end{aligned} \tag{85}$$

Entsprechend folgt mit $Q_1(x) = c_{n-4}x^{n-4} + \cdots + c_1x + c_0$ aus (81)

$$\begin{aligned}
c_{n-4} &= b_{n-2},\\
c_{n-5} &= b_{n-3} + c_{n-4}r,\\
c_{i-2} &= b_i + c_{i-1}r + c_iq \quad \text{für } i = n-4(-1)2,\\
A_1 &= b_1 + c_0r + c_1q,\\
B_1 &= b_0 + c_0q.
\end{aligned} \tag{86}$$

Mit Hilfe von (83) bis (86) lassen sich die Iterationen (28) des Newtonschen Verfahrens zur Lösung des Gleichungssystems (78) berechnen. Man kann zeigen, daß dieses für Startwerte r, q, die hinreichend nahe bei einer Lösung liegen, konvergiert.

Die Rekursionen (85) und (86) sind gemäß 5.2.(112) gebildet, wenn man die c_i im doppelzeiligen Hornerschema mit den Koeffizienten der Polynome P bzw. Q und s, t mit q bzw. r identifiziert. Die Größen A, B, A_1, B_1 können demnach mit der (geringfügig zu ändernden) Prozedur *POLKA* (5.2.5.) bestimmt werden.

7. Lineare Optimierung

7.1. Formulierung des LO-Problems

In MfL Bd. 9, 1.1., haben wir die optimale Planung eines Produktionsprozesses als eine Aufgabe der Systemsynthese vorgestellt und auf die Besonderheiten des entsprechenden mathematischen Modells hingewiesen. Wir betrachten als ein weiteres Beispiel für eine derartige Organisationsaufgabe das folgende *Transportproblem*:

In m Depots lagern a_i, $i = 1(1)m$, Mengeneinheiten eines Produktes; davon sind b_j, $j = 1(1)n$, Einheiten an n ortsgebundene Verbraucher zu befördern. Bei einer Versorgung bedeuten

x_{ij} die Menge des Produktes, die vom i-ten Depot an den j-ten Verbraucher geliefert wird,

c_{ij} die entsprechenden Transportkosten für eine Mengeneinheit.

Mit minimalen Transportkosten ist eine ausgeglichene Belieferung zu organisieren, bei welcher jedes Depot geräumt und der Bedarf aller Verbraucher gedeckt wird.

Eine konkrete Aufgabe dieser Art ergibt sich etwa bei der Versorgung von m Baustellen mit Kies aus n Gruben innerhalb eines Territoriums, wenn für einen bestimmten Zeitraum das Förderaufkommen dem Gesamtbedarf angeglichen ist.

Auch in diesem Beispiel wird durch Festlegung der x_{ij} eine Strukturierung einer Objektmenge (Verbraucher und Depots) zu einem System vorgenommen. Diese soll im Hinblick auf ein gegebenes Ziel optimal gestaltet werden. Die mathematische Formulierung des Problems ist offensichtlich. Die lineare Zielfunktion

$$Z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

ist unter den Nebenbedingungen

$$\sum_{j=1}^{n} x_{ij} = a_i \qquad (i = 1, 2, \ldots, m), \tag{2}$$

$$\sum_{i=1}^{m} x_{ij} = b_j \qquad (j = 1, 2, \ldots, n), \tag{3}$$

$$x_{ij} \geqq 0 \qquad (i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n) \tag{4}$$

zum Minimum zu machen.

Sehen wir von speziellen Bezeichnungsweisen ab, so lassen sich die Probleme MfL Bd. 9, 1.1.(7)—(10) und (1)—(4), einheitlich so formulieren:

Problem 1. *Eine lineare Zielfunktion*

$$Z_1(x) = \sum_{j=1}^{n} a_{0j}x_j \tag{5}$$

ist unter den Nebenbedingungen (eines LO-Problems)

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leqq a_{10}, \\ &a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leqq a_{20}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leqq a_{m0}, \end{aligned} \tag{6}$$

$$x_1 \geqq 0,\ x_2 \geqq 0, \ldots, x_n \geqq 0 \tag{7}$$

zum Maximum zu machen, d. h., es ist ein Vektor ξ *in der Erfüllungsmenge* B_1 *der linearen Ungleichungen* (6), (7) *so zu bestimmen, daß*

$$Z_1(x) \leqq Z_1(\xi) \quad \textit{für alle } x \in B_1 \tag{8}$$

gilt.

B_1 heißt der *Zulässigkeitsbereich* des Problems; die a_{ij} im Ausdruck der Zielfunktion und in den Nebenbedingungen sind als gegeben anzusehen. Die Bedeutung von m, n ist eine andere als in den zuvor betrachteten Beispielen. Problem 1 ist die Grundaufgabe der linearen Optimierung (LO-Problem), die wir noch mit einigen Bemerkungen kommentieren.

Bemerkung 1. Mit dem Maximumproblem beherrscht man auch die entsprechende Minimumaufgabe. Nimmt nämlich Z für ξ in B_1 das Maximum an, so besitzt $-Z$ dort ein Minimum und umgekehrt. Jedes derartige Minimumproblem läßt sich demnach in ein äquivalentes LO-Problem (5)—(8) überführen.

Bemerkung 2. Lineare Nebenbedingungen in Gleichungsform — wie etwa (2) und (3) — können durch Paare linearer Ungleichungen ausgedrückt werden. Beispielsweise ist

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = a$$

äquivalent mit

$$\begin{aligned} &a_1x_1 + a_2x_2 + \cdots + a_nx_n \geqq a, \\ &a_1x_1 + a_2x_2 + \cdots + a_nx_n \leqq a. \end{aligned}$$

Bemerkung 3. Jede mit $\geqq$ gebildete lineare Ungleichung läßt sich durch Multiplikation mit -1 äquivalent in eine solche mit $\leqq$ umwandeln. Damit ist alles gesagt,

was nötig ist, um Praxisaufgaben von der Art unserer Beispiele auf die Form (5)—(8) zu bringen.

Bei der mathematischen Behandlung legt man LO-Probleme in einer Gestalt zugrunde, in welcher die Nebenbedingungen (6), die im Vergleich mit (7) gelegentlich die nichttrivialen genannt werden, lineare Gleichungen sind. Ist das wie bei dem oben betrachteten Transportproblem nicht ohnehin schon der Fall, so läßt es sich durch Einführung sogenannter *Schlupfvariabler* (je einer für die in (6) auftretenden linearen Ungleichungen) erreichen. Wir bezeichnen diese mit

$$x_{n+1}, x_{n+2}, \ldots, x_{n+m} \tag{9}$$

und ergänzen die in (5) und (6) auftretenden Koeffizienten a_{ij}, $i = 0(1)m$, $j = 0(1)n$, noch durch die Größen

$$a_{0,n+1} := a_{0,n+2} := \cdots := a_{0,n+m} := 0. \tag{10}$$

Damit formulieren wir folgendes

Problem 2. *Für die lineare Zielfunktion*

$$Z_2 = \sum_{j=1}^{n+m} a_{0j}x_j$$

ist unter den Nebenbedingungen

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + x_{n+1} &&= a_{10},\\
&a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \quad + x_{n+2} &&= a_{20},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots &&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \quad + x_{n+m} &&= a_{m0},
\end{aligned} \tag{11}$$

$$x_1 \geqq 0,\ x_2 \geqq 0, \ldots, x_n \geqq 0, x_{n+1} \geqq 0, \ldots, x_{n+m} \geqq 0 \tag{12}$$

das Maximum zu bestimmen.

Diejenigen Vektoren $x \in \mathbf{R}^{n+m}$, die (11) und (12) genügen, bilden den Zulässigkeitsbereich B_2 des Problems.

Die Probleme 1 und 2 sind in folgendem Sinne äquivalent:

Behauptung 1. *Ist* $\xi^{(1)} \in \mathbf{R}^n$ *eine Lösung des Problems* 1 *und setzt man*

$$\xi_{n+i} := a_{i0} - \sum_{j=1}^{n} a_{ij}\xi_j, \qquad i = 1(1)m, \tag{13}$$

so ist

$$\xi^{(2)} := (\xi_1, \ldots, \xi_n, \xi_{n+1}, \ldots, \xi_{n+m})^{\mathrm{T}}$$

Lösung des Problems 2.

Behauptung 2. *Ist*

$$\xi^{(2)} = (\xi_1, \ldots, \xi_n, \xi_{n+1}, \ldots, \xi_{n+m})^T \in \mathbb{R}^{n+m}$$

eine Lösung des Problems 2, *so ist*

$$\xi^{(1)} := (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$$

Lösung des Problems 1.

Wir beweisen etwa Behauptung 1. Aus der Voraussetzung folgt, daß die Größen (13) nicht negativ sind und definitionsgemäß den Gleichungen (11) genügen. Daher ist $\xi^{(2)} \in B_2$. Gäbe es einen Vektor $\hat{\xi} \in B_2$, für den

$$Z_2(\hat{\xi}) > Z_2(\xi^{(2)})$$

ist, so würde auf Grund von (10) mit $\hat{\xi}^{(1)} := (\hat{\xi}_1, \ldots, \hat{\xi}_n)^T$ auch

$$Z_1(\hat{\xi}^{(1)}) > Z_1(\xi^{(1)})$$

gelten. Da offensichtlich $\hat{\xi}_1 \in B_1$, widerspricht diese Ungleichung der Voraussetzung, daß $\xi^{(1)}$ Lösung des Problems 1 ist.

Entsprechend beweist man die Behauptung 2.

Wir gehen nun noch einen Schritt weiter und lösen uns von der speziellen Form der Koeffizientenmatrix in (11); dabei wird $n + m$ in n umbenannt:

Problem 3. *Es sei* $A = (a_{ij})$, $i = 1(1)m$, $j = 1(1)n$, *eine reelle Matrix des Typs* $m \times n$ *und* Rang $A = m \leqq n$. *Gegeben seien ferner zwei Vektoren*

$$a_{0*} = (a_{01}, a_{02}, \ldots, a_{0n})^T,$$

$$a_{*0} = (a_{10}, a_{20}, \ldots, a_{m0})^T$$

des $\mathbb{R}^n$. *Zu bestimmen ist das absolute Maximum der linearen Funktion*

$$Z_3(x) = a_{0*}^T x \tag{14}$$

über dem Zulässigkeitsbereich

$$B_3 = \{x : Ax = a_{*0} \wedge x \geqq 0\}.^1) \tag{15}$$

Unsere Betrachtungen haben gezeigt, daß man jedes LO-Problem durch Einführung von Schlupfvariablen in den Typ 3 überführen kann. Weiterhin bezeichnen wir die Aufgabe, (14) über dem Zulässigkeitsbereich (15) zu maximieren, als *Normalform des linearen Optimierungsproblems* und entwickeln dafür im Rahmen seiner theoretischen Untersuchung einen Lösungsalgorithmus. Der Index bei Z und B wird dann weggelassen.

[1]) $x \geqq 0$ bedeutet, daß alle Komponenten von x nichtnegativ sind.

7.2. Konvexität des Zulässigkeitsbereichs. Basislösungen

7.2.1. Einführendes Beispiel

Besonders für die Behandlung der linearen Optimierung im fakultativen Unterricht der Schule ist es hilfreich, gewisse Grundvorstellungen über den Zulässigkeitsbereich und die Lage des gesuchten Extremums an zweidimensionalen Problemen des Typs 1 zu entwickeln. Wir betrachten etwa die Aufgabe, das Maximum von

$$Z = 0{,}5x_1 + x_2 \tag{1}$$

unter den Nebenbedingungen

$$\begin{aligned} -x_1 + 3x_2 &\leqq 21, \\ 2x_1 - 3x_2 &\leqq 6, \\ x_1 + x_2 &\leqq 11, \\ x_1 &\geqq 0, \\ x_2 &\geqq 0 \end{aligned} \tag{2}$$

zu berechnen. Der Zulässigkeitsbereich B_1 des Problems — die Erfüllungsmenge der Ungleichungen (2) — ist der Durchschnitt von Halbebenen und mit den Hilfsmitteln der analytischen Geometrie leicht zu bestimmen. Man findet dafür den polygonal berandeten konvexen Bereich der Abb. 7.1. Die Niveaulinien $Z = c = \text{const}$ sind

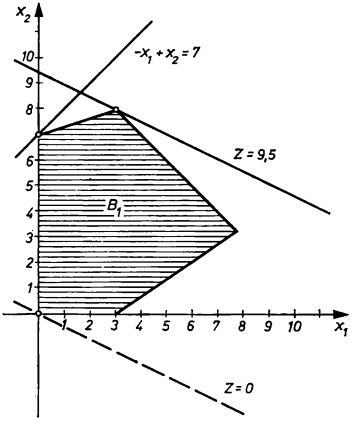


Abb. 7.1

parallele Geraden, und das LO-Problem läuft geometrisch betrachtet darauf hinaus, die Niveaulinie mit dem größten c zu bestimmen, die noch Punkte mit B_1 gemeinsam hat. Das ist, wie man durch Parallelverschiebung der Geraden $Z = 0$ erkennt,

$$Z = 0{,}5x_1 + x_2 = 9{,}5;$$

diese Gerade hat mit B_1 den Eckpunkt

$$x_1 = 3, \qquad x_2 = 8$$

und nur diesen gemeinsam.

Im folgenden wird gezeigt, daß der an diesem Beispiel wahrgenommene Sachverhalt unter einer gewissen Voraussetzung für alle LO-Probleme Gültigkeit hat: Der Zulässigkeitsbereich ist konvex, und das Maximum der Zielfunktion wird in einem Eckpunkt desselben angenommen. Dieser ist durch die Extremalforderung im allgemeinen nicht eindeutig bestimmt. Wenn man etwa in dem betrachteten Beispiel die Zielfunktion (1) durch

$$Z = -x_1 + 3x_2$$

ersetzt, würde diese auf B_1 ihr Maximum in allen Punkten der Verbindungsstrecke von (0, 7) und (3, 8) erreichen.

7.2.2. Konvexe Mengen

In MfL Bd. 4, 1.5.2., wurden die Begriffe der Strecke und konvexen Menge in endlichdimensionalen Zahlenräumen eingeführt, die schon im fakultativen Teil von 5.1.2. bei der Erörterung der Einzigkeitsfrage des linearen Approximationsproblems eine Rolle spielten. Wir beginnen die Untersuchung des Zulässigkeitsbereichs eines LO-Problems mit einer Wiederholung der entsprechenden Definitionen. Es sei noch angemerkt, daß sich die folgenden Betrachtungen auf unendlichdimensionale lineare Räume ausdehnen lassen.

Definition 1. Unter der *Verbindungsstrecke* $S[\boldsymbol{x}, \boldsymbol{y}]$ zweier Punkte $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$ versteht man die durch

$$\boldsymbol{z} \in S[\boldsymbol{x}, \boldsymbol{y}] \Leftrightarrow \bigvee_{\theta \in \{0,1\}} \left(\boldsymbol{z} = \theta \boldsymbol{x} + (1 - \theta)\, \boldsymbol{y}\right)$$

charakterisierte Menge.

Definition 2. $M \subseteq \mathbf{R}^n$ *konvex* $:\Leftrightarrow \bigwedge \boldsymbol{x}, \boldsymbol{y} (\boldsymbol{x}, \boldsymbol{y} \in M \Rightarrow S[\boldsymbol{x}, \boldsymbol{y}] \subseteq M)$.

Offenbar gilt

Satz 1. *Der Durchschnitt beliebig vieler konvexer Mengen ist konvex.*

Der Beweis sei dem Leser als leichte Übungsaufgabe empfohlen.

Im $\mathbf{R}^n$ bezeichnet man die Gesamtheit der Punkte $\boldsymbol{x}$, die mit ihren Koordinaten einer linearen Gleichung

$$l(\boldsymbol{x}) := \sum_{i=1}^{n} a_i x_i = a, \qquad a_i, a \in \mathbf{R}, \qquad \sum_{i=1}^{n} a_i^2 > 0, \tag{3}$$

genügen, als eine *Hyperebene* und die Erfüllungsmenge der Ungleichung

$$l(\boldsymbol{x}) \leqq a \tag{4}$$

allgemein als einen *Halbraum* (Halbebene im Falle $n = 2$).

Hilfssatz 1. *Jede Hyperebene und jeder Halbraum ist konvex.*

Beweis. $\boldsymbol{x}, \boldsymbol{y}$ seien zwei Punkte der Hyperbenene (3) oder des Halbraumes (4). Dann gilt auf Grund der Linearität des Funktionals l (vgl. MfL Bd. 3, 3.2.) für $0 \leqq \theta \leqq 1$

$$l\big(\theta \boldsymbol{x} + (1-\theta)\,\boldsymbol{y}\big) = \theta l(\boldsymbol{x}) + (1-\theta)\, l(\boldsymbol{y}) = \theta a + (1-\theta)\, a = a$$

bzw.

$$l\big(\theta \boldsymbol{x} + (1-\theta)\,\boldsymbol{y}\big) \leqq \theta a + (1-\theta)\, a = a;$$

in jedem Fall gehört also $S[\boldsymbol{x}, \boldsymbol{y}]$ der Menge an.

Da die Zulässigkeitsbereiche der in 7.1. betrachteten LO-Probleme Durchschnitte von Halbräumen bzw. von Halbräumen und Hyperebenen sind, folgt aus Satz 1 und Hilfssatz 1

Satz 2. *Der Zulässigkeitsbereich eines LO-Problems ist konvex.*

Die kleinste konvexe Menge, die eine Menge $U \subseteq \mathbf{R}^n$ umfaßt, heißt deren *konvexe Hülle*. Die folgenden Betrachtungen zu ihrer Charakterisierung sind denen ähnlich, die in MfL Bd. 3, 3.4., bezüglich der linearen Hülle einer Teilmenge des $\mathbf{R}^n$ angestellt wurden.

Definition 3. Es sei $U \subseteq \mathbf{R}^n$. Die *konvexe Hülle* $K(U)$ ist der Durchschnitt aller U umfassenden konvexen Teilmengen des $\mathbf{R}^n$:

$$K(U) := \bigcap K \quad (K \text{ konvexe Teilmenge des } \mathbf{R}^n \text{ mit } K \supseteq U).$$

Definition 4. Es seien $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ Elemente des $\mathbf{R}^n$. Dann heißt jedes Element $\boldsymbol{x} \in \mathbf{R}^n$ von der Form

$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i, \qquad \lambda_i \geqq 0, \qquad i = 1(1)k, \qquad \sum_{i=1}^{k} \lambda_i = 1$$

eine *konvexe Linearkombination* der $\boldsymbol{x}_i$.

Damit läßt sich die konvexe Hülle einer Menge $U \subseteq \mathbf{R}^n$ so charakterisieren:

Satz 3. $\boldsymbol{x} \in K(U) \Leftrightarrow \boldsymbol{x}$ *ist konvexe Linearkombination endlich vieler Elemente aus* U.

