Rostocker Mathematisches Kolloquium

Heft 12

Wilhelm-Pieck-Universität
Rostock

ROSTOCKER MATHEMATISCHES KOLLOQUIUM

Heft 12

Gewidmet

dem 30. Jahrestag der Gründung der

DEUTSCHEN DEMOKRATISCHEN REPUBLIK

1979

Wilhelm-Pieck-Universität Rostock
Sektion Mathematik

Redaktion: Abt. Wissenschaftspublizistik der Wilhelm-Pieck-Universität Rostock, 25 Rostock, Vogelsang 13/14
Fernruf 369 577

Verantwortlicher Redakteur: Dipl.-Ges.-Wiss. Bruno Schrage
Fachredakteur: Doz. Dr. sc. nat. Gerhard Maeß, Sektion Mathematik

Herausgegeben von der Wilhelm-Pieck-Universität Rostock unter Genehmigungs-Nr. C 124/79

Druck: Ostsee-Druck Rostock, Werk II

Inhalt

Rostock. Math. Kolloq. 12, 5 - 11 (1979)

Lothar Berg

Lösungsdarstellungen für eine Rekursionsformel der Kombinatorik

Für die Anzahl H_{nk} der Elemente einer gewissen Auswahlmenge wurden in /3/ die Rekursionsformel mit variabler Gliederzahl

$$H_{n+1,k} = H_{nk} + \sum_{i=1}^{k} \binom{q}{i} H_{n-1,k-i} \tag{1}$$

und die Anfangsbedingungen

$$H_{0k} = \binom{0}{k}, \quad H_{1k} = \binom{q}{k} \tag{2}$$

aufgestellt, wobei n,k nichtnegative, ganze Zahlen sind und q ein Parameter ist, der hier nicht weiter eingeschränkt zu werden braucht. Proberechnungen zeigen, daß die ersten H_{nk} die Darstellung

$$H_{nk} = \sum_{m=0}^{[(n+1)/2]} h_{nm} \binom{mq}{k} \tag{3}$$

mit den anschließend angeführten von q unabhängigen Koeffizienten h_{nm} besitzen.

m \ n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	0	-1	-1	0	1	1	0	-1	-1	0	1	1	0	-1	-1
1		1	2	1	-2	-4	-2	3	6	3	-4	-8	-4	5	10	5
2				1	3	3	-2	-9	-9	3	18	18	-4	-30	-30	5
3						1	4	6	0	-15	-24	-6	36	60	20	-70
4								1	5	10	5	-20	-49	-35	50	145
5										1	6	15	14	-21	-84	-98
6												1	7	21	28	-14

Im folgenden sollen für die h_{nm} vier verschiedene Darstellungen hergeleitet und das asymptotische Verhalten dieser Koeffizienten diskutiert werden. Eine dieser Darstellungen wurde in /3/

auf anderem Wege hergeleitet, und im Spezialfall q = 1 wurde bereits von I. Kaplansky

$$H_{nk} = \binom{n-k+1}{k} \text{ für } n \geq 2k - 1 \tag{4}$$

sowie $H_{nk} = 0$ für $n < 2k-1$ berechnet.

1. Erzeugende Funktionen

Führen wir die erzeugende Funktion

$$f_n(z) = \sum_{k=0}^{\infty} H_{nk} z^k \tag{5}$$

ein und benutzen wir die Bezeichnung

$$p = (1+z)^q, \tag{6}$$

so erhalten wir aus (1) die Rekursionsformel zweiter Ordnung

$$f_{n+1}(z) = f_n(z) + (p-1) f_{n-1}(z)$$

und aus (2) die zugehörigen Anfangsbedingungen $f_0(z) = 1$, $f_1(z) = p$. Für die weitere erzeugende Funktion

$$F(w) = \sum_{n=0}^{\infty} f_n(z) w^n \tag{7}$$

ergibt sich hieraus die Darstellung

$$F(w) = \frac{1-w+pw}{1-w+w^2-pw^2}$$

mit der Entwicklung

$$F(w) = \frac{1-w}{1-w+w^2} + \sum_{i=1}^{\infty} \frac{w^{2i-1}}{(1-w+w^2)^{i+1}} p^i. \tag{8}$$

Offenbar besitzt die Funktion (7) auch eine zweidimensionale Entwicklung der Form

$$F(w) = \sum_{n,m=0}^{\infty} h_{nm} w^n p^m,$$

wobei wir hierdurch die h_{nm} unabhängig von (3) einführen, und daher besitzt die Funktion (5) auch die Entwicklung

$$f_n(z) = \sum_{m=0}^{\infty} h_{nm} p^m. \qquad (9)$$

Durch Vergleich von (8) mit der anschließenden Gleichung folgt

$$\sum_{n=0}^{\infty} h_{no} w^n = \frac{1-w^2}{1+w^3} \qquad (10)$$

und für $m \geq 1$

$$\sum_{n=0}^{\infty} h_{nm} w^n = \frac{w^{2m-1}}{(1-w+w^2)^{m+1}} . \qquad (11)$$

Durch Einsetzen von (6) in (9) und Vergleich mit (5) erkennt man, daß für die hier eingeführten h_{nm} die Gleichung (3) ohne Einschränkung gültig ist, da aus (11) insbesondere $h_{nm} = 0$ für $n < 2m-1$ hervorgeht. Die Entwicklung (10) bestätigt die in der obenstehenden Tabelle für h_{no} angegebenen Werte und zeigt darüber hinaus, daß diese Werte in n die Periode 6 besitzen. Aus (9) und der Rekursionsformel für $f_n(z)$ folgt die Differenzengleichung

$$h_{n+1,m} = h_{nm} - h_{n-1,m} + h_{n-1,m-1},$$

mit deren Hilfe man unter Beachtung der Anfangswerte h_{no} und $h_{om} = 0$ für $m \geq 1$ leicht die Tabelle ergänzen kann.