Beweis. a) Wir nehmen $\boldsymbol{x} \in K(U)$ an und zeigen zunächst, daß die Menge M aller konvexen Linearkombinationen über U konvex ist. Dazu werden zwei Elemente $\xi, \eta \in M$ betrachtet.

Für diese gelten Darstellungen der Form

$$\xi = \sum_{i=1}^{k} \lambda_i x_i, \qquad \lambda_i \geqq 0, \qquad \sum_{i=1}^{k} \lambda_i = 1, \qquad x_i \in U,$$

$$\eta = \sum_{j=1}^{l} \mu_j y_j, \qquad \mu_j \geqq 0, \qquad \sum_{j=1}^{l} \mu_j = 1, \qquad y_j \in U,$$

so daß mit $\theta \in \mathbf{R}$

$$\theta\xi + (1-\theta)\,\eta = \sum_{i=1}^{k} \theta\lambda_i x_i + \sum_{j=1}^{l} (1-\theta)\,\mu_j y_j$$

ist. Für $0 \leqq \theta \leqq 1$ ist das eine konvexe Linearkombination der Elemente x_i, y_j, $i = 1(1)k$, $j = 1(1)l$, denn es gilt

$$\theta\lambda_i \geqq 0, \qquad (1-\theta)\,\mu_j \geqq 0$$

und

$$\sum_{i=1}^{k} \theta\lambda_i + \sum_{j=1}^{l} (1-\theta)\,\mu_j = \theta + (1-\theta) = 1.$$

Es gehört also auch die Verbindungsstrecke von ξ und η zu M, d. h., diese Menge ist konvex. Da jedes Element $x \in U$ auf Grund der Darstellung $x = 1 \cdot x$ eine konvexe Linearkombination über U ist, gilt weiter $U \subseteq M$ und nach Definition 3 auch $K(U) \subseteq M$, also

$$x \in K(U) \Rightarrow x \in M.$$

b) Die Umkehrung dieser Implikation wird durch Induktion nach der Länge k der konvexen Linearkombinationen über U bewiesen. Für $k = 1$ sind das die Elemente von U selbst, die offenbar zu $K(U)$ gehören. Wir nehmen nun an, daß alle konvexen Linearkombinationen, deren Länge kleiner oder gleich k ($k > 1$) ist, in $K(U)$ liegen, und betrachten eine konvexe Linearkombination der Länge $k+1$:

$$x = \sum_{i=1}^{k+1} \lambda_i x_i, \qquad x_i \in U, \qquad \lambda_i \geqq 0, \qquad \sum_{i=1}^{k+1} \lambda_i = 1.$$

Beim Nachweis, daß x Element von $K(U)$ ist, können wir uns auf den Fall beschränken, daß alle Koeffizienten λ_i positiv sind, da sonst $x \in K(U)$ schon aus der Induktionsvoraussetzung folgen würde. Für x wird eine Darstellung der Form

$$x = \sum_{i=1}^{k+1} \lambda_i x_i = \theta \sum_{i=1}^{k} \mu_i x_i + (1-\theta)\,x_{k+1} \tag{5}$$

gesucht, in der

$$\mu_i > 0, \qquad \sum_{i=1}^{k} \mu_i = 1 \tag{5a}$$

und $0 < \theta < 1$ ist. Es ist naheliegend, die Bestimmung von μ_i, $i = 1(1)k$, und θ auf Grund der Beziehungen

$$\lambda_i = \theta\mu_i, \qquad i = 1(1)k, \tag{6}$$

und

$$\lambda_{k+1} = 1 - \theta \tag{7}$$

vorzunehmen. Aus (7) folgt wegen $\lambda_i > 0$, $i = 1(1)k+1$, und $\sum_{i=1}^{k+1} \lambda_i = 1$

$$\theta = \sum_{i=1}^{k} \lambda_i \quad \text{und} \quad 0 < \theta < 1. \tag{8}$$

Mit dem nach (8) berechneten Wert θ erhält man aus (6) μ_i-Werte, die auch (5a) erfüllen und zusammen mit θ die gewünschte Darstellung (5) liefern. x erweist sich demnach als Element der Verbindungsstrecke von $\sum_{i=1}^{k} \lambda_i x_i$ und x_{k+1}. Die Elemente $\sum_{i=1}^{k} \lambda_i x_i$ und x_{k+1} sind konvexe Linearkombinationen über U der Länge k bzw. 1 und gehören nach Induktionsannahme zu $K(U)$. Wegen der Konvexität dieser Menge gilt das folglich auch für x.

Bemerkung 4. Speziell kann U eine endliche Menge $U = \{x_1, x_2, \ldots, x_n\}$ sein. Dann besteht $K(U)$ aus allen Elementen der Form

$$x = \sum_{i=1}^{n} \lambda_i x_i, \qquad \lambda_i \geqq 0, \qquad \sum_{i=1}^{n} \lambda_i = 1. \tag{9}$$

Von besonderem Interesse sind die Punkte einer konvexen Menge K, die nicht dem Inneren der Verbindungsstrecke irgend zweier Elementen von K angehören. Diese sogenannten *Eckpunkte* sind also folgendermaßen zu definieren:

Definition 5. x *Eckpunkt einer konvexen Menge* K: $\Leftrightarrow$

$$x \in K \wedge \neg \bigvee_{x_1, x_2 \in K} \bigvee_{\theta \in]0,1[} \left(x_1 \neq x_2 \wedge x = \theta x_1 + (1-\theta) x_2\right).$$

Beispielsweise sind die Punkte, die man im Hinblick auf Abb. 7.1 anschaulich als Ecken des Zulässigkeitsbereichs des LO-Problems aus 7.2.1. bezeichnen würde, auch Eckpunkte im Sinne von Definition 5.

Definition 6. Eine beschränkte konvexe Menge mit nur endlich vielen Eckpunkten heißt ein *konvexes Polyeder*.

Satz 4. *Die konvexe Hülle einer endlichen Menge* $U = \{x_1, x_2, \ldots, x_n\}$ *ist ein konvexes Polyeder, dessen Eckpunkte in* U *enthalten sind.*

Beweis. Wir zeigen zunächst, daß $K(U)$ beschränkt ist. In MfL Bd. 4, 1.5.3., wurde die Beschränktheit einer Menge des $\mathbf{R}^n$ mit Hilfe der euklidischen Norm charakterisiert. Auf Grund des Satzes 11, 6.1.3., kann man $\|\cdot\|_2$ durch eine beliebige Norm dieses Raumes ersetzen. Bedeutet dann

$$m := \max_{i \in \{1,2,\ldots,n\}} \|x_i\|,$$

so gilt für ein x aus $K(U)$ auf Grund von (9)

$$\|x\| = \left\| \sum_{i=1}^{n} \lambda_i x_i \right\| \leqq \sum_{i=1}^{n} |\lambda_i| \cdot \|x_i\| \leqq m \sum_{i=1}^{n} \lambda_i = m,$$

woraus die Beschränktheit von $K(U)$ folgt.

Nun betrachten wir einen Eckpunkt x von $K(U)$ und wählen dafür unter den Darstellungen (9) eine solche minimaler Länge $l \leqq n$:

$$x = \sum_{i=1}^{l} \lambda_{j_i} x_{j_i}, \qquad \lambda_{j_i} > 0, \qquad \sum_{i=1}^{l} \lambda_{j_i} = 1.$$

Es wird gezeigt, daß $l = 1$ ist, also $x \in U$. Wäre $l > 1$, so würde aus

$$x = \sum_{i=1}^{l-1} \lambda_{j_i} x_{j_i} + \lambda_{j_l} x_{j_l}$$

wie beim Beweis von Satz 3b) eine Darstellung der Form

$$x = \theta \sum_{i=1}^{l-1} \mu_i x_{j_i} + (1 - \theta)\, x_{j_l}$$

mit $\theta \in]0, 1[$ und $\mu_i > 0$, $\sum_{i=1}^{l-1} \mu_i = 1$ folgen, d. h., x könnte nicht Eckpunkt von $K(U)$ sein.

7.2.3. Basislösungen

Wir betrachten das LO-Problem in der Normalform und definieren bezüglich des Zulässigkeitsbereiches 7.1.(15)

Definition 7. x *Basislösung*: $\Leftrightarrow$

$$Ax = a_{*0} \wedge x \text{ enthält genau } m \text{ von Null verschiedene Komponenten } x_{i_1}, x_{i_2}, \ldots, x_{i_m} \wedge \text{die Spaltenvektoren } a_{i_1}, a_{i_2}, \ldots, a_{i_m} \text{ der Matrix } A \text{ sind linear unabhängig.}$$

Eine Basislösung x heißt *zulässig*, wenn $x \geqq 0$ ist.

Es sei daran erinnert, daß A eine Matrix vom Typ $m \times n$ ist, für die Rang $A = m \leqq n$ gilt.

Zulässige Basislösungen gehören dem Zulässigkeitsbereich B des LO-Problems an. Sofern nicht ausdrücklich etwas anderes gesagt wird, beziehen sich auch im folgenden alle Aussagen auf das LO-Problem in Normalform und die in 7.1.(14), (15) eingeführten Bezeichnungen.

Zur Erläuterung der in Definition 7 eingeführten Begriffe transformieren wir das LO-Problem des Beispiels (1), (2) durch Einführung von Schlupfvariablen x_3, x_4, x_5 in Normalform. Für den Zulässigkeitsbereich des äquivalenten Problems 2 ergibt sich gemäß 7.1.(11), (12) folgendes System linearer Gleichungen und Ungleichungen:

$$\begin{aligned} -x_1 + 3x_2 + x_3 &= 21, \\ 2x_1 - 3x_2 + x_4 &= 6, \\ x_1 + x_2 + x_5 &= 11, \\ x_i \geqq 0, \quad i &= 1(1)5. \end{aligned} \tag{10}$$

Man überprüft sofort, daß die den Ecken des Bereichs der Abb. 7.1 entsprechenden Punkte zulässige Basislösungen sind. Beispielsweise erhält man aus der Ecke $x_1 = 0$, $x_2 = 7$ den Punkt x mit den (10) genügenden Koordinaten

$$x_1 = 0, \quad x_2 = 7, \quad x_3 = 0, \quad x_4 = 27, \quad x_5 = 4. \tag{11}$$

Die zu betrachtenden Spaltenvektoren von A sind

$$a_2 = \begin{pmatrix} 3 \\ -3 \\ 1 \end{pmatrix}, \quad a_4 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad a_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix};$$

man überzeugt sich leicht, daß diese linear unabhängig sind. Schließlich treten in (11) genau $m = 3$ von Null verschiedene Koordinaten auf, und es ist $x \geqq 0$. Damit erweist sich x als zulässige Basislösung.

Wir wollen den Nebenbedingungen (2) noch die Ungleichung

$$-x_1 + x_2 \leqq 7 \tag{12}$$

hinzufügen. Die hierdurch bestimmte Halbebene enthält den Zulässigkeitsbereich der Abb. 7.1, und dieser wird folglich durch das Hinzufügen von (12) nicht verändert. Das neue LO-Problem hat dieselbe Lösung wie die in 7.2.1. betrachtete Aufgabe; es wird sich aber zeigen, daß die überflüssige Nebenbedingung (12) Schwierigkeiten bei deren Bestimmung verursacht. Um diese genauer zu charakterisieren, gehen wir wieder gemäß 7.1.(11), (12) zur Normalform über. Da eine nichttriviale Nebenbedingung hinzugekommen ist, muß auch eine weitere Schlupfvariable x_6 eingeführt werden, und an Stelle von (10) ergibt sich

$$\begin{aligned} -x_1 + 3x_2 + x_3 &= 21, \\ 2x_1 - 3x_2 + x_4 &= 6, \\ x_1 + x_2 + x_5 &= 11, \\ -x_1 + x_2 + x_6 &= 7, \\ x_i \geqq 0, \quad i &= 1(1)6. \end{aligned} \tag{13}$$

Der Ecke $x_1 = 0$, $x_2 = 7$ entspricht jetzt der Punkt x mit den (13) genügenden Koordinaten

$$x_1 = 0, \quad x_2 = 7, \quad x_3 = 0, \quad x_4 = 27, \quad x_5 = 4, \quad x_6 = 0.$$

Damit ist die Existenz einer Lösung des linearen Gleichungssystems (13) mit weniger als m (= 4) von Null verschiedenen Komponenten nachgewiesen. Wegen dieser Erscheinung bezeichnet man das betrachtete LO-Problem als *ausgeartet* (oder *entartet*) und definiert allgemein:

Definition 8. Ein LO-Problem heißt *ausgeartet*, wenn das lineare Gleichungssystem $Ax = a_{*0}$ Lösungen mit weniger als m von Null verschiedenen Komponenten besitzt.

Bei der Entwicklung eines Lösungsalgorithmus für das LO-Problem werden wir voraussetzen, daß dieses nicht ausgeartet ist.

Das Beispiel legt die Vermutung nahe, daß zwischen den Ecken von B und den zulässigen Basislösungen ein Zusammenhang besteht. In der Tat gilt

Satz 5. *x ist zulässige Basislösung $\Rightarrow$ x ist Eckpunkt des Zulässigkeitsbereiches B.*

Beweis. Wenn nötig, kann man durch Umbenennung der Variablen erreichen, daß die ersten m Koordinaten der betrachteten Basislösung $\boldsymbol{x}$ von Null verschieden sind. Dann gilt

$$x_1 > 0, x_2 > 0, \ldots, x_m > 0, x_{m+1} = x_{m+2} = \cdots = x_n = 0, \tag{14}$$

und die Spaltenvektoren $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$ von $\boldsymbol{A}$ sind linear unabhängig. Nehmen wir nun im Sinne eines indirekten Beweises an, daß $\boldsymbol{x}$ nicht Eckpunkt von B ist, dann existieren nach Definition 5 zwei verschiedene Punkte $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ in B, für die mit einem gewissen $\theta \in]0, 1[$

$$\boldsymbol{x} = \theta \boldsymbol{x}_1 + (1 - \theta)\, \boldsymbol{x}_2 \tag{15}$$

gilt. Wir bezeichnen deren Koordinaten mit $x_i^{(1)}$ bzw. $x_i^{(2)}$, $i = 1(1)n$. Da $x_i^{(1)}, x_i^{(2)} \geqq 0$ und θ, $1 - \theta$ positiv sind, folgt aus (15) mit Beachtung von (14)

$$x_j^{(1)} = x_j^{(2)} = 0 \quad \text{für } j = m + 1(1)n, \tag{16}$$

also

$$\boldsymbol{a}_1 x_1^{(1)} + \boldsymbol{a}_2 x_2^{(1)} + \cdots + \boldsymbol{a}_m x_m^{(1)} = \boldsymbol{a}_{*0},$$

$$\boldsymbol{a}_1 x_1^{(2)} + \boldsymbol{a}_2 x_2^{(2)} + \cdots + \boldsymbol{a}_m x_m^{(2)} = \boldsymbol{a}_{*0}$$

und

$$\boldsymbol{a}_1(x_1^{(2)} - x_1^{(1)}) + \boldsymbol{a}_2(x_2^{(2)} - x_2^{(1)}) + \cdots + \boldsymbol{a}_m(x_m^{(2)} - x_m^{(1)}) = \boldsymbol{0}. \tag{17}$$

Auf Grund der linearen Unabhängigkeit der $\boldsymbol{a}_j$, $j = 1(1)m$, folgt aus (17) $x_j^{(1)} = x_j^{(2)}$ und mit (16)

$$\boldsymbol{x}_1 = \boldsymbol{x}_2.$$

Das aber ist ein Widerspruch zur vorausgesetzten Verschiedenheit dieser Punkte.

Satz 5 läßt sich umkehren, wenn das LO-Problem nicht ausgeartet ist:

Satz 6. *LO-Problem nicht ausgeartet $\wedge$ x Eckpunkt von $B \Rightarrow$ x zulässige Basislösung.*

Beweis. $\boldsymbol{x}$ sei Eckpunkt von B und r die Anzahl seiner von Null verschiedenen Koordinaten. Aus $\boldsymbol{x} \in B$ folgt

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}_{*0} \tag{18}$$

und — wegen der ausgeschlossenen Ausartung — nach Definition 8

$$r \geqq m. \tag{19}$$

Ohne Beschränkung der Allgemeinheit sei wieder angenommen, daß die ersten r Koordinaten von $\boldsymbol{x}$ nicht verschwinden, also positiv sind. Dann kann (18) in der Form

$$\boldsymbol{a}_1 x_1 + \boldsymbol{a}_2 x_2 + \cdots + \boldsymbol{a}_r x_r = \boldsymbol{a}_{*0} \tag{20}$$

ausgedrückt werden. Die in (20) auftretenden Spaltenvektoren $\boldsymbol{a}_j$ sind linear unabhängig. Im Sinne eines indirekten Beweises dieser Behauptung wird eine nichttriviale Linearkombination

$$\boldsymbol{a}_1 y_1 + \boldsymbol{a}_2 y_2 + \cdots + \boldsymbol{a}_r y_r = \boldsymbol{0} \tag{21}$$

dieser Vektoren zum Nullvektor betrachtet. Darin sei etwa $y_\varrho \neq 0$. Wir multiplizieren (21) mit einer positiven Zahl s und addieren und subtrahieren diese Gleichung zu bzw. von (20). Auf diese Weise ergibt sich

$$\begin{aligned} &\boldsymbol{a}_1(x_1 + sy_1) + \boldsymbol{a}_2(x_2 + sy_2) + \cdots + \boldsymbol{a}_\varrho(x_\varrho + sy_\varrho) + \cdots + \boldsymbol{a}_r(x_r + sy_r) = \boldsymbol{a}_{*0}, \\ &\boldsymbol{a}_1(x_1 - sy_1) + \boldsymbol{a}_2(x_2 - sy_2) + \cdots + \boldsymbol{a}_\varrho(x_\varrho - sy_\varrho) + \cdots + \boldsymbol{a}_r(x_r - sy_r) = \boldsymbol{a}_{*0}. \end{aligned} \tag{22}$$

Durch Wahl eines genügend kleinen s kann erreicht werden, daß sämtliche der in (22) auftretenden Größen $x_i \pm sy_i$, $i = 1(1)r$, positiv sind. Wir bilden damit die Vektoren $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ mit den Koordinaten

$$\begin{aligned} x_i^{(1)} &= x_i + sy_i \quad \text{für } i = 1(1)r, \\ x_i^{(1)} &= 0 \qquad\quad \text{für } i = r + 1(1)n \text{ [1]} \end{aligned}$$

bzw.

$$\begin{aligned} x_i^{(2)} &= x_i - sy_i \quad \text{für } i = 1(1)r, \\ x_i^{(2)} &= 0 \qquad\quad \text{für } i = r + 1(1)n. \end{aligned}$$

An der ϱ-ten Koordinate erkennt man, daß $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, und wegen (22) gilt $\boldsymbol{x}_1, \boldsymbol{x}_2 \in B$; außerdem ist

$$\boldsymbol{x} = \frac{1}{2}(\boldsymbol{x}_1 + \boldsymbol{x}_2). \tag{23}$$

(23) widerspricht der Voraussetzung, daß $\boldsymbol{x}$ Eckpunkt von B ist. Die dem $\mathbf{R}^m$ angehörenden $\boldsymbol{a}_j$, $j = 1(1)r$, sind also tatsächlich linear unabhängig, und es ist

$$r \leqq m, \tag{24}$$

da mehr als m Vektoren dieses Raumes stets linear abhängig sind. Aus (19) und (24) folgt $r = m$, d. h., $\boldsymbol{x}$ ist Basislösung.

Satz 7. *Es gibt nur endlich viele Basislösungen zu einem LO-Problem.*

Beweis. In dem n-Tupel $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ kann man $\binom{n}{m}$ Systeme $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ als *Basisvariable* auszeichnen, womit diejenigen Variablen gemeint sind, die in einer Basislösung nicht verschwinden. Sind dann noch die Spaltenvektoren $\boldsymbol{a}_{i_1}, \boldsymbol{a}_{i_2}, \ldots, \boldsymbol{a}_{i_m}$ linear unabhängig, so gibt es genau eine Lösung des Systems $\boldsymbol{Ax} = \boldsymbol{a}_{*0}$ mit ver-

[1] Auch für das Folgende sei vereinbart: Ist in einem Laufbereich die untere Grenze größer als die obere, so ist dieser als leer zu betrachten.

schwindenden Komponenten, $x_{j_1}, x_{j_2}, \ldots, x_{j_{n-m}}$, wenn

$$\{i_1, i_2, \ldots, i_m\} \cup \{j_1, j_2, \ldots, j_{n-m}\} = \{1, 2, \ldots, n\}.$$

Diese ist Basislösung, falls $x_{i_k} \neq 0$ für $k = 1(1)m$. Damit erweist sich $\binom{n}{m}$ als eine obere Schranke für die Anzahl der Basislösungen.

Mit Satz 7 folgt aus Satz 6

Satz 8. *Der Zulässigkeitsbereich B eines nicht ausgearteten LO-Problems besitzt nur endlich viele Eckpunkte.*

und — in Verbindung mit Definition 6 und Satz 2 —

Satz 9. *Ist der Zulässigkeitsbereich B eines nicht ausgearteten LO-Problems beschränkt, so ist B ein konvexes Polyeder.*

7.3. Das Fundamentaltheorem der linearen Optimierung

Man überlegt sich leicht, daß der Zulässigkeitsbereich eines LO-Problems unbeschränkt und dieses dann unlösbar sein kann. Setzen wir jedoch seine Lösbarkeit voraus, so wird im folgenden gezeigt, daß eine Lösung — oder, wie man auch sagt, ein *optimaler Vektor* — bereits in der endlichen Menge der Basislösungen enthalten ist, sofern keine Ausartung vorliegt. Wir schicken dem Beweis dieses Lokalisierungssatzes Betrachtungen über speziell gebildete Elemente des Zulässigkeitsbereiches voraus, die auch für die anschließende Erörterung des *Simplexalgorithmus* wichtig sind.

Es sei angenommen, daß der Spaltenvektor $\boldsymbol{a}_k$ von den Vektoren $\boldsymbol{a}_i$, $i = 1(1)r$, $r < k$, linear abhängig ist:

$$\boldsymbol{a}_k = \sum_{i=1}^{r} \lambda_{ik} \boldsymbol{a}_i, \qquad \lambda_{ik} \in \mathbf{R}. \tag{1}$$

Für jedes $\boldsymbol{x} \in B$ gilt

$$\sum_{i=1}^{n} x_i \boldsymbol{a}_i = \sum_{i=1}^{r} x_i \boldsymbol{a}_i + \sum_{i=r+1}^{n} x_i \boldsymbol{a}_i = \boldsymbol{a}_{*0}. \tag{2}$$

Multipliziert man (1) mit einem beliebigen $h \in \mathbf{R}$ und subtrahiert diese Gleichung von (2), so folgt

$$\sum_{i=1}^{r} (x_i - h\lambda_{ik})\, \boldsymbol{a}_i + \sum_{i=r+1}^{k-1} x_i \boldsymbol{a}_i + (x_k + h)\, \boldsymbol{a}_k + \sum_{i=k+1}^{n} x_i \boldsymbol{a}_i = \boldsymbol{a}_{*0},$$

d. h., der Vektor $\boldsymbol{x}^{(h)}$ mit den Koordinaten

$$x_i^{(h)} = \begin{cases} x_i - h\lambda_{ik} & \text{für } i = 1(1)r, \\ x_i & \text{für } i = r+1(1)k-1, \\ x_i + h & \text{für } i = k, \\ x_i & \text{für } i = k+1(1)n \end{cases} \tag{3}$$

genügt der Gleichung

$$A\boldsymbol{x}^{(h)} = \boldsymbol{a}_{*0}; \tag{4}$$

der entsprechende Wert der Zielfunktion ist

$$Z(\boldsymbol{x}^{(h)}) = \sum_{i=1}^{n} a_{0i}x_i^{(h)} = Z(\boldsymbol{x}) + h\left(a_{0k} - \sum_{i=1}^{r} a_{0i}\lambda_{ik}\right)$$

oder, wenn

$$d_k := a_{0k} - \sum_{i=1}^{r} a_{0i}\lambda_{ik} \tag{5}$$

gesetzt wird,

$$Z(\boldsymbol{x}^{(h)}) = Z(\boldsymbol{x}) + hd_k. \tag{6}$$

Die beiden folgenden Hilfssätze lassen erkennen, daß die Größe (5) im weiteren eine wichtige Rolle spielen wird. *Darin bedeutet k einen Index, für den* (1) *mit* $k > r$ *gilt.*

Hilfssatz 1. $\boldsymbol{x} \in B \wedge x_i > 0$ für $i = 1(1)r$ und $i = k \wedge d_k < 0 \Rightarrow \neg$ (*x ist optimaler Vektor*).

Beweis. Für die durch

$$h' := \max_{\substack{i=1(1)r \\ \lambda_{ik}<0}} \left\{\frac{x_i}{\lambda_{ik}}, -x_k\right\} \tag{7}$$

definierte Größe gilt

$$h' < 0 \quad \text{und} \quad \boldsymbol{x}^{(h')} \geqq 0. \tag{8}$$

Mit Beachtung von (4) ist also $\boldsymbol{x}^{(h')} \in B$, und nach (6) hat man auf Grund der Voraussetzung

$$Z(\boldsymbol{x}^{(h')}) > Z(\boldsymbol{x}).$$

Daraus folgt die Behauptung.

Hilfssatz 2. *LO-Problem lösbar* $\wedge\ d_k > 0 \Rightarrow \{\lambda_{ik}: i = 1(1)r \wedge \lambda_{ik} > 0\} \neq \emptyset$.

Beweis. Im Sinne eines indirekten Beweises wird angenommen, daß die in der Konklusion auftretende Menge leer ist. Dann definiert (3) für ein fest gewähltes $\boldsymbol{x} \in B$ und jedes positive h einen Vektor des Zulässigkeitsbereiches. Für solche $\boldsymbol{x}^{(h)}$ gilt nach (6) wegen $d_k > 0$

$$\lim_{h\to\infty} Z(\boldsymbol{x}^{(h)}) = +\infty.$$

Die Zielfunktion nimmt also in B beliebige große Werte an, und das LO-Problem besitzt im Widerspruch zur Voraussetzung keine Lösung.

Nunmehr sind wir in der Lage, für ein nicht ausgeartetes LO-Problem das folgende *Fundamentaltheorem* zu beweisen.

Satz 1.

$$\bigvee_{x \in B} (x \text{ optimaler Vektor}) \Rightarrow \bigvee_{x \in B} (x \text{ optimale zulässige Basislösung}).$$

Beweis. $\tilde{x}$ bedeutet einen der Voraussetzung dieses Satzes genügenden Vektor mit k von Null verschiedenen — also positiven — Komponenten. Nötigenfalls durch Umnumerierung von Variablen sei dafür gesorgt, daß $\tilde{x}_i > 0$ für $i = 1(1)k$ und — falls maximal r der Spaltenvektoren $a_1, a_2, \ldots, a_k$ linear unabhängig sind — dieses auf a_i, $i = 1(1)r$, zutrifft. Wegen der ausgeschlossenen Entartung gilt

$$k \geqq m \tag{9}$$

und wegen $a_i \in \mathbf{R}^m$, $i = 1(1)n$,

$$m \geqq r. \tag{10}$$

Es ist also

$$k \geqq r. \tag{11}$$

$k = r$ gilt genau dann, wenn in (1) und (2) das Gleichheitszeichen steht, und das charakterisiert $\tilde{x}$ als Basislösung. In diesem Fall bleibt nichts zu beweisen. Ist aber $k > r$, so konstruieren wir ausgehend von $\tilde{x}$ einen optimalen Vektor x' mit $k' < k$ von Null verschiedenen Komponenten (s. u.), wobei wieder ohne Beschränkung der Allgemeinheit $x'_i > 0$, $i = 1(1)k'$, angenommen werden kann. Gemäß (11) gilt

$$k' \geqq r', \tag{12}$$

wenn r' die maximale Anzahl linear unabhängiger Vektoren im System der a_i, $i = 1(1)k'$, bedeutet. Gilt in (12) das Gleichheitszeichen, so ist x' optimaler Basisvektor, anderenfalls konstruiert man nach dem Prinzip, das zu x' geführt hat, einen optimalen Vektor x'' mit $k'' < k'$ von Null verschiedenen Koordinaten und so fort. Wegen $k > k' > k'' > \cdots$ wird nach endlich vielen Schritten ein optimaler Vektor bestimmt, der (11) mit dem Gleichheitszeichen erfüllt, also optimaler Basisvektor ist.

Konstruktion von x'. Wenn $r < k$ und somit a_k von den a_i, $i = 1(1)r$, linear abhängig ist, folgt aus der Kontraposition des Hilfssatzes 1 mit $x = \tilde{x}$

$$d_k \geqq 0.$$

Wir wollen

$$d_k = 0 \tag{13}$$

zeigen und nehmen im Sinne eines indirekten Beweises $d_k > 0$ an. Dann ergibt sich aus Hilfssatz 2 unter Berücksichtigung der Lösbarkeit des LO-Problems

$$L := \{\lambda_{ik} : i = 1(1)r \wedge \lambda_{ik} > 0\} \neq \emptyset.$$

Auf Grund dessen existiert

$$h' := \min_{\substack{i=1(1)r \\ \lambda_{ik}>0}} \left\{\frac{\tilde{x}_i}{\lambda_{ik}}\right\} \tag{14}$$

und ist positiv. Für den mit (14) gemäß (3) gebildeten Vektor $x' := x^{(h')}$ ist $x' \geqq 0$, also wegen (4) $x' \in B$ und nach (6)

$$Z(x') = Z(\tilde{x}) + h'd_k > Z(\tilde{x}).$$

Dem widerspricht, daß $\tilde{x}$ optimal ist, und folglich gilt (13).

Nunmehr wird h' gemäß (7) bzw. (14) gebildet, je nachdem, ob $L = \emptyset$ oder $L \neq \emptyset$. Dann ist $x' := x^{(h')}$ ein zulässiger Vektor, für den nach (6) wegen (13)

$$Z(x') = Z(\tilde{x})$$

gilt; außerdem verschwindet bei x' nach (3) mindestens eine Koordinate mehr als bei $\tilde{x}$.

7.4. Der Simplexalgorithmus

Dieses Verfahren ist eine effektive Methode zur Bestimmung einer optimalen Basislösung. Wir setzen voraus, daß keine Ausartung vorliegt und für das vorgelegte LO-Problem schon eine zulässige Basislösung bekannt ist.

In der kombinatorischen Topologie (vgl. [41]) bezeichnet man gewisse durch konvexe Hüllenbildung erzeugte Punktmengen als *Simplexe*. Sie stellen n-dimensionale Verallgemeinerungen der Begriffe Dreieck ($n = 2$) und Tetraeder ($n = 3$) dar. Wegen ihrer Beziehung zu den konvexen Polyedern und damit den Zulässigkeitsbereichen von LO-Problemen wurde die Bezeichnung „Simplex" mit dem dafür maßgebenden Lösungsalgorithmus verknüpft.

7.4.1. Simplexkriterium und Austauschverfahren

Vorgelegt sei ein nicht ausgeartetes LO-Problem und eine seiner zulässigen Basislösungen $\tilde{x}$. Ohne Beschränkung der Allgemeinheit können wir bezüglich der Koordinaten wieder

$$\tilde{x}_1 > 0, \quad \ldots, \; \tilde{x}_m > 0, \quad \tilde{x}_{m+1} = \tilde{x}_{m+2} = \cdots = \tilde{x}_n = 0 \tag{1}$$

annehmen. Es gilt also

$$\sum_{i=1}^{m} \tilde{x}_i a_i = a_{*0}, \tag{2}$$

und die Spaltenvektoren $a_1, a_2, \ldots, a_m$ von A sind linear unabhängig. Für einen Index $k > m$ sei

$$a_k = \sum_{i=1}^{m} \lambda_{ik} a_i, \tag{3}$$

und entsprechend 7.3.(5) wird

$$d_k := a_{0k} - \sum_{i=1}^{m} a_{0l}\lambda_{ik} \tag{4}$$

definiert. Dann gilt bezüglich (1) der

Satz 1.

$$\bigvee_{k\in\{m+1,\ldots,n\}} (d_k > 0) \Rightarrow \bigvee_{x'\in B} \big(x' \textit{ Basislösung} \wedge Z(x') > Z(\tilde{x})\big) \vee \textit{LO-Problem nicht lösbar.}$$

Beweis. Bei den folgenden Überlegungen werden wir uns auf Definitionen und Sätze von 7.3. beziehen; dabei ist stets $r = m$ und $x = \tilde{x}$ zu setzen. Nach Voraussetzung gilt $d_k > 0$ für ein gewisses $k \in \{m+1, \ldots, n\}$; außerdem sei angenommen, daß das LO-Problem eine Lösung besitzt. Dann ist auf Grund von 7.3., Hilfssatz 2, die Menge

$$\{\lambda_{ik} : i = 1(1)m \wedge \lambda_{ik} > 0\}$$

nicht leer und die positive Zahl h' gemäß 7.3.(14) bestimmbar. Für den nach 7.3.(3) gebildeten Vektor $x' := x^{(h')} \in B$ ergibt sich nach 7.3.(6)

$$Z(x') > Z(\tilde{x}). \tag{5}$$

Wir zeigen noch, daß x' Basislösung ist: Nach Konstruktion hat dieser Vektor nicht mehr als m von Null verschiedene Koordinaten, wegen der ausgeschlossenen Entartung aber auch nicht weniger. Wegen $x'_k > 0$ muß genau eine der Koordinaten x'_i, $i = 1(1)m$, verschwinden. Bezeichnet man deren Index mit l, so ist

$$\lambda_{lk} > 0 \quad \text{und} \quad h' = \frac{\tilde{x}_l}{\lambda_{lk}}.$$

Mit (3) ergibt sich für eine beliebige Linearkombination der zu den Basisvariablen gehörenden Spaltenvektoren von A

$$y_1 a_1 + y_2 a_2 + \cdots + y_{l-1} a_{l-1} + y_{l+1} a_{l+1} + \cdots + y_m a_m + y_k a_k$$

$$= \sum_{\substack{i=1 \\ i\neq l}}^{m} (y_i + y_k\lambda_{ik})\, a_i + \lambda_{lk} y_k a_l .$$

Ist diese dem Nullvektor gleich, so folgt wegen der linearen Unabhängigkeit der a_i, $i = 1(1)m$, und $\lambda_{lk} \neq 0$ zunächst $y_k = 0$ und weiter für alle $i = 1(1)m$, $i \neq l$ auch $y_i = 0$. Die Spaltenvektoren bei den Basisvariablen von x' sind also linear unabhängig. Wegen (5) stellt x' eine zulässige Basislösung dar, für welche die Zielfunktion einen größeren Wert annimmt als bei $\tilde{x}$.

In kontraponierter Form lautet Satz 1:

$$\textit{LO-Problem lösbar} \wedge \bigwedge_{x'\in B} \big(x' \textit{ Basislösung} \Rightarrow Z(x') \leqq Z(\tilde{x})\big) \Rightarrow \bigwedge_{k\in\{m+1,\ldots,n\}} (d_k \leqq 0).$$

Offenbar ist die Prämisse dieser Implikation aus der Aussage

$$\bar{x} \text{ ist optimale Basislösung des LO-Problems } (\bar{x}_i > 0,\ i = 1(1)m)$$

ableitbar, so daß folgendes notwendiges Kriterium für eine optimale Basislösung gilt:

Satz 2.

$$\bar{x} \textit{ optimale Basislösung } (\bar{x}_i > 0,\ i = 1(1)m) \Rightarrow \bigwedge_{k\in\{m+1,\ldots,n\}} (\bar{d}_k \leqq 0).$$

Dieses ist auch hinreichend:

Satz 3.

$$\bigwedge_{k\in\{m+1,\ldots,n\}} (\bar{d}_k \leqq 0) \Rightarrow \bar{x} \textit{ optimale Basislösung } (\bar{x}_i > 0,\ i = 1(1)m).$$

Beweis. Es sei x ein beliebiges Element von B, d. h., es ist

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = a_{*0}, \qquad x_i \geqq 0, \quad i = 1(1)n. \tag{6}$$

Wir dehnen die Bestimmung der λ_{ik} gemäß (3) auch auf die Basisvektoren $a_1, a_2, \ldots, a_m$ aus und erhalten dafür

$$\lambda_{ik} = \delta_{ik}, \qquad k = 1(1)m; \tag{7}$$

zur Abkürzung wird noch

$$z_k := \sum_{i=1}^{m} a_{0i}\lambda_{ik} \tag{8}$$

gesetzt. Dann ist nach (7) $a_{0k} = z_k$ für $k = 1(1)m$, und auf Grund der Voraussetzung von Satz 3 ist $a_{0k} \leqq z_k$ für $k = m + 1(1)n$, in jedem Fall also

$$a_{0k} \leqq z_k, \qquad k = 1(1)n. \tag{9}$$

Mithin gilt

$$Z(x) = \sum_{k=1}^{n} a_{0k}x_k \leqq \sum_{k=1}^{n} z_k x_k. \tag{10}$$

Substituiert man (3) in (6), so folgt

$$x_1 \sum_{i=1}^{m} \lambda_{i1}a_i + x_2 \sum_{i=1}^{m} \lambda_{i2}a_i + \cdots + x_n \sum_{i=1}^{m} \lambda_{in}a_i = a_{*0}$$

und nach Umordnung

$$\left(\sum_{j=1}^{n} x_j\lambda_{1j}\right) a_1 + \left(\sum_{j=1}^{n} x_j\lambda_{2j}\right) a_2 + \cdots + \left(\sum_{j=1}^{n} x_j\lambda_{mj}\right) a_m = a_{*0}. \tag{11}$$

Aus (10) resultiert nach Einsetzen von (8) durch entsprechende Umordnung (die a_{0i} spielen die Rolle der a_i)

$$Z(x) \leqq \left(\sum_{j=1}^{n} x_j\lambda_{1j}\right) a_{01} + \left(\sum_{j=1}^{n} x_j\lambda_{2j}\right) a_{02} + \cdots + \left(\sum_{j=1}^{n} x_j\lambda_{mj}\right) a_{0m}. \tag{12}$$

Wegen der linearen Unabhängigkeit der Vektoren $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$ kann $\boldsymbol{a}_{*0}$ nur auf eine Weise aus diesen linear kombiniert werden. Daher gilt auf Grund von (2) und (11)

$$\bar{x}_i = \sum_{j=1}^{n} \lambda_{ij} x_j, \qquad i = 1(1)m, \tag{13}$$

und in Verbindung mit (12)

$$Z(\boldsymbol{x}) \leqq a_{01}\bar{x}_1 + a_{02}\bar{x}_2 + \cdots + a_{0m}\bar{x}_m = Z(\bar{\boldsymbol{x}}),$$

was zu beweisen war.

Nach Satz 2 und 3 ist

$$\bigwedge_{k\in\{m+1,\ldots,n\}} (d_k \leqq 0) \tag{14}$$

also notwendig und hinreichend dafür, daß eine Basislösung $\boldsymbol{x}$, deren erste m Koordinaten von Null verschieden sind, optimal ist. (14) heißt deshalb das *Simplexkriterium*. In Verbindung damit konzipieren wir ein *Austauschverfahren*, das nach endlich vielen Schritten eine optimale Basislösung liefert. Zunächst wird diese Grundform des Simplexalgorithmus in einem PAP dargestellt und dann im einzelnen erläutert. Die theoretische Begründung des Austauschverfahrens ist in der Bemerkung 5 enthalten.

Bemerkungen zum PAP der Abb. 7.2.

1. Die Eingabe am Anfang betrifft die Daten des LO-Problems in Normalform. $\boldsymbol{x}$ bedeutet eine zulässige Basislösung mit $x_i > 0$, $i = 1(1)m$, zu der gemäß der folgenden Anweisung der Wert der Zielfunktion zu berechnen ist.

2. Für $i = 1(1)m$ und $k = m + 1(1)n$ sind dann die Größen λ_{ik} und d_k zu bestimmen. Im Anschluß an diese Erläuterungen befassen wir uns mit einem Verfahren zur Berechnung der λ_{ik}.

3. Das folgende Entscheidungskästchen enthält das Simplexkriterium (14). Wie schon bemerkt, bedeutet seine Erfüllung, daß die im PAP mit $\boldsymbol{x}$ bezeichnete Basislösung optimal ist. Diese Information wird zusammen mit dem Wert der Zielfunktion im Ja-Zweig ausgegeben, und der Algorithmus bricht ab.

4. Anderenfalls wird im Hinblick auf Satz 1 ein $k \in \{m + 1, \ldots, n\}$ bestimmt, für welches $d_k > 0$. Ist dann die Aussage

$$\{\lambda_{ik} : i = 1(1)m \wedge \lambda_{ik} > 0\} = \emptyset$$

wahr, so folgt aus der Kontraposition des Hilfssatzes 2 in 7.3. die Unlösbarkeit des LO-Problems. Nach Ausgabe dieser Information im Ja-Zweig bricht der Algorithmus ab.

5. Im Nein-Zweig wird wie beim Beweis von Satz 1 eine Basislösung $\boldsymbol{x}'$ konstruiert, für die $Z(\boldsymbol{x}') \geqq Z(\boldsymbol{x})$ ist. Dabei sind die Koordinaten x'_i, $i = 1(1)m$, $i \neq l$, und x'_k positiv, während die übrigen verschwinden. Die neuen Basisvektoren sind also

$$\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{l-1}, \boldsymbol{a}_{l+1}, \ldots, \boldsymbol{a}_m, \boldsymbol{a}_k. \tag{15}$$

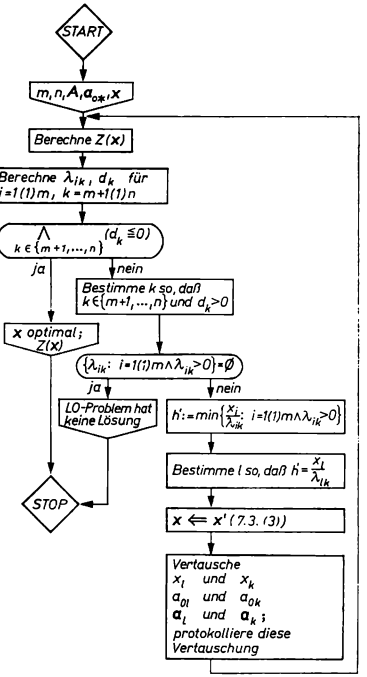


Abb. 7.2

6. Die im Verfahren benutzte zulässige Basislösung wird mit dem in 5. bestimmten Basisvektor x' aktualisiert. Dabei ist durch eine Umformung des Problems dafür zu sorgen, daß der zu Anfang erhobenen Forderung des Nichtverschwindens der ersten m Komponenten genügt wird. Man erreicht das durch Vertauschung der Vektorkoordinaten x_l und x_k, der Spaltenvektoren $\mathfrak{a}_l$ und $\mathfrak{a}_k$ sowie der Koeffizienten a_{0l} und a_{0k} in der Zielfunktion. Diese Transformation ist in geeigneter Weise zu protokollieren und bei der Ausgabe der Endlösung zu berücksichtigen.

7. Da sich der Wert der Zielfunktion beim Austausch der Basislösungen vergrößert hat und nur endlich viele Basislösungen existieren, muß das Verfahren nach endlich vielen Schritten mit der Bestimmung eines optimalen Basisvektors oder der Feststellung der Nichtlösbarkeit des LO-Problems abbrechen. Man wird bemüht

sein, in jedem Schritt einen möglichst großen Zuwachs der Zielfunktion zu gewinnen. Mit Rücksicht auf 7.3.(6) bedeutet das, $k \in \{m+1, \ldots, n\}$ so zu wählen, daß d_k unter den möglichen dieser Werte maximal ist.

Wir befassen uns jetzt mit der Berechnung der λ_{ij}, $j = m + 1(1)n$, $i = 1(1)m$, in (3). Dabei bedeuten $l \in \{1, 2, \ldots, m\}$ und $k \in \{m+1, \ldots, n\}$ weiterhin die im PAP der Abb. 7.2 bestimmten Indizes. Wir gehen von der Annahme aus, daß die Basisdarstellung (3) für die am Anfang des Verfahrens vorliegenden $\boldsymbol{a}_j$, $j = m + 1(1)n$, bekannt ist. Mit Hilfe der so gegebenen λ_{ij} wollen wir die Koeffizienten λ'_{ij} in (3) nach Ausführung der durch l und k bestimmten Vertauschung berechnen. Konsequent wird dabei an folgender Indizierung festgehalten: λ'_{ij} bedeutet den in der Basisdarstellung von $\boldsymbol{a}_j$ bei $\boldsymbol{a}_i$ stehenden Koeffizienten. Überhaupt werden die in 7. eingeführten Größenbezeichnungen nicht aktualisiert, d. h., eine bestimmte Bezeichnung behält ihre ursprüngliche Bedeutung bei. Das gilt nicht für die symbolischen Adressen der für die Abarbeitung des Simplexalgorithmus reservierten Speicherplätze und insofern auch nicht für den PAP der Abb. 7.2.

Zunächst bestimmten wir λ'_{il} für

$$i = 1, 2, \ldots, l-1, k, l+1, \ldots, m. \tag{16}$$

Es ist $\boldsymbol{a}_k = \sum_{i=1}^{m} \lambda_{ik} \boldsymbol{a}_i$ und folglich

$$\lambda_{lk} \boldsymbol{a}_l = \boldsymbol{a}_k - \sum_{\substack{i=1 \\ i \neq l}}^{m} \lambda_{ik} \boldsymbol{a}_i. \tag{17}$$

Da $\lambda_{lk} > 0$ ist und die Vektoren auf der rechten Seite von (17) linear unabhängig sind, gilt

$$\lambda'_{kl} = \frac{1}{\lambda_{lk}} \quad \text{und} \quad \lambda'_{il} = -\frac{\lambda_{ik}}{\lambda_{lk}} \tag{18}$$

für die übrigen Indizes (16). Nun sei

$$j \in \{m+1, \ldots, n\}, \quad j \neq k. \tag{19}$$

Dann ist

$$\boldsymbol{a}_j = \sum_{\substack{i=1 \\ i \neq l}}^{m} \lambda'_{ij} \boldsymbol{a}_i + \lambda'_{kj} \boldsymbol{a}_k = \sum_{\substack{i=1 \\ i \neq l}}^{m} \lambda'_{ij} \boldsymbol{a}_i + \lambda'_{kj} \sum_{i=1}^{m} \lambda_{ik} \boldsymbol{a}_i = \sum_{i=1}^{m} \lambda_{ij} \boldsymbol{a}_i,$$

und die lineare Unabhängigkeit der Vektoren $\boldsymbol{a}_i$, $i = 1(1)m$, erlaubt folgenden Koeffizientenvergleich:

$i = l$: $\lambda'_{kj} \lambda_{lk} = \lambda_{lj}$, also

$$\lambda'_{kj} = \frac{\lambda_{lj}}{\lambda_{lk}}; \tag{20}$$

$i \neq l$: $\lambda'_{ij} + \lambda'_{kj}\lambda_{ik} = \lambda_{ij}$ und wegen (20)

$$\lambda'_{ij} = \lambda_{ij} - \frac{\lambda_{lj}\lambda_{ik}}{\lambda_{lk}}. \tag{21}$$

7.4.2. Rechenschema

Wir entwickeln ein Rechenschema für die Abarbeitung des Simplexalgorithmus von Hand. Das diesem zugrunde liegende Formular spiegelt etwa die Speicherplatzverteilung der ALGOL-Prozedur in 7.4.3. wider. Die dabei auftretenden Felder werden im Ablauf des Verfahrens aktualisiert, d. h. sukzessive mit neuen Werten überschrieben. Das als *Simplextabelle* (*-tableau*) bezeichnete Schema enthält m Zeilen und $n-m$ Spalten zur Aufnahme der jeweils aktuellen λ-Koeffizienten in der Basisdarstellung (3). Zusätzlich findet man am linken und oberen Tabelleneingang von 1 bis m bzw. $m+1$ bis n indizierte eindimensionale Felder, die der Protokollierung der im PAP ausgewiesenen Vertauschung dienen, und solche, welche die in einem Zyklus bestimmten Werte der neuen Basisvariablen und der entsprechenden Koeffizienten der Zielfunktion enthalten. In einer $(m+1)$-ten Zeile erscheinen die Werte der Zielfunktion für die aktuelle zulässige Basislösung und der d_j, $j = m+1(1)n$. Diese gewinnt man gemäß 7.1.(14) und (4) wesentlich als Skalarprodukte von Spaltenvektoren des Schemas. Das Anfangstableau hat danach die Form von Tabelle 7.1. Alle in Tabelle 7.1 einzutragenden Werte beziehen sich auf die Startgrößen des Simplexalgorithmus. Sollte das Simplexkriterium nicht erfüllt sein, so wird die im Zyklus des PAP bestimmte neue Basislösung mit den entsprechenden Vertauschungen und den gemäß (18), (20) und (21) zu berechnenden λ-Koeffizienten der Basisdarstel-

				$m+1$	$m+2$	...	k	...	n
				$m+1$	$m+2$	...	k	...	n
	i	x_i	a_{0i}	$a_{0,m+1}$	$a_{0,m+2}$	...	a_{0k}	...	a_{0n}
1	1	x_1	a_{01}	$\lambda_{1,m+1}$	$\lambda_{1,m+2}$		λ_{1k}		λ_{1n}
2	2	x_2	a_{02}	$\lambda_{2,m+1}$	$\lambda_{2,m+2}$		λ_{2k}		λ_{2n}
⋮	⋮	⋮	⋮						
l	l	x_l	a_{0l}	$\lambda_{l,m+1}$	$\lambda_{l,m+2}$		λ_{lk}		λ_{ln}
⋮	⋮	⋮	⋮						
m	m	x_m	a_{0m}	$\lambda_{m,m+1}$	$\lambda_{m,m+2}$		λ_{mk}		λ_{mn}
$m+1$		$Z(x)$		d_{m+1}	d_{m+2}	...	d_k	...	d_n

Tabelle 7.1

				$m+1$	$m+2$	...	k	...	n
				$m+1$	$m+2$	...	l	...	n
	i	x_i	a_{0i}	$a_{0,m+1}$	$a_{0,m+2}$	...	a_{0l}	...	a_{0n}
1	1	x'_1	a_{01}	$\lambda'_{1,m+1}$	$\lambda'_{1,m+2}$		λ'_{1l}		λ'_{1n}
2	2	x'_2	a_{02}	$\lambda'_{2,m+1}$	$\lambda'_{2,m+2}$		λ'_{2l}		λ'_{2n}
⋮	⋮	⋮	⋮						
l	k	x'_k	a_{0k}	$\lambda'_{k,m+1}$	$\lambda'_{k,m+2}$		λ'_{kl}		λ'_{kn}
⋮	⋮	⋮	⋮						
m	m	x'_m	a_{0m}	$\lambda'_{m,m+1}$	$\lambda'_{m,m+2}$		λ'_{ml}		λ'_{mn}
$m+1$		$Z(x')$		d'_{m+1}	d'_{m+2}		d'_l		d'_n

Tabelle 7.2

lung der Vektoren $a_{m+1}, a_{m+2}, \ldots, a_{k-1}, a_l, a_{k+1}, \ldots, a_n$ durch die Vektoren (15) in einer umgeformten Simplextabelle (vgl. Tabelle 7.2) festgehalten. In Tabelle 7.2 ist

$$d'_j = a_{0j} - \sum_{\substack{i=1 \\ i \neq l}}^{m} a_{0i}\lambda'_{ij} - a_{0k}\lambda'_{kj}, \tag{22}$$

wobei j einem der Indizes (19) oder l gleich sein kann. Für diese j ist wiederum das Simplexkriterium $d'_j \leqq 0$ zu überprüfen und gegebenenfalls ein weiterer Umformungsschritt mit der Tabelle durchzuführen, wobei jedesmal die für die Bestimmung der neuen Basislösung und das Austauschverfahren erforderlichen Größen h', k, l zu berechnen sind, usw.; mit Rücksicht auf die Verwendung in Programmen wird h' weiterhin mit h bezeichnet.

Wir erläutern das Vorgehen an dem Beispiel 7.2.1. in der durch 7.2.3.(10) gegebenen Normalform. Startbasislösung sei

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 21, \quad x_4 = 6, \quad x_5 = 11. \tag{23}$$

Um der Annahme zu entsprechen, daß die ersten m $(= 3)$ Koordinaten positiv sind, formulieren wir die Aufgabe so, daß die eingeführten Schlupfvariablen die Indizes 1, 2 und 3 erhalten: Es ist dann das Maximum der Zielfunktion

$$Z = 0{,}5x_4 + x_5 \tag{24}$$

über dem Zulässigkeitsbereich

$$\begin{aligned} x_1 \quad &- x_4 + 3x_5 = 21, \\ x_2 \quad &+ 2x_4 - 3x_5 = 6, \\ x_3 + \; &x_4 + x_5 = 11, \\ x_i \geqq 0, \quad &i = 1(1)5, \end{aligned} \tag{25}$$

zu bestimmen. Der Basislösung (23) entspricht

$$x_1 = 21, \quad x_2 = 6, \quad x_3 = 11, \quad x_4 = 0, \quad x_5 = 0, \tag{26}$$

und die Anfangstabelle 7.1 hat die Gestalt der Tabelle 7.3. Wegen $d_4 = 0{,}5$ und $d_5 = 1$ ist das Simplexkriterium nicht erfüllt, und wir müssen zur Gewinnung der umgeformten Simplextabelle die Größen h, k, l berechnen. Mit Beachtung der Bemerkung 7 findet man dafür die Werte der ersten Zeile in Tabelle 7.4. Nach 7.3.(3) ($r = m = 3$) ergibt sich als neue Basislösung

$$x_1' = 0, \quad x_2' = 27, \quad x_3' = 4, \quad x_4' = 0, \quad x_5' = 7, \tag{27}$$

und die umgeformte Simplextabelle erhält auf Grund von (18), (20) und (21) die Gestalt der Tabelle 7.5. Auf Grund des Simplexkriteriums ist die Basislösung (27) noch nicht optimal, und es werden aus Tabelle 7.5 die h,k,l-Größen der zweiten Zeile von Tabelle 7.4 ermittelt, um damit einen weiteren Austauschschritt durchzurechnen. Behalten wir die mit (24) und (25) eingeführte Koordinatenumerierung bei, so ist gemäß 7.3.(3)

$$\begin{aligned} x_5' &= 7 - 3\lambda_{54}' = 8, \\ x_2' &= 27 - 3\lambda_{24}' = 24, \\ x_3' &= 4 - 3\lambda_{34}' = 0, \\ x_4' &= 0 + 3 \quad = 3, \\ x_1' &= 0 \end{aligned} \tag{28}$$

die neue Basislösung, und man gewinnt diesbezüglich mit dem Austauschverfahren die Tabelle 7.6. Man erkennt: Das Simplexkriterium ist erfüllt, und (28) stellt eine optimale Basislösung dar. In der ursprünglichen Formulierung der Aufgabe entspricht dieser die Ecke der Abb. 7.1, welche auf Grund einer anschaulich-geometrischen Betrachtungsweise in 7.2.1. als optimal ermittelt wurde. (27) ist der vom Ursprung verschiedenen Ecke des Zulässigkeitsbereichs auf der x_2-Achse zuzuordnen.

				4	5
				4	5
	i	x_i	a_{0i}	1/2	1
1	1	21	0	−1	3
2	2	6	0	2	−3
3	3	11	0	1	1
4		0		1/2	1

Tabelle 7.3

h	k	l
7	5	1
3	4	3

Tabelle 7.4

				4	5
				4	1
	i	x_i	a_{0i}	1/2	0
1	5	7	1	−1/3	1/3
2	2	27	0	1	1
3	3	4	0	4/3	−1/3
4		7		5/6	−1/3

Tabelle 7.5

				4	5
				3	1
	i	x_i	a_{0i}	0	0
1	5	8	1	1/4	1/4
2	2	24	0	−3/4	5/4
3	4	3	1/2	3/4	−1/4
4		19/2		−5/8	−1/8

Tabelle 7.6

7.4.3. ALGOL-Prozedur zum Simplexalgorithmus

Wir entwickeln eine Prozedur *SIMPLEX* zur Lösung eines LO-Problems nach dem in 7.4.2. erörterten Rechenschema. Diesem liegt die in Abb. 7.3 angegebene Modifikation des PAP der Abb. 7.2 zugrunde. In der Prozedur treten als formale Parameter auf:

Bezeichnung	*Benennung*	*Bedeutung*
m,n	**integer**	Gemäß Darstellung des LO-Problems in Normalform.
A,X	**array**	Eindimensionale Felder zur Speicherung der Koeffizienten der Zielfunktion bzw. Koordinaten der Startbasislösung. Nach Abarbeitung der Prozedur steht auf *X* eine optimale Basislösung.
ST	**array**	Zweidimensionales Feld zur Speicherung der λ-Koeffizienten in der Basisdarstellung 7.4. (3) (Simplextabelle).
w	**Boolean**	*w* ist mit **false** belegt, wenn das LO-Problem nicht lösbar ist, sonst mit **true**.
z	**real**	*z* ist nach Abarbeitung der Prozedur mit dem Maximum der Zielfunktion über dem Zulässigkeitsbereich belegt.

Nehmen wir an, daß in einem Programm die den formalen Parametern entsprechenden aktuellen Felder die gleichen Bezeichnungen tragen, so sind diese in der Form *A,X*[1:*n*] bzw. *ST*[1:*m*,*m* + 1:*n*] zu vereinbaren.

Lokal werden in *SIMPLEX* folgende Größen benutzt:

Variable *l*, *k*, *h* mit der oben eingeführten Bedeutung;

i, *j* sind Laufvariable;

hh ist reellwertig und dient zum Zwischenspeichern.

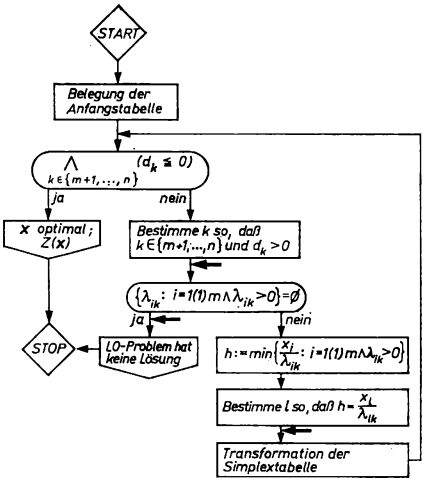


Abb. 7.3

Auf dem Feld P werden die Vertauschungen bei der Bildung neuer Basissysteme protokolliert; D speichert die für das Simplexkriterium maßgebenden Größen d_k, und auf dem Feld Y werden sukzessive nach dem Simplexalgorithmus zulässige Basislösungen gebildet, deren letzte optimal ist.

```
procedure SIMPLEX(m,n,A,X,ST,w,z); value m,n;
Boolean w; integer m,n; real z; array A,X,ST;
begin
      integer i,j,l,k; real h,hh;
      integer array P[1:n]; array Y[1:n], D[m + 1:n];
      w := true;
      for i := 1 step 1 until n do begin P[i] := i; Y[i] := X[i] end;
LA:   for j := m + 1 step 1 until n do begin D[j] := A[j];
      for i := 1 step 1 until m do D[j] := D[j] − A[i] × ST[i,j]
                                   end;
      k := m + 1;
      for i := 1 step 1 until n do if D[i] > D[k] then k := i;
      if D[k] ≦ 0 then begin z := 0;
      for i := 1 step 1 until m do begin z := z + A[i] × Y[i]; X[P[i]] := Y[i]
                                   end;
```

```
        for i := m + 1 step 1 until n do X[P[i]] := Y[i]; goto LC
                                  end;
        for i := 1 step 1 until m do if ST[i,k] > 0 then
        begin h := Y[i]/ST[i,k]; l := i;
        for j := l + 1 step 1 until m do if ST[j,k] > 0 then begin
        hh := Y[j]/ST[j,k]; if hh < h then begin h := hh; l := j end
                                                          end;
        goto LB
        end;
        w := false; goto LC;
LB:     for i := 1 step 1 until m do Y[i] := Y[i] - h × ST[i,k];
        Y[l] := h; Y[k] := 0;
        j := P[l]; P[l] := P[k]; P[k] := j;
        hh := A[l]; A[l] := A[k]; A[k] := hh;
        hh := 1/ST[l,k]; ST[l,k] := hh;
        for j := m + 1 step 1 until k - 1 do ST[l,j] := ST[l,j] × hh;
        for j := k + 1 step 1 until n do ST[l,j] := ST[l,j] × hh;
        for i := 1 step 1 until l - 1 do begin
        for j := m + 1 step 1 until k - 1 do ST[i,j] := ST[i,j] - ST[i,k] × ST[l,j];
        for j := k + 1 step 1 until n do ST[i,j] := ST[i,j] - ST[i,k] × ST[l,j];
        ST[i,k] := -ST[i,k] × hh        end;
        for i := l + 1 step 1 until m do begin
        for j := m + 1 step 1 until k - 1 do ST[i,j] := ST[i,j] - ST[i,k] × ST[l,j];
        for j := k + 1 step 1 until n do ST[i,j] := ST[i,j] - ST[i,k] × ST[l,j];
        ST[i,k] := -ST[i,k] × hh        end;
        goto LA;
LC:     end
```

Abschließend noch einige Erläuterungen zur Prozedur *SIMPLEX*: Zu Anfang werden die Boolesche Variable w mit **true** und die Felder P, Y mit den Zahlen $1, 2, \ldots, n$ in der natürlichen Anordnung bzw. den Koordinaten der Startbasislösung belegt. Die folgenden Anweisungen erzeugen wesentlich die Anfangstabelle und bestimmen k. Im Fall $D[k] \leqq 0$ ist das Simplexkriterium erfüllt, und Y enthält bis eventuell auf Koordinatenvertauschungen eine optimale Basislösung; diese wird nach Berechnung des Wertes der Zielfunktion und Herstellung der ursprünglichen Koordinatenfolge dem Feld X übermittelt, und der Algorithmus bricht ab. Sonst wird die auf ST gespeicherte Simplextabelle transformiert. Dieser Programmteil beginnt mit einer Laufanweisung, welche die durch Hinweisungspfeile herausgehobene Struktur des PAP der Abb. 7.3 realisiert. Damit werden h und l bestimmt. Sind sämtliche $ST[i,k]$, $i = 1(1)m$, kleiner oder gleich Null, so hat das LO-Problem keine Lösung und der Algorithmus bricht nach Belegung von w mit **false** ab. Anderenfalls werden diese Anweisungen übersprungen, und es erfolgt die Bildung eines neuen zulässigen Basisvektors gemäß 7.3.(3) und der k-l-Positionstausch. Die weiteren

Anweisungen betreffen die Umformung der Simplextabelle gemäß den Formeln 7.4.1.(18), (20) und (21). Der Rücksprung zur Marke LA leitet einen neuen durch das Simplexkriterium gesteuerten Zyklus ein.

7.4.4. Bestimmung einer zulässigen Basislösung

Der von uns erörterte Simplexalgorithmus setzt die Kenntnis einer zulässigen Basislösung voraus. Die Bestimmung einer solchen ist sehr einfach bei einem Problem 1, wenn die in 7.1.(6) auftretenden Größen a_{i0}, $i = 1(1)m$, positiv sind. Nach Einführung von Schlupfvariablen gewinnt man hier für das Problem in Normalform zu den Nebenbedingungen 7.1.(11) mit

$$x_1 = x_2 = \cdots = x_n = 0, \quad x_{n+1} = a_{10}, \quad x_{n+2} = a_{20}, \ldots, x_{n+m} = a_{m0} \tag{29}$$

eine zulässige Basislösung. Für das Beispiel aus 7.2.1. erhält man so auf Grund von 7.2.(10) den in 7.4. benutzten Startvektor (23).

Es gibt mehrere Verfahren, nach denen im allgemeinen Fall eine zulässige Basislösung konstruiert werden kann. Das im folgenden betrachtete geht von der Annahme aus, daß das LO-Problem nicht ausgeartet und schon ein Element $\tilde{\boldsymbol{x}}$ seines Zulässigkeitsbereichs mit k ($\geqq m$) von Null verschiedenen Koordinaten bekannt ist. Letzteres bedeutet für die Praxis keine erhebliche Einschränkung der Allgemeinheit, da das zu optimierende System mit seinen (vor der Optimierung) gegebenen Parametern ein solches Element liefert.

Der zu erörternde Algorithmus ist im PAP der Abb. 7.4 dargestellt; wegen der fortlaufenden Aktualisierung wurde $\tilde{\boldsymbol{x}}$ mit $\boldsymbol{x}$ bezeichnet. Bei der Begründung des Verfahrens beziehen wir uns auf die in 7.3. entwickelte Theorie.

Zunächst wird durch Koordinatenvertauschung dafür gesorgt, daß $x_i > 0$ für $i = 1(1)k$ und — wenn maximal r der Vektoren $\boldsymbol{a}_i$, $i = 1(1)k$, linear unabhängig sind — dieses für $\boldsymbol{a}_i$, $i = 1(1)r$, zutrifft. Nach 7.3.(11) ist dann

$$k \geqq r,$$

und $k = r$ gilt genau dann, wenn $\boldsymbol{x}$ zulässige Basislösung ist. Es sei also $k > r$. Nach Bestimmung der Basisdarstellung von $\boldsymbol{a}_k$ durch die Vektoren $\boldsymbol{a}_i$, $i = 1(1)r$, und der Größe d_k verzweigt sich das Programm, je nachdem, ob diese negativ ist oder nicht. Im zweiten Fall ist zu prüfen, ob die in 7.3. eingeführte Menge L leer ist. Trifft dies zu, so folgt aus 7.3., Hilfssatz 2, die Unlösbarkeit des Problems, sofern $d_k > 0$. In allen anderen Fällen wird gemäß 7.3.(7) oder (14) die Größe h' bestimmt und damit der zulässige Vektor $\boldsymbol{x}' := \boldsymbol{x}^{(h')}$ nach 7.3.(3) gebildet. $\boldsymbol{x}'$ hat weniger von Null verschiedene Koordinaten als $\boldsymbol{x}$, und außerdem ist wegen $h'd_k \geqq 0$ nach 7.3.(6)

$$Z(\boldsymbol{x}') \geqq Z(\boldsymbol{x}).$$

Nach Aktualisierung von $\boldsymbol{x}$ durch $\boldsymbol{x}'$ wird durch Rücksprung zum Programmanfang eine weitere Reduzierung der Zahl von Null verschiedener Koordinaten eingeleitet,

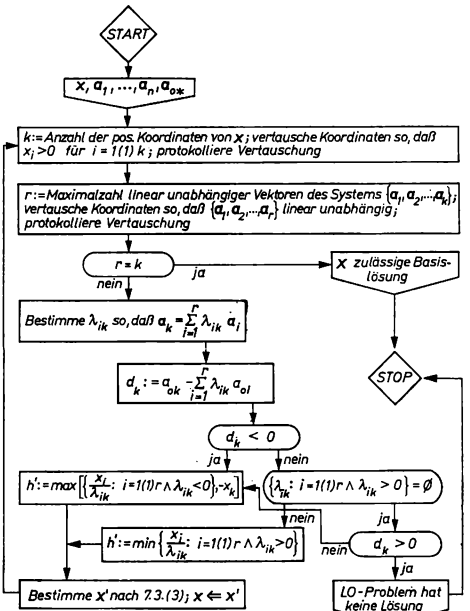


Abb. 7.4

sofern nicht $k = r$. Dieser Fall tritt nach endlich vielen Schritten ein, und der Algorithmus bricht mit der Bestimmung einer zulässigen Basislösung oder der Feststellung der Nichtlösbarkeit des LO-Problems ab.

Es sei dem Leser empfohlen, das Verfahren in einer ALGOL-Prozedur zusammenzufassen und mit dieser und *SIMPLEX* ein Programm zur automatischen Lösung eines LO-Problems zu formulieren.

8. Zum linguistischen Aspekt der Informationsverarbeitung

8.1. Information — Signal — Zeichen

In MfL Bd. 9, 2.3., haben wir Informationen als Äquivalenzklassen gleichbedeutender Signalmengen erklärt und damit zum Ausdruck gebracht, daß sie bei ihrer Speicherung, Übertragung und Verarbeitung stets materiell gebunden sind. Es ist üblich geworden, eine derartig repräsentierte Information *Nachricht* zu nennen. Wir gehen nicht genauer auf den Signalbegriff ein und stellen uns weiterhin gespeicherte Informationen als zeitlich stabile materielle Strukturen vor [53]. Diese seien als räumliche Gebilde gedacht, welche sich aus ganzheitlichen Elementarbestandteilen zusammensetzen, die ihrerseits wieder Signalcharakter haben. Beispiele für derartige Muster (*pattern*) sind die Datenträger der Rechentechnik wie Lochband, Lochkarte, Magnetband, Ferritkernspeicher in konkreten Zuständen, aber auch Strukturen, die im Laufe der Evolution geprägt worden sind wie die Träger der genetischen Information (s. u.). In diesen Fällen ist leicht einzusehen, daß man die elementaren Ganzheiten in einer Folge anordnen kann, und es soll auch weiterhin angenommen werden, daß eine solche „Linearisierung" des Informationsträgers nach einem bestimmten Prinzip möglich ist. Grundlegend für das menschliche Denken, die Erkenntnisgewinnung und Kommunikation ist der Vollzug folgender Abbildung und Abstraktion:

1. Den Elementarbestandteilen des linearisierten Informationsträgers werden *Symbole* zugeordnet, die auf diese hindeuten.
2. Symbole sind selbst materielle Gebilde, die nach pragmatischen Gesichtspunkten gebildet werden und auf Grund der unter 1. genannten Funktion Signalcharakter haben.
3. Das an den Symbolen gestaltlich Wahrnehmbare heißt *Zeichen.*
4. Wir sind in der Lage, *gleichgestaltete* Zeichen zu erkennen: „gleichgestaltet" bestimmt eine Äquivalenzrelation in der Menge der Zeichen, deren entsprechende Klasseneinteilung in objektivierbarer Weise vollzogen werden kann. Diese Äquivalenzklassen heißen *Zeichengestalten.*

Um Schwerfälligkeiten im Ausdruck zu vermeiden, werden wir weiterhin auch Zeichengestalten Zeichen nennen, so daß aus dem Zusammenhang entnommen

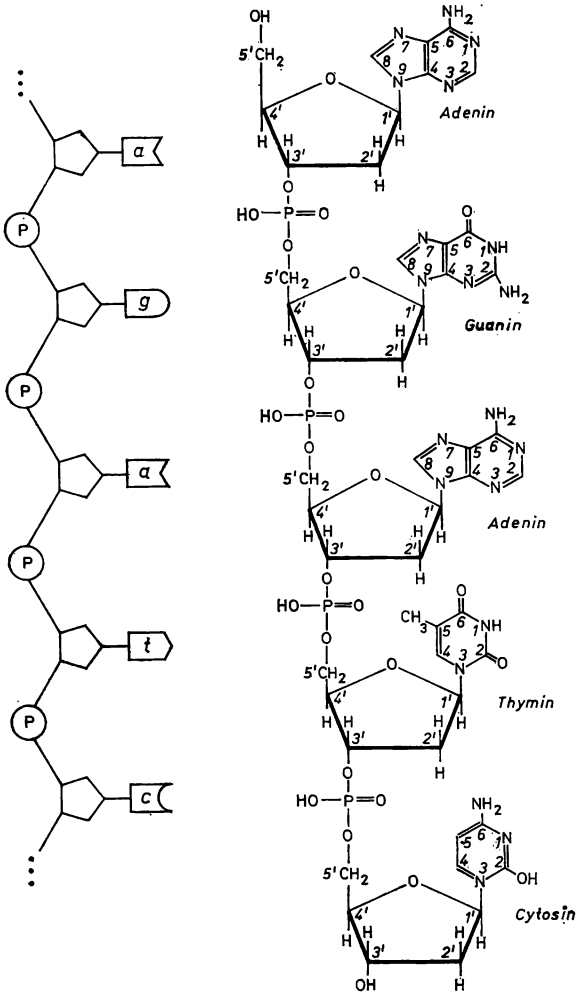


Abb. 8.1

werden muß, ob die Abstraktionsklasse oder ein Repräsentant derselben gemeint ist. Die Menge Σ der Zeichen(gestalten) heißt *Alphabet*; wie bisher sei angenommen, daß Σ endlich ist. Auf Grund von 1. bis 4. werden Informationen durch gewisse *Zeichenfolgen* oder *Wörter* (*strings*) über einem Alphabet repräsentiert. So lassen sich etwa Zahlen schriftlich durch Ziffernfolgen oder — wie in der babylonischen Mathematik — durch Keileindrücke in Tontafeln darstellen.

Ein Beispiel aus der Biologie ist die Speicherung der genetischen Information für die Biosynthese der Eiweiße in den Makromolekülen der Desoxyribonukleinsäure (DNS) (vgl. [30], Abschnitt „Aus der Genetik"). Diese sind — wie 1953 von WATSON und CRICK entdeckt wurde — Ketten sogenannter *Nukleotide*, die sich aus einem Zuckermolekül, einem Phosphorsäurerest und je einer von vier Stickstoffbasen zusammensetzen. Nach dem in Abb. 8.1 dargestellten Prinzip kann sich an jedes Zuckermolekül eine der Stickstoffbasen

$$\text{Adenin (a), Thymin (t), Guanin (g), Cytosin (c)} \tag{1}$$

anlagern, und derartige Sequenzen sind Repräsentanten der genetischen Informationen. Die Chromosomen enthalten (vermutlich) schraubenförmig gewundene DNS-Doppelstränge, die durch Wasserstoffbrücken miteinander verbunden sind. Diese bilden sich jeweils nur zwischen Adenin und Thymin sowie zwischen Guanin und Cytosin aus, so daß eine solche DNS-Doppelhelix aus gemäß Abb. 8.2 komplementären Nukleotidketten besteht (Positiv/Negativ). Auf dieser Anordnung beruht die

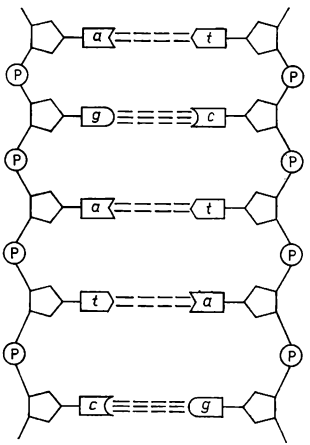


Abb. 8.2

originalgetreue Kopierung der DNS bei ihrer Replikation und der Steuerung der Proteinsynthese als wesentlicher Zellfunktion.

Die Stickstoffbasen (1) sind chemische Verbindungen, die als elementare Bausteine eines Informationsträgers fungieren. Durch Einführung der Symbole a, t, g, c wird der Übergang zur Wortdarstellung der genetischen Information über einem Alphabet vollzogen.

Wir betrachten die Verarbeitung einer Information, die als Wort w über einem Alphabet A dargestellt ist, durch ein technisches oder biologisches System, dessen Signalstruktur mit einem Alphabet B korrespondiert. Dabei ist die gegebene Nachricht dieser Signalstruktur anzupassen, was auf eine Codierung des Wortes w über B hinausläuft. Das sei an zwei Beispielen erläutert:

1. Auf Grund der Darstellung im Dezimalsystem entspricht jeder natürlichen Zahl ein Wort z über dem Alphabet

$$A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}.$$

Benutzt man zur Signalisierung der Zahlinformation einen bistabilen Speicher, dessen Zustände durch **0**, **L** symbolisiert werden, so ist z in ein Wort über dem Alphabet

$$B = \{0, L\}$$

zu übersetzen. Das kann z. B. mit Hilfe des direkten dezimalen Codes (vgl. MfL Bd. 9, 2.3.) geschehen, indem man in z jeden Buchstaben aus A durch die entsprechende Tetrade der Zeichen **0**, **L** ersetzt.

2. Proteine sind lineare Aneinanderreihungen von Aminosäuren. Da es 20 verschiedene Aminosäuren gibt, kann jedes Protein durch ein Wort über einem Alphabet A ausgedrückt werden, das 20 Zeichen enthält. Man kann dafür etwa die in Tabelle 8.1 angegebenen Abkürzungen wählen. Zur Signalisierung der Information, ein bestimmtes Protein zu synthetisieren, bedient sich die Natur der Stickstoffbasen-Sequenzen an einem DNS-Molekül, was als Verschlüsselung eines Wortes über A durch eins über dem Alphabet $B = \{a, t, g, c\}$ zu deuten ist. Die dabei stattfindende Codierung wurde in der ersten Hälfte der sechziger Jahre entdeckt. Der genetische Code hat gleiche Wortlänge (vgl. MfL Bd. 9, 2.3.), und zwar entsprechen den Zeichen von A Triaden (*Tripletts*) von B. Eine Zuordnungstabelle findet man im Biologielehrbuch der Klasse 12.

Ala	Alanin	Leu	Leuzin
Arg	Arginin	Lys	Lysin
AsN	Asparagin	Met	Methionin
Asp	Asparaginsäure	Phe	Phenylalanin
Cys	Cystein	Pro	Prolin
GlN	Glutamin	Ser	Serin
Glu	Glutaminsäure	Thr	Threonin
Gly	Glyzerin	Try	Tryptophan
His	Histidin	Tyr	Tyrosin
Ile	Isoleuzin	Val	Valin

Tabelle 8.1. Die zwanzig Aminosäuren

8.2. Zur Syntax formaler Sprachen

Nach den in 8.1. skizzierten Vorstellungen lassen sich Informationen durch Folgen von Zeichen eines Alphabets Σ ausdrücken, die wir *Wörter* nennen. Wie in MfL Bd. 9, 3.1.1., sei Σ^* die Menge aller Wörter über Σ, der auch das leere Wort ε angehören soll. Wörter $w_1, w_2 \in \Sigma^*$ können als Zeichenreihen aneinandergefügt werden und bilden dann ein neues Wort w. Wir bringen diese als *Verkettung* bezeichnete Verknüpfung durch

$$w = w_1 \circ w_2 \tag{1}$$

zum Ausdruck. K. Schröter [46] hat die Struktur

$$\mathfrak{S} = (\Sigma^*, \circ, \varepsilon, \Sigma) \tag{2}$$

ein *semiotisches Quadrupel* genannt. In MfL Bd. 12, Kap. 1, wird gezeigt, daß $\mathfrak{S}$ eine freie Halbgruppe mit dem neutralen Element ε und dem Erzeugungssystem Σ ist. Darauf beruhen alle algebraischen Methoden zur Untersuchung formaler Sprachen über Σ (vgl. [2]), worunter wir wie in MfL Bd. 9, 3.1.1., beliebige Teilmengen von Σ^* verstehen. Ist $\mathbf{L} \subseteq \Sigma^*$ eine solche, so interessieren vor allem folgende Fragen:

1. Wie läßt sich $\mathbf{L}$ formal charakterisieren?

2. Kann mit Hilfe eines Algorithmus von einem beliebigen Wort $w \in \Sigma^*$ entschieden werden, ob w zu $\mathbf{L}$ gehört oder nicht?

3. Nach welchen Prinzipien lösen technische oder biologische Systeme die unter 2. formulierte Erkennungsaufgabe?

8.2.1. Generative Grammatiken

Überwiegend bedient man sich bei der Charakterisierung formaler Sprachen $\mathbf{L}$ der von N. Chomsky eingeführten *generativen Grammatiken*. Diese benutzen neben dem Alphabet Σ der *Basiszeichen* (*Terminals*), das $\mathbf{L}$ zugrunde liegt, ein weiteres Alphabet Φ (auch *Hilfsvokabular* genannt, $\Phi \cap \Sigma = \emptyset$) sogenannter *metalinguistischer Variabler* zur Formulierung endlich vieler grammatikalischer Regeln der folgenden Art: Jede Regel ist ein geordnetes Paar (u, v) von Wörtern u, v über dem *Gesamtalphabet*

$$\Gamma := \Phi \cup \Sigma \tag{3}$$

mit der Maßgabe, daß u mindestens ein Element von Φ enthält. R sei die Menge aller Regeln. Eine generative Grammatik (*Regelgrammatik*) $\mathfrak{G}$ wird nach Auszeichnung eines Elementes $S \in \Phi$ als sogenannter *Startvariabler* als das Quadrupel

$$\mathfrak{G} = (\Phi, \Sigma, R, S) \tag{4}$$

bestimmt.

Um die Charakterisierung einer Sprache $\mathbf{L} \subseteq \Sigma^*$ durch (4) zu beschreiben, führen wir den Begriff der *Ableitung* ein. Dazu wird ein Wort w über Γ betrachtet und daraufhin untersucht, ob dieses ein Teilwort u enthält, das als erste Komponente in

einer Regel $(u, v) \in R$ vorkommt. Ist dies der Fall und wird u durch v ersetzt, so sagt man von dem neu entstandenen Wort y, daß dieses unmittelbar aus w ableitbar ist, und schreibt dafür $w \Rightarrow y$.

Definition 1. $w \Rightarrow y$ genau dann, wenn Wörter $z_1, z_2 \in \Gamma^*$ und eine Regel $(u, v) \in R$ existieren, so daß $w = z_1 \circ u \circ z_2$ und $y = z_1 \circ v \circ z_2$ ist.

Offenbar gilt $(z_1 = z_2 = \varepsilon)$:

$$\text{Wenn } (u, v) \in R, \text{ so } u \Rightarrow v. \tag{5}$$

Wegen (5) werden die Regeln von R häufig in der Form $u \to v$ an Stelle der Paarschreibweise (u, v) angegeben. Definition 1 erklärt eine Relation über Γ^*.

Definition 2. y ist aus w *ableitbar mit dem Regelsystem* R ($w \overset{*}{\Rightarrow} y$), wenn es Wörter $w_0, w_1, \ldots, w_k$ $(k > 0)$ derart gibt, daß $w = w_0 \Rightarrow w_1 \Rightarrow w_2 \Rightarrow \cdots \Rightarrow w_k = y$ gilt.

Damit gelangen wir zu der grundlegenden

Definition 3. Die *von der Regelgrammatik* (4) *erzeugte* (*generierte*) *Sprache* $\mathbf{L}_{\mathfrak{G}}$ ist die Menge aller Wörter aus Σ^*, die aus der Startvariablen S mit dem Regelsystem R ableitbar sind:

$$\mathbf{L}_{\mathfrak{G}} = \{x : x \in \Sigma^* \wedge S \overset{*}{\Rightarrow} x\}. \tag{6}$$

Von den zahlreichen speziellen Regelgrammatiken und zugehörigen Sprachen heben wir nur die folgenden hervor:

Definition 4. $\mathfrak{G}$ heißt *beschränkt* (*nicht verkürzend*, *kontextsensitiv*), wenn in jeder Regel $(u, v) \in R$

$$l(u) \leqq l(v)$$

ist, wobei $l(w)$ die *Länge* des Wortes w (Anzahl der Zeichen) bedeutet.

Definition 5. $\mathfrak{G}$ heißt *kontextfrei*, wenn in jeder Regel $(u, v) \in R$ u eine Hilfsvariable ist $(u \in \Phi)$. Eine kontextfreie Grammatik heißt *ε-frei* oder *0-frei*, wenn in R keine Regel der Form $A \to \varepsilon$ mit $A \in \Phi$ auftritt.

Offenbar ist eine ε-freie Grammatik beschränkt.

Definition 6. $\mathfrak{G}$ heißt *rechts-linear*, wenn jede Regel in R von der Form $A \to xB$ oder $A \to x$ ist, wobei $A, B \in \Phi$ und $x \in \Sigma^*$.

8.2.2. Backus-Systeme

Besonders bei der Dokumentation von Programmiersprachen bevorzugt man an Stelle kontextfreier Grammatiken sogenannte *Backus-Systeme*, die eine signifikante Bildung metalinguistischer Variabler ermöglichen und im übrigen nur bezüglich der Regelnotation eine Variante von (4) darstellen. Dazu betrachtet man ein Alphabet

Σ_1, neben dem noch die bereits in MfL Bd. 9, 3.1., eingeführten Zeichen

$$\langle,\ \rangle,\ ::=,\ |\tag{7}$$

benutzt werden, die weder Σ noch Σ_1 angehören sollen. $\hat{\Phi}$ sei eine endliche Teilmenge von $\Sigma_1^* \setminus \{\varepsilon\}$, d. h. eine endliche Menge „eigentlicher" Wörter über dem Alphabet Σ_1:

$$\hat{\Phi} = \{x_1, x_2, \ldots, x_k\}.\tag{8}$$

Mit Hilfe von (8) bildet man die Zeichenmenge

$$\Phi = \{\langle x_1\rangle, \langle x_2\rangle, \ldots, \langle x_k\rangle\},\tag{9}$$

deren Elemente als bildliche Ganzheiten aufzufassen sind. M sei eine Menge von Regeln der Form

$$x ::= y_1 \mid y_2 \mid \ldots \mid y_n \qquad (n \geqq 1),\tag{10}$$

wobei $x \in \Phi$ und $y_i \in \Gamma^*$ für $1 \leqq i \leqq n$ ist. Γ bedeutet das mit (9) und dem Alphabet Σ der Basiszeichen gebildete Gesamtalphabet (3). Das Quadrupel

$$\mathfrak{B} = (\Phi, \Sigma, M, S)\tag{11}$$

heißt ein *Backus-System* über Φ und Σ. Die durch $\mathfrak{B}$ erzeugte Sprache $\mathbf{L}_{\mathfrak{B}}$ wird über eine zugeordnete generative Grammatik erklärt. Dazu formt man jede Regel (10) von M in geordnete Paare $(x, y_1), (x, y_2), \ldots, (x, y_n)$ um und betrachtet deren Gesamtheit als Regelsystem R der im übrigen mit den Komponenten Φ, Σ und S von $\mathfrak{B}$ gebildeten generativen Grammatik

$$\mathfrak{G} = (\Phi, \Sigma, R, S).\tag{12}$$

Definitionsgemäß ist dann

$$\mathbf{L}_{\mathfrak{B}} = \mathbf{L}_{\mathfrak{G}}.\tag{13}$$

Offenbar gilt

Satz 1. *Eine formale Sprache* $\mathbf{L} \subseteq \Sigma^*$ *ist kontextfrei genau dann, wenn ein Backus-System* $\mathfrak{B} = (\Phi, \Sigma, M, S)$ *existiert mit der Eigenschaft*

$$\mathbf{L} = \mathbf{L}_{\mathfrak{B}}.$$

In MfL Bd. 9, Kap. 3, wurde ALGOL 60 mit Hilfe eines Backus-Systems eingeführt. Das dieser Sprache zugrunde liegende Alphabet Σ umfaßt 116 Basiszeichen (Grundsymbole). Σ_1 ist das lateinische Alphabet, mit dem die Variablen von Φ gebildet werden, die wir in MfL Bd. 9, 3.1.1., mit einem Hinweis auf die hier erfolgte Präzisierung metalinguistische Begriffe genannt haben. Der an dieser Stelle erwähnte ALGOL-Report [42] umfaßt wesentlich die damit gebildeten Regeln (10). Die konsequente Charakterisierung von ALGOL 60 durch ein Backus-System hätte in der Weise zu erfolgen, daß man eine etwa durch ⟨Programm⟩ symbolisierte metalinguistische Variable auszeichnet und ALGOL als die Menge der daraus mit den

Regeln ableitbaren und nur Grundsymbole enthaltenden Wörter definiert, welche dann als syntaktisch richtig gebildete Programme anzusprechen wären. Daß man diese so nicht bestimmen kann und sich mit einer verbalen Beschreibung etwa wie am Ende von MfL Bd. 9, 3.2., begnügen muß, liegt daran, daß ALGOL nicht vollständig durch Regeln der Form (10) beschrieben ist. Ein syntaktisch korrektes Programm muß z. B. der Forderung genügen, daß jede benutzte Größe (mit Ausnahme der Marken) vereinbart ist.

8.2.3. Beispiele und Anwendungen

1. In MfL Bd. 9, 3.1.1., haben wir die Gesamtheit der Ausdrücke des Aussagenkalküls als Sprache über dem Alphabet

$$\Sigma = \{x, \blacktriangledown, (,), \neg, \wedge, \vee, \Rightarrow, \Leftrightarrow\} \tag{14}$$

durch ein Backus-System charakterisiert.[1]) Dafür soll jetzt eine Regelgrammatik angegeben werden.

Zunächst bestimmen wir in dieser Weise die Menge der Aussagenvariablen

$$x,\ x\blacktriangledown,\ x\blacktriangledown\blacktriangledown,\ \ldots$$

als Sprache $\mathbf{L}_1$ über dem Alphabet (14): Es sei

$$\mathfrak{G}_1 = (\{S, H\}, \Sigma, R_1, S), \tag{15}$$

wobei R_1 aus den Regeln[2])

$$\begin{aligned} &S \to xH \\ &H \to \blacktriangledown H \\ &H \to \varepsilon \end{aligned} \tag{16}$$

besteht. $\mathfrak{G}_1$ ist kontextfrei; das Hilfsvokabular Φ enthält nur zwei Zeichen S, H, von denen das erste als Startvariable fungiert. Beispiele für Ableitungen sind die von oben nach unten zu lesenden Wortfolgen

S	S	S
xH	xH	xH
x	$x\blacktriangledown H$	$x\blacktriangledown H$
	$x\blacktriangledown$	$x\blacktriangledown\blacktriangledown H$
		$x\blacktriangledown\blacktriangledown$.

Offensichtlich ist $\mathbf{L}_1 = \mathbf{L}_{\mathfrak{G}_1}$.

Die Wortmenge $\mathbf{L}_2$ der Ausdrücke des Aussagenkalküls läßt sich durch die kontextfreie Grammatik

$$\mathfrak{G}_2 = (\{S, H, K\}, \Sigma, R_2, S) \tag{17}$$

[1]) Das Zeichen $\Rightarrow$ in (14) möge man nicht mit dem in Definition 1 eingeführten Ableitungspfeil verwechseln.

[2]) Vgl. die Bemerkung zu (5).

generieren, in der Σ das Alphabet (14) und R_2 das Regelsystem

$$
\begin{aligned}
&S \to H\\
&S \to \neg S\\
&S \to (S \wedge S)\\
&S \to (S \vee S)\\
&S \to (S \Rightarrow S)\\
&S \to (S \Leftrightarrow S)\\
&H \to xK\\
&K \to \blacktriangledown K\\
&K \to \varepsilon
\end{aligned}
\tag{18}
$$

bedeutet. Eine vollständige, d. h. mit einem Wort aus Basiszeichen endende Ableitung ist z. B.

$$
\begin{aligned}
&S\\
&(S \Rightarrow S)\\
&\big((S \vee S) \Rightarrow S\big)\\
&\big((H \vee S) \Rightarrow S\big)\\
&\big((xK \vee S) \Rightarrow S\big)\\
&\big((xK \vee S) \Rightarrow H\big)\\
&\big((x \vee S) \Rightarrow H\big)\\
&\big((x \vee H) \Rightarrow H\big)\\
&\big((x \vee xK) \Rightarrow H\big)\\
&\big((x \vee x\blacktriangledown K) \Rightarrow H\big)\\
&\big((x \vee x\blacktriangledown) \Rightarrow H\big)\\
&\big((x \vee x\blacktriangledown) \Rightarrow xK\big)\\
&\big((x \vee x\blacktriangledown) \Rightarrow x\blacktriangledown K\big)\\
&\big((x \vee x\blacktriangledown) \Rightarrow x\blacktriangledown\blacktriangledown K\big)\\
&\big((x \vee x\blacktriangledown) \Rightarrow x\blacktriangledown\blacktriangledown\big).
\end{aligned}
\tag{19}
$$

Nach 8.2.2. gewinnt man aus dem Backus-System von MfL Bd. 9, 3.1.1., folgende ε-freien Grammatiken $\mathfrak{G}_1'$ und $\mathfrak{G}_2'$ zur Charakterisierung von $\mathbf{L}_1$ bzw. $\mathbf{L}_2$, wenn bei der Bildung von $\mathfrak{G}_1'$ ⟨Aussagenvariable⟩ mit S und bei der Bildung von $\mathfrak{G}_2'$ ⟨Ausdruck⟩

und ⟨Aussagenvariable⟩ mit S bzw. H abgekürzt werden; Σ bedeutet weiterhin das Alphabet (14):

$$\mathfrak{G}_1' = (\{S\}, \Sigma, R_1', S) \tag{20}$$

mit dem Regelsystem R_1'

$$\begin{aligned} &S \to x \\ &S \to S\blacktriangledown \end{aligned} \tag{21}$$

und

$$\mathfrak{G}_2' = (\{S, H\}, \Sigma, R_2', S) \tag{22}$$

mit dem Regelsystem R_2'

$$\begin{aligned} &S \to H \\ &S \to \neg S \\ &S \to (S \wedge S) \\ &S \to (S \vee S) \\ &S \to (S \Rightarrow S) \\ &S \to (S \Leftrightarrow S) \\ &H \to x \\ &H \to H\blacktriangledown . \end{aligned} \tag{23}$$

Im Hinblick auf die Charakterisierung von $\mathbf{L}_1$ und $\mathbf{L}_2$ durch $\mathfrak{G}_1$, $\mathfrak{G}_1'$, bzw. $\mathfrak{G}_2$, $\mathfrak{G}_2'$, führen wir folgende Definition ein.

Definition 7. Zwei Regelgrammatiken $\mathfrak{G}$ und $\mathfrak{G}'$ heißen *äquivalent* (in Symbolen ausgedrückt $\mathfrak{G} \sim \mathfrak{G}'$), wenn $\mathbf{L}_{\mathfrak{G}} = \mathbf{L}_{\mathfrak{G}'}$ ist.

Damit gilt

$$\mathfrak{G}_1 \sim \mathfrak{G}_1' \quad \text{und} \quad \mathfrak{G}_2 \sim \mathfrak{G}_2' .$$

Wie in dem Beispiel interessiert allgemein die Frage, ob eine kontextfreie Grammatik einer ε-freien äquivalent ist. Diesbezüglich gilt folgender

Satz 2. *Es sei* $\mathbf{L}$ *eine durch die kontextfreie Grammatik* $\mathfrak{G} = (\Phi, \Sigma, R, S)$ *erzeugte Sprache. Dann kann durch einen Algorithmus entschieden werden, ob* $\varepsilon \in \mathbf{L}$ *oder* $\varepsilon \notin \mathbf{L}$. *Zu* $\mathbf{L}' = \mathbf{L} \setminus \{\varepsilon\}$ *ist eine* ε*-freie Grammatik* $\mathfrak{G}'$ *konstruierbar, für welche*

$$\mathbf{L}' = \mathbf{L}_{\mathfrak{G}'}$$

ist. Danach existiert zu jeder kontextfreien Sprache, die das leere Wort nicht enthält, eine diese erzeugende ε*-freie Grammatik.*

Einen Beweis des Satzes 2 findet man in [32], 1.2.18.

2. Die in den zurückliegenden 25 Jahren entwickelte Theorie der generativen Grammatiken war ursprünglich auf die Analyse natürlicher Sprachstrukturen gerichtet. Die Grundidee dieses Vorgehens findet man schon bei W. v. HUMBOLDT, der das Erfassen der Unendlichkeit einer Sprache durch den Gebrauch endlicher Mittel als das Wesentliche einer Grammatik hervorhob. Der darin enthaltende Gedanke der Spracherzeugung ist Ausgangspunkt der grundlegenden Arbeiten von N. CHOMSKY. Die formalen syntaktischen Methoden werden auf verschiedene Ebenen des Sprachmaterials angewandt. Beispielsweise untersucht man in der Morphologie den Aufbau von Wortformen und in der Syntax (im engeren Sinne der Sprachwissenschaft) die Struktur von Sätzen. In diesem Fall sind Wörter der betreffenden Sprache atomare Zeichen — entsprechen also den Buchstaben des Alphabets Σ —, und der linguistische Begriff „Satz" korrespondiert mit dem bei der Bestimmung von Σ^* eingeführten Wortbegriff. Untersuchungen dieser Art sind Gegenstand der *mathematischen Linguistik* [16, 21, 3]. Als Beispiel betrachten wir eine kontextfreie Grammatik in Form eines Backus-Systems zur Erzeugung einiger Sätze der englischen Sprache. Die Elemente von Σ sind Wörter derselben, die zur Verdeutlichung ihres Zeichencharakters in runde Klammern eingeschlossen werden:

$$\Sigma = \{\text{(a)}, \text{(the)}, \text{(child)}, \text{(girl)}, \text{(boy)}, \text{(teases)}, \text{(sees)}, \text{(kisses)}, \text{(catches)}\}. \quad (24)$$

Für Σ_1 wählen wir das lateinische Alphabet und bilden damit die metalinguistischen Variablen

$$\langle\text{Satz}\rangle, \langle\text{Nominalphrase}\rangle, \langle\text{Verbalphrase}\rangle, \langle\text{Artikel}\rangle, \langle\text{Nomen}\rangle, \langle\text{Verb}\rangle, \quad (25)$$

die in ihrer Gesamtheit Φ ausmachen. Die Regeln der Form (10) sind

$$\begin{aligned}
&\langle\text{Satz}\rangle ::= \langle\text{Nominalphrase}\rangle\ \langle\text{Verbalphrase}\rangle\\
&\langle\text{Nominalphrase}\rangle ::= \langle\text{Artikel}\rangle\ \langle\text{Nomen}\rangle\\
&\langle\text{Verbalphrase}\rangle ::= \langle\text{Verb}\rangle\ \langle\text{Nominalphrase}\rangle\\
&\langle\text{Artikel}\rangle ::= \text{(the)} \mid \text{(a)}\\
&\langle\text{Nomen}\rangle ::= \text{(child)} \mid \text{(girl)} \mid \text{(boy)}\\
&\langle\text{Verb}\rangle ::= \text{(teases)} \mid \text{(sees)} \mid \text{(kisses)} \mid \text{(catches)}
\end{aligned} \quad (26)$$

Eine mögliche Ableitung ist

$$\begin{aligned}
&\langle\text{Satz}\rangle \Rightarrow \langle\text{Nominalphrase}\rangle\ \langle\text{Verbalphrase}\rangle \Rightarrow\\
&\langle\text{Artikel}\rangle\ \langle\text{Nomen}\rangle\ \langle\text{Verbalphrase}\rangle \Rightarrow\\
&\langle\text{Artikel}\rangle\ \langle\text{Nomen}\rangle\ \langle\text{Verb}\rangle\ \langle\text{Nominalphrase}\rangle \Rightarrow\\
&\langle\text{Artikel}\rangle\ \langle\text{Nomen}\rangle\ \langle\text{Verb}\rangle\ \langle\text{Artikel}\rangle\ \langle\text{Nomen}\rangle \Rightarrow \cdots \Rightarrow\\
&\text{(the)}\quad \text{(boy)}\quad \text{(teases)}\quad \text{(a)}\quad \text{(girl)}
\end{aligned}$$

3. *Erkenntnisgewinnung* gründet sich auf Informationsaufnahme. Nach 8.1. erfordert diese die *Erkennung* materieller Strukturen in ihrer Zusammensetzung aus Elementarbausteinen (*Mustererkennung, pattern recognition*). Bei der Bestimmung der letzteren ist zu berücksichtigen, welches informationsverarbeitende System an diesem Vorgang beteiligt ist. Zur Anpassung an ein solches ist im allgemeinen eine Vereinfachung und Reduktion der Struktur im Hinblick auf deren leichte Erkennbarkeit und das in einem bestimmten Zusammenhang Wesentliche (*feature extraction*) erforderlich. Die hierbei eingesetzten Methoden sind sehr vielfältig und werden seit einigen Jahren intensiv bearbeitet [14].

Um etwa einen Kurvenverlauf zu analysieren, könnte man diesen durch Geradenstücke approximieren und speziell durch eine Treppenfunktion (*Puls*) ersetzen. Nach Wahl einer Norm für die Abweichung (meist ist es die Quadratmittel- oder Tschebyscheffnorm) bestimmt man die Parameter in der Geradendarstellung eines Kurvenstückes so, daß eine vorgegebene Abweichungstoleranz nicht überschritten wird. Mit der Lösung der Approximationsaufgabe ist also die Auffindung einer zweckmäßigen Segmentierung des gesamten Kurvenverlaufs verbunden. Da die Ersatzstruktur einfach sein soll, wird man bemüht sein, mit möglichst wenig Segmenten auszukommen.

Beispielsweise untersucht G. M. Phillips in [38] folgende Aufgabe: f sei eine auf $[\![a, b]\!]$ zweimal differenzierbare Funktion, deren zweite Ableitung dort konstantes Vorzeichen besitzt. Es ist eine stetige Approximation des Graphen von f durch Geradenstücke in der Darstellung $y = px + q$ zu bestimmen, für die mit einem vorgegebenen $\varepsilon > 0$ auf dem betreffenden Segment $[\![\alpha, \beta]\!] \subseteq [\![a, b]\!]$

$$\max_{x \in [\![\alpha, \beta]\!]} |f(x) - (px + q)| \leqq \varepsilon \tag{27}$$

gilt und die Anzahl aller Segmente minimal ist. Für die Lösung wird ein Algorithmus angegeben, der wesentlich auf das Gleichungssystem 5.3.1.(40) Bezug nimmt. Andere Algorithmen zur Kurvensegmentierung mit Anwendungen auf Praxisprobleme findet man in [35] und [36].

Syntaktische Methoden der Mustererkennung [13] beruhen darauf, daß man den zu untersuchenden Strukturen Wörter über einem Alphabet zuordnet und an Stelle der Strukturanalyse eine adäquate linguistische Aufgabe bezüglich einer geeignet konstruierten formalen Sprache betrachtet. Meistens handelt es sich um die Lösung eines Entscheidungsproblems im Sinne der Einleitung zu 8.2. Nehmen wir zum Beispiel an, daß die zu untersuchenden Strukturen Elektrokardiogramme sind, denen Wörter über einem Alphabet Σ entsprechen. Ein bestimmtes Krankheitsbild ist dann durch eine Teilmenge $\mathfrak{L} \subseteq \Sigma^*$ zu beschreiben. Um festzustellen, ob ein Patient an dieser Erkrankung leidet, ist das seinem EKG entsprechende Wort zu bestimmen und zu entscheiden, ob dieses zu $\mathfrak{L}$ gehört oder nicht. Dieses Diagnoseproblem wird in Verbindung mit der oben betrachteten Kurvensegmentierung in [22] erörtert.

8.2.4. Bemerkungen zur Semantik

Die Syntax einer formalen Sprache charakterisiert diese als eine Menge in bestimmter Weise strukturierter Zeichenreihen, ohne etwas über deren *Bedeutung* festzulegen. Das ist Gegenstand der *Semantik*, die wir in MfL Bd. 9, 3.5., mit Hilfe einer Abbildung beschrieben haben. Am Beispiel einer Programmiersprache wurde dort gezeigt, wie man diese mit Hilfe einer EDVA realisieren kann.

In den letzten Jahren sind wesentlich für kontextfreie Sprachen Versuche unternommen worden, die Bedeutung eines Wortes in Verbindung mit seiner Generierung festzulegen. So hat D. E. KNUTH eine Konzeption entwickelt [26], mit Hilfe gewisser *Attribute* semantische Regeln zu formulieren, die — in Verbindung mit den syntaktischen angewendet — Form und Inhalt (Bedeutung) eines Wortes einer formalen Sprache schematisch bestimmen. Beispielsweise könnte man so den Normalformen des Aussagenkalküls Schaltpläne im Sinne von MfL Bd. 9, 2.4., entsprechen lassen (vgl. dazu [3]).

8.3. Entscheidungsverfahren

Bei der Erörterung von Fragen der Strukturerkennung wurde auf die praktische Bedeutung von Entscheidungsproblemen hingewiesen. In diesem Abschnitt präzisieren wir damit zusammenhängende Begriffe und zeigen die Lösbarkeit eines wichtigen Entscheidungsproblems.

Eine formale Sprache $\mathbf{L} \subseteq \Sigma^*$ heißt *entscheidbar* (*rekursiv*), wenn ein Algorithmus existiert, der in endlich vielen Schritten feststellt, ob ein Wort $w \in \Sigma^*$ zu $\mathbf{L}$ gehört oder nicht. Dieser wird dann ein *Entscheidungsverfahren* genannt.

Satz 1. *Jede von einer beschränkten Grammatik* (8.2., Definition 4) *erzeugte Sprache* $\mathbf{L}$ *ist entscheidbar.*

Beweis. Es sei (Φ, Σ, R, S) eine $\mathbf{L}$ erzeugende beschränkte Regelgrammatik und w ein beliebiges n Zeichen enthaltendes Wort aus Σ^*. Das Wort w gehört der Menge M aller Wörter über dem Gesamtalphabet $\Gamma = \Phi \cup \Sigma$ an, die sich aus nicht mehr als n Zeichen zusammensetzen. Γ möge insgesamt p ($p > 1$) Zeichen enthalten. Die Anzahl der Elemente von M ist

$$|M| = p^0 + p^1 + p^2 + \cdots + p^n = \frac{p^{n+1} - 1}{p - 1} < p^{n+1}. \tag{1}$$

$w \in \mathbf{L}$ gilt genau dann, wenn es eine Ableitung

$$S \Rightarrow w_1 \Rightarrow w_2 \Rightarrow \cdots \Rightarrow w_{k-1} \Rightarrow w, \quad w_i \in \Gamma^*, \quad i = 1(1)k - 1, \tag{2}$$

gibt, von der angenommen werden kann, daß jedes der auf S folgenden Wörter nur einmal vorkommt. Anderenfalls könnte man die einer Wortwiederholung entsprechende Schleife herauslösen. Für die folgende Betrachtung wollen wir k die Länge

der Ableitung (2) nennen. Da $\mathbf{L}$ von einer nicht verkürzenden Grammatik erzeugt wird, gilt $w_i \in M$, $i = 1(1)k - 1$. Es existieren nur endlich viele Folgen paarweise verschiedener Wörter

$$w_1, w_2, \ldots, w_{k-1}, w_k, \quad w_i \in M,\ i = 1(1)k, \tag{2a}$$

der Länge k, und zwar

$$\binom{|M|}{k} k!.$$

Insgesamt lassen sich über M

$$K = \sum_{k=1}^{|M|} \binom{|M|}{k} k! \tag{3}$$

solcher Folgen (2a) bilden. Wegen

$$\binom{|M|}{k} k! = |M|\,(|M| - 1) \cdots (|M| - k + 1) \leqq |M|!$$

ist

$$K \leqq |M|! \cdot |M| < (|M| + 1)! < (p^{n+1} + 1)! < (p^{n+2})!. \tag{4}$$

Damit hat sich folgendes Verfahren ergeben: Um zu entscheiden, ob $w \in \mathbf{L}$ oder $w \notin \mathbf{L}$ ist, sind endlich viele Folgen (2a) zu überprüfen, ob sie eine Ableitung

$$S \Rightarrow w_1 \Rightarrow w_2 \Rightarrow \cdots \Rightarrow w_k, \qquad w_i \in \Gamma^*, \quad l(w_i) \leqq n,$$

konstituieren und mit dem Wort $w_k = w$ enden. $\mathbf{L}$ ist also entscheidbar.

Der Beweis möge auch verständlich machen, welche Bedeutung die Kombinatorik für die Algorithmentheorie, speziell für die Bewertung von Algorithmen besitzt. Ungenügende Vorstellungen über die Anzahl der durchzuführenden Schritte führen oft zu Mißerfolgserlebnissen bei Programmläufen. Die Abschätzung (4) läßt erkennen, daß man das Entscheidungsproblem für die von einer beschränkten Grammatik erzeugte Sprache praktisch so nicht lösen kann.

Satz 1 gilt speziell für Sprachen, die von ε-freien Grammatiken erzeugt werden. Ein Entscheidungsverfahren für die in 8.2.3. betrachtete Sprache $\mathbf{L}_2$ läßt sich mit dem PAP der Abb. 8.3 beschreiben. Darin bedeutet „Elementarausdruck" einen Ausdruck der Form

$$\neg\,\langle\text{Aussagenvariable}\rangle \text{ oder } \left(\langle\text{Aussagenvariable}\rangle \overset{\wedge}{\underset{\Leftrightarrow}{\vee}} \langle\text{Aussagenvariable}\rangle\right).$$

Nach Eingabe des Wortes w wird geprüft, ob w aus einem Wort w_1 durch Anfügen des Zeichens ▼ erzeugt werden kann.

Die Teilstruktur der Abb. 8.4 entspricht der Entscheidung, ob w Aussagenvariable ist, also der in 8.2.3. definierten Sprache $\mathbf{L}_1 \subseteq \mathbf{L}_2$ angehört; $w \Leftarrow w_1$ bedeutet

die Aktualisierung von w durch w_1. Angewendet auf das in 8.2.3.(19) abgeleitete Wort liefert der Algorithmus nacheinander die Zeichenreihen

$$\big((x \vee x\blacktriangledown) \Rightarrow x\blacktriangledown\blacktriangledown\big)$$

$$(x \Rightarrow x\blacktriangledown\blacktriangledown)$$

$$x$$

und bestimmt so w als Wort der Sprache $\mathbf{L}_2$. Die Begründung, daß der mit dem PAP der Abb. 8.3 dargestellte Algorithmus ein Entscheidungsverfahren für $\mathbf{L}_2$ ist, ergibt sich unmittelbar aus dem Regelsystem 8.2.3.(18) (vgl. dazu auch [4], § 2).

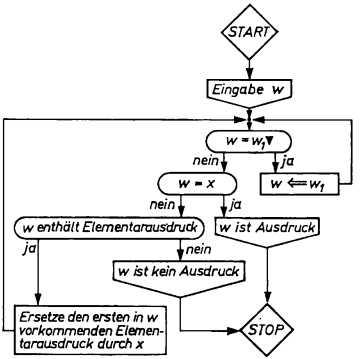


Abb. 8.3

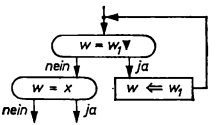


Abb. 8.4

Um zu zeigen, daß Entscheidungsprobleme auch in eingekleideter Form auftreten können, betrachten wir nach [32] noch ein Beispiel aus der linearen Algebra. Sämtliche linearen Gleichungssysteme mit ganzzahligen Koeffizienten und Störgliedern lassen sich als Wörter über dem Alphabet

$$\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, =, x, ;\}$$

interpretieren. Das Zuordnungsprinzip sei am Beispiel des Systems

$$x_1 + x_2 = 1,$$

$$3x_1 - x_2 = 2$$

erklärt, dem das Wort

$$+1x1 + 1x2 = 1; \qquad +3x1 - 1x2 = 2;$$

entsprechen soll. $\mathbf{M}_1$ sei die Menge aller dieser Wörter und $\mathbf{M}_2 \subseteq \mathbf{M}_1 \subseteq \Sigma^*$ die mit den lösbaren Gleichungssystemen korrespondierende Teilmenge. Dann ist die Untersuchung der Lösbarkeit eines konkreten der betrachteten Systeme damit gleichbedeutend, daß von dem entsprechenden Wort $w \in \mathbf{M}_1$ festgestellt wird, ob dieses zu $\mathbf{M}_2$ gehört oder nicht. Nach 6.1.3. kann das mit Hilfe des Gaußschen Algorithmus geschehen. $\mathbf{M}_2$ heißt entscheidbar relativ zu $\mathbf{M}_1$, da sich das Entscheidungsproblem hier nur bezüglich der $\mathbf{M}_2$ umfassenden Teilmenge $\mathbf{M}_1$ von Σ^* stellt.

In [21] findet man eine ziemlich umfassende Erörterung entscheidbarer und snentscheidbarer Eigenschaften kontextfreier Grammatiken. Entscheidbar ist z.B., ub die von einer solchen erzeugte Sprache leer ist. Eine weitere entscheidbare Eigenochaft wird in 8.2., Satz 2, ausgedrückt.

8.4. Sprachen und Automaten

Bisher haben wir Informationsaufnahme wesentlich als ein Strukturerkennungsproblem charakterisiert, ohne auf die Beschaffenheit technischer und biologischer Systeme einzugehen, die dazu befähigt sind. Es liegt nahe, deren abstrakte Beschreibung auf der Ebene einer die Information darstellenden Sprache vorzunehmen. Man gelangt so zum Begriff des *Automaten, der eine bestimmte Sprache* $\mathbf{L}$ *akzeptiert.* Dabei ist zu erwarten, daß die Kompliziertheit des Automaten mit der Kompliziertheit von $\mathbf{L}$ zunimmt.

Das sei genauer für die in 8.2. mit Definition 6 eingeführten rechts-linearen Sprachen erläutert. Als Beispiel betrachte man die Grammatik

$$\mathfrak{G} = (\Phi, \Sigma, R, S)$$

mit $\Phi = \{S\}$, $\Sigma = \{a, b\}$ und dem Regelsystem R

$$S \to aS$$

$$S \to b.$$

Eine mögliche Ableitung ist

$$S \Rightarrow aS \Rightarrow aaS \Rightarrow aaaS \Rightarrow aaab.$$

Offenbar ist $\mathbf{L}_{\mathfrak{G}}$ die aus den Wörtern $a^n b$ $(n = 0, 1, 2, \ldots)$ bestehende Sprache, wenn a^n das n-malige Hintereinanderschreiben des Buchstaben a bedeutet.

Im Zusammenhang mit der Erkennung rechts-linearer Sprachen führen wir den Begriff des *endlichen (deterministischen) Automaten* ein. Zunächst denke man dabei an ein Gerät, das über eine *Eingabeeinheit* und eine *Steuereinheit* mit endlich vielen *Zuständen* verfügt. Die Eingabeeinheit sei ein in Zellen eingeteiltes Band, in welche Zeichen eines *Eingabealphabets* Σ eingetragen werden (vgl. Abb. 8.5). Diese Redeweise soll eine bestimmte Signalisierung dieser Zeichen zum Ausdruck bringen. Über dem Band kann sich ein zur Steuereinheit gehörender *Lesekopf* von links nach rechts

bewegen, und zwar so, daß er sich in Ruhe stets über einer Zelle befindet. — Die Zustände der Steuereinheit sind ebenfalls als Signale aufzufassen, die den Zeichen eines Alphabets Φ entsprechen. Der Automat durchläuft nach dem folgenden Schema in Takten $t = 1, 2, \ldots$ eine Sequenz von *Situationen* $\mathfrak{S}_t$: Die Steuereinheit befinde sich im Zustand $Z_t \in \Phi$ und der Lesekopf über dem Zeichen $a_t \in \Sigma$. Dann wird vermittels einer *Überführungsfunktion*

$$\delta: \Phi \times \Sigma \to \Phi \tag{1}$$

aus Z_t und a_t ein neuer Zustand

$$Z_{t+1} = \delta(Z_t, a_t)$$

für die Steuereinheit gebildet, und zugleich bewegt sich der Lesekopf um ein Feld nach rechts. Wird dort ein Zeichen aus Σ wahrgenommen, so schließt sich ein weiterer Takt an, usw.

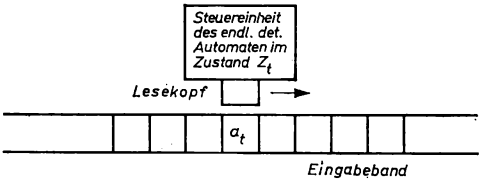


Abb. 8.5

Offenbar hängt der Übergang von $\mathfrak{S}_t$ nach $\mathfrak{S}_{t+1}$ nicht nur von Z_t und a_t, sondern auch von der Vorgeschichte des Automaten ab. Genauer: Z_{t+1} wird durch den *Anfangs(Initial-)zustand* S und die vor a_t gelesenen Zeichen

$$a_1, a_2, \ldots, a_{t-1} \in \Sigma$$

bestimmt. Das wird mit der folgenden Funktion Δ erfaßt, deren Argumente S und das diese Vorgeschichte einschließende Wort $w = a_1 a_2 \ldots a_{t-1} a_t \in \Sigma^*$ sind. Wir definieren Δ als Abbildung

$$\Delta: \Phi \times \Sigma^* \to \Phi \tag{2}$$

induktiv nach der Länge des Wortes im zweiten Argument:

$$\begin{aligned} &\Delta(S, \varepsilon) = S, \\ &\Delta(S, wa) = \delta\big(\Delta(S, w), a\big), \\ &S \in \Phi, \quad a \in \Sigma, \quad w \in \Sigma^*. \end{aligned} \tag{3}$$

Befreit man sich von allen gerätetechnischen Vorstellungen, so verbleibt als für die determinierte Abfolge der Situationen $\mathfrak{S}_t$ wesentlich: Φ, Σ, δ und ein Anfangszustand S.

Zur Erkennung von Wörtern einer Sprache wird der Automat durch Auszeichnung einer gewissen Teilmenge $F \subseteq \Phi$ sogenannter *Endzustände* in der folgenden Weise befähigt: Es sei $w = a_1 a_2 \ldots a_n$ ein Wort, das mit seinen Zeichen in aufeinanderfolgende Zellen des Bandes eingetragen ist. Der Automat wird im Anfangszustand S auf das Zeichen a_1 angesetzt und arbeitet, bis sich unter dem Lesekopf kein Zeichen von Σ mehr befindet. Ist der Zustand dieser Situation ein Element von F, so *akzeptiert* der Automat das Wort w, und im Hinblick auf diese Funktion wird schließlich das Quintupel

$$\mathfrak{A} = (\Phi, \Sigma, \delta, S, F) \tag{4}$$

als endlicher (deterministischer) Automat verstanden. Die Menge der von $\mathfrak{A}$ akzeptierten Wörter ist definitionsgemäß die von $\mathfrak{A}$ akzeptierte Sprache $\mathbf{L}_{\mathfrak{A}} \subseteq \Sigma^*$:

$$\mathbf{L}_{\mathfrak{A}} := \{w : w \in \Sigma^* \wedge \Delta(S, w) \in F\}. \tag{5}$$

Bezüglich (5) kann man für das System (4) im Sinne von MfL Bd. 9, 1.2., eine Analyse- und Syntheseaufgabe formulieren:

Welche Sprache wird von einem konkreten Automaten $\mathfrak{A}$ akzeptiert?
Welcher Automat $\mathfrak{A}$ akzeptiert eine vorgegebene Sprache? (6)

Es gilt folgender

Satz 1 [1, 13]. *Für jeden endlichen Automaten $\mathfrak{A}$ ist $\mathbf{L}_{\mathfrak{A}}$ eine rechts-lineare Sprache, und zu jeder rechts-linearen Sprache $\mathbf{L}$ kann ein endlicher Automat $\mathfrak{A}$ konstruiert werden, für den $\mathbf{L} = \mathbf{L}_{\mathfrak{A}}$ gilt.*

Beispielsweise findet man für die von der Grammatik (1) generierte Sprache als Akzeptor den endlichen Automaten (4) mit

$$\Phi = \{A_1, A_2, A_3\}, \qquad \Sigma = \{a, b\}, \qquad S = \{A_1\}, \qquad F = \{A_2\}$$

und der durch Tabelle 8.2 bestimmten Überführungsfunktion.

δ	a	b
A_1	A_1	A_2
A_2	A_3	A_3
A_3	A_3	A_3

Tabelle 8.2

Für die Lösung der Aufgaben (6) ist es nützlich, die Überführungsfunktion in der folgenden Weise graphisch darzustellen. Ist

$$\Phi = \{A_1, A_2, \ldots, A_m\} \quad \text{und} \quad \Sigma = \{a_1, a_2, \ldots, a_n\},$$

so wird jedem Zeichen von Φ ein Punkt zugeordnet und mit diesem markiert. Sodann verbindet man die A_i, A_k, $1 \leq i, k \leq m$, entsprechenden Punkte durch einen mit a_j, $1 \leq j \leq n$, markierten und von A_i nach A_k gerichteten Bogen genau dann, wenn $A_k = \delta(A_i, a_j)$ ist. Man erhält so einen gerichteten Graphen, der als *Zustandsdiagramm* (*Zustandsgraph*) des Automaten bezeichnet wird.

Beispiel. Wir betrachten nach [50] eine Mausefalle als endlichen Automaten (4). Es sei $\Phi = \{S, T\}$ und $\Sigma = \{a, b\}$ je ein Binäralphabet, wobei folgende Signalisierung der Zeichen stattfindet:

$$\Phi: \begin{cases} S & \text{Falle gespannt} \\ T & \text{Falle nicht gespannt} \end{cases} \qquad \Sigma: \begin{cases} a & \text{Maus geht in die Falle} \\ b & \text{Maus geht nicht in die Falle} \end{cases}$$

Die Übergangsfunktion läßt sich mit der folgenden Tabelle beschreiben:

δ	a	b
S	T	S
T	T	T

Der Anfangszustand sei S und die Menge der Endzustände $F = \{T\}$. In Abb. 8.6 ist das Zustandsdiagramm des Automaten dargestellt; man erkennt mit einem Blick, daß dieser alle Wörter akzeptiert, die mit einem Wort der Form $b^n a$ ($n = 0, 1, 2, \ldots$) beginnen.

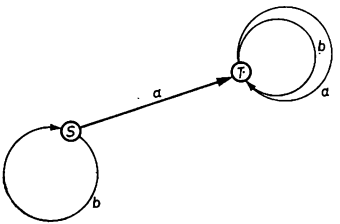


Abb. 8.6

Beispiele der Biologie ergeben sich aus dem Verhalten niederer Organismen, die in ihrer Reaktion auf Umweltsignale als endliche Automaten beschrieben werden können.

Die mit (4) gegebene abstrakte Beschreibung eines endlichen Automaten $\mathfrak{A}$ kann als Grundlage für den Entwurf technischer Systeme dienen, welche die Arbeitsweise von $\mathfrak{A}$ realisieren. Dabei werden Hilfsmittel der Schaltalgebra benutzt, die insofern einer Ergänzung bedürfen, als für eine taktgerechte Rückführung des gemäß

$$Z_{t+1} = \delta(Z_t, a_k)$$

gebildeten Outputzustandes gesorgt werden muß.

Literatur

[1] Aho, A. V., and J. D. Ullman, The Theory of Parsing, Translation, and Compiling, Prentice-Hall, Inc., Englewood Cliffs (N. J.) 1973.

[2] Arbib, M. A., Algebraische Theorie abstrakter Automaten, formaler Sprachen und Halbgruppen, Akademie-Verlag, Berlin 1973 (Übersetzung aus dem Englischen).

[3] Ascheb, L., und H. Kaiser, Exemplarische Betrachtungen zur Syntax und Semantik formaler Sprachen, Wiss. Z. PH „Karl Liebknecht" Potsdam 22 (1978), 165–181.

[4] Asser, G., Einführung in die mathematische Logik, Teil I (4. Aufl.), II, BSB B. G. Teubner Verlagsgesellschaft, Leipzig 1972.

[5] Bauer, F. L., und G. Goos, Informatik — eine einführende Übersicht. Springer-Verlag, Berlin–Heidelberg–New York 1973.

[6] Beresin, I. S., und N. P. Shidkow, Numerische Methoden, Bd. 1, 2, VEB Deutscher Verlag der Wissenschaften, Berlin 1970, 1971 (Übersetzung aus dem Russischen).

[7] Collatz, L., und J. Albrecht, Aufgaben aus der Angewandten Mathematik, Bd. 2, Akademie-Verlag, Berlin 1973.

[8] Collatz, L., Das Hornersche Schema bei komplexen Wurzeln algebraischer Gleichungen, ZAMM 20 (1940), 235–236.

[9] Demidowitsch, B. P., and I. A. Maron, Computational Mathematics, MIR Publishers, Moskau 1973 (Übersetzung aus dem Russischen).

[10] Demidowitsch, B. P., I. A. Maron und E. S. Schuwalowa, Numerische Methoden der Analysis, VEB Deutscher Verlag der Wissenschaften, Berlin 1968 (Übersetzung aus dem Russischen).

[11] Döring, B., Das Fixpunktprinzip in der Analysis, in: D. Laugwitz (Hrsg.), Überblicke Mathematik II, Bibliographisches Institut, Mannheim 1969, S. 151–210.

[12] Faddejew, D. K., und W. N. Faddejewa, Numerische Methoden der linearen Algebra, 5. Aufl., VEB Deutscher Verlag der Wissenschaften, Berlin/R. Oldenbourg, München 1979 (Übersetzung aus dem Russischen).

[13] Fu, K. S., Syntactic Methods in Pattern Recognition, Academic Press, New York–London 1974.

[14] Fu, K. S., and A. Rosenfeld, Pattern recognition and image processing, IEEE Trans. Comp. C-25 (1976), 1336–1346.

[15] Gastinel, N., Lineare numerische Analysis, Friedr. Vieweg & Sohn GmbH, Braunschweig/ VEB Deutscher Verlag der Wissenschaften, Berlin 1972 (Übersetzung aus dem Französischen).

[16] Gladkij, A. V., und I. A. Mel'čuk, Elemente der mathematischen Linguistik, VEB Deutscher Verlag der Wissenschaften, Berlin/Fink-Verlag, München 1973 (Übersetzung aus dem Russischen).

[17] GORTZEL, G., An algorithm for the evaluation of finite trigonometric series, Amer. Math. Monthly **65** (1958), 34–35.
[18] GREVILLE, T. N. E., Data fitting by spline functions, MRC Technical Summary Report # 893, Madison (Wisc.) 1968.
[19] GREVILLE, T. N. E., Spline functions, interpolation, and numerical quadrature, in: A. RALSTON and H. S. WILF, Mathematical methods for digital computers II, John Wiley & Sons, New York–London–Sidney 1967, Chapter 8.
[20] GREVILLE, T. N. E., Introduction to spline functions, in: T. N. E. GREVILLE (Hrsg.), Theory and Applications of Spline Functions, Academic Press, New York–London 1969, p. 1–35.
[21] GROS, M., und A. LENTHIN, Mathematische Linguistik, Springer-Verlag, Berlin–Heidelberg–New Yok 1971.
[22] HOROWITZ, S. L., A syntactic algorithm for peak detection in waveforms with applications to cardiography, Commun. ACM **18** (1975), 281–285.
[23] ISAACSON, E., und H. B. KELLER, Analyse numerischer Verfahren, Edition, Leipzig 1972 (Übersetzung aus dem Englischen).
[24] KANTOROWITSCH, L. W., und G. P. AKILOW, Funktionalanalysis in normierten Räumen, Akademie-Verlag, Berlin 1964 (Übersetzung aus dem Russischen).
[25] KLAUS, G., Kybernetik und Erkenntnistheorie, VEB Deutscher Verlag der Wissenschaften, Berlin 1967.
[26] KNUTH, D. E., Semantics of context-free languages, Math. Systems Theory **2** (1968), 127–145.
[27] KRASNOSELSKI, M. A., u. a., Näherungsverfahren zur Lösung von Operatorgleichungen, Akademie-Verlag, Berlin 1973 (Übersetzung aus dem Russischen).
[28] KÜNZI, H. P., H. G. TZSCHACH and C. A. ZEHNDER, Numerical Methods of Mathematical Optimization, Academic Press, London–New York 1968.
[29] LAGUERRE, E., Sur une méthode pour obtenir par approximation les racines d'une équation algébraique qui a toutes ses racines réelles, Nouv. Ann. math., 2. Sér., **19** (1880), 161–171.
[30] Lehrbuch – Biologie (Klasse 12), Verlag Volk und Wissen, Berlin 1973.
[31] LUDWIG, R., Methoden der Fehler- und Ausgleichsrechnung, Friedr. Vieweg & Sohn, Braunschweig/VEB Deutscher Verlag der Wissenschaften, Berlin 1969.
[32] MAURER, H., Theoretische Grundlagen der Programmiersprachen – Theorie der Syntax, Bibliographisches Institut, Mannheim–Wien–Zürich 1969.
[33] MERZ, G., Erzeugende Funktionen bei Spline-Interpolation mit äquidistanten Knoten, Computing **12** (1974), 195–201.
[34] OBRESCHKOFF, N., Verteilung und Berechnung der Nullstellen reeller Polynome, VEB Deutscher Verlag der Wissenschaften, Berlin 1963.
[35] PAVLIDIS, T., Waveform segmentation through functional approximation, IEEE Trans. Comp. **C-22** (1973), 689–697.
[36] PAVLIDIS, T., and S. L. HOROWITZ, Segmentation of plane curves, IEEE Trans. Comp. **C-23** (1974), 860–870.
[37] PENNINGTON, R. H., Introductory Computer Methods and Numerical Analysis, The Mac Millan Company, New York 1970.
[38] PHILLIPS, G. M., Algorithms for piecewise straight line approximations, Computer J. **11** (1968), 211–212.
[39] PIEHLER, J., Einführung in die lineare Optimierung, Teubner-Verlagsgesellschaft, Leipzig 1964.
[40] RABIN, M. O., Complexity of computations, Commun. ACM **20** (1977), 625–633.
[41] REIDEMEISTER, K., Topologie der Polyeder und kombinatorische Topologie der Komplexe, Akademische Verlagsgesellschaft Geest & Portig, Leipzig 1953.
[42] Report on the Algorithmic Language ALGOL 60, Hrsg. P. NAUR, Numer. Math. **2** (1960), 106–136.

[43] Rice, J. R., The Approximation of Functions, Addison-Wesley Publ. Comp., Reading (Mass.)—Palo Alto—London 1964.
[44] Schoenberg, I. J., Monosplines and quadrature formulae, in: T. N. E. Greville (Hrsg.), Theory and Application of Spline Functions, Academic Press, New York—London 1969, p. 158—206.
[45] Schumaker, L. L., Approximation by splines (p. 65—85), Some algorithms for the computation of interpolating and approximating spline functions (p. 87—102), in: T. N. E. Greville (Hrsg.), Theory and Application of Spline Functions, Academic Press, New York—London 1969.
[46] Schröter, K., Was ist eine mathematische Theorie, in: K. Berka und L. Kreiser (Hrsg.), Logik-Texte, Akademie-Verlag, Berlin 1971.
[47] Scott, D. S., Logic and programming languages, Commun. ACM 20 (1977), 634—641.
[48] Stoer, J., Einführung in die Numerische Mathematik, Springer-Verlag, Berlin—Heidelberg—Wien—New York 1976.
[49] Strassen, V., Gaussian elimination is not optimal, Numer. Math. 13 (1969), 345—356.
[50] Studienmaterial: Einführung in die Grundlagen der Kybernetik (Autorenkollektiv), Ministerium f. Volksbildung, Hauptabteilung Lehrerbildung, 1974.
[51] Tschebyscheff, P. L., Werke, Band I, St. Petersburg 1899.
[52] Tschebyscheff, P. L., Werke, Band II, St. Petersburg 1907.
[53] Völz, H., Informationsspeicherung in Natur und Technik, Wissenschaft und Fortschritt, Heft 6 (1976), 242—248.
[54] Weber, H., Lehrbuch der Algebra, Vieweg & Sohn, Braunschweig 1912.

Namen- und Sachverzeichnis