2. Drei Lösungsvarianten

Zur Berechnung der Koeffizienten von (11) mit $m \geq 1$ erweitern wir erstens die rechte Seite mit $(1+w)^{m+1}$ und entwickeln die Klammerausdrücke der Erweiterung

$$w^{2m-1}(1+w)^{m+1}(1+w^3)^{-m-1}$$

in Binomialreihen. Dann gelangen wir zu der Darstellung

$$h_{nm} = \sum_j \binom{m+j}{j}\binom{m+1}{3m+3j-n}(-1)^j, \qquad (12)$$

wobei über alle j mit $n-3m \leq 3j \leq n-2m+1$ zu summieren ist, also über höchstens $[(m+4)/3]$ Glieder.

Zweitens können wir aus (11) durch Anwendung der Entwicklung

$$(1-w+w^2)^{-m-1} = \sum_{j=0}^{\infty} \binom{m+j}{j} w^j (1-w)^j$$

und nachträglicher Entwicklung von $(1-w)^j$ die Darstellung

$$h_{nm} = \sum_{j} \binom{m+j}{j} \binom{j}{n-2m-j+1} (-1)^{n+j+1} \tag{13}$$

herleiten, wobei über alle j mit $n-2m+1 \leq 2j \leq 2n-4m+2$ zu summieren ist, also über höchstens $[(n+3)/2] - m$ Glieder. Mit Hilfe der Indexverschiebung $k = n-m-j+1$ entsteht aus (13) der auch aus /3/ zu entnehmende Ausdruck

$$h_{nm} = \sum_{k=m}^{[(n+1)/2]} \binom{n-k+1}{k} \binom{k}{m} (-1)^{m+k}. \tag{14}$$

Drittens gelangen wir mit Hilfe der Partialbruchzerlegung (vgl. /1/)

$$\frac{1}{(1-w+w^2)^{m+1}} = \sum_{j=0}^{m} \binom{m+j}{j} \sqrt{3}^{-m-j-1} \left(\left(\frac{\varepsilon-w}{i}\right)^{j-m-1} + \left(\frac{\bar{\varepsilon}-w}{-i}\right)^{j-m-1} \right),$$

wobei $\varepsilon = (1+i\sqrt{3})/2$ eine sechste Einheitswurzel ist, nach erneuter Anwendung der Binomialentwicklung zu dem Ergebnis

$$h_{nm} = 2 \sum_{j=0}^{m} \binom{m+j}{j} \binom{n-m-j+1}{m-j} \sqrt{3}^{-m-j-1} \cos\left((5m-2n-j-1)\frac{\pi}{6}\right). \tag{15}$$

Unter Beachtung der bekannten Orthogonalitätsrelationen

$$\sum_{m=j}^{k} \binom{k}{m} \binom{m}{j} (-1)^{m+k} = \delta_{kj}$$

ergibt sich aus (14) für $2 \leq 2j \leq n+1$ die Beziehung

$$\sum_{m=j}^{[(n+1)/2]} \binom{m}{j} h_{nm} = \binom{n-j+1}{j}, \tag{16}$$

die wegen

$$\sum_{m \geq 0} h_{nm} = f_n(0)$$

und

$$\sum_{n=0}^{\infty} f_n(0)w^n = \frac{1}{1-w},$$

d. h. $f_n(0) = 1$, auch für $j = 0$ gültig ist und sich bei der Berechnung der h_{nm} gut zur Kontrolle der Ergebnisse eignet. Wegen (3) besagt die Beziehung (16) nichts anderes als die Gleichung (4), und ebenfalls wegen (3) geht (16) für $j = 0$ und $j = 1$ in die in /3/ angegebenen Gleichungen $H_{no} = 1$ bzw. $H_{n1} = nq$ über.

3. Asymptotische Abschätzungen

Abschließend sollen für die Koeffizienten h_{nm} noch einige asymptotische Approximationen aufgestellt werden. Aus (15) folgt bei festem $m > 0$ für $n \to \infty$

$$h_{nm} = \frac{2n^m}{m!\sqrt{3}^{m+1}} \cos((5m-2n-1)\frac{\pi}{6}) + O(n^{m-1}), \tag{17}$$

wobei die Kosinusfunktion in n die Periode 6 besitzt.

Aus (13) folgt bei festem $n-2m+1 = k \geq 0$ für $m \to \infty$

$$h_{nm} = \frac{1}{k!} m^k + O(m^{k-1}). \tag{18}$$

Um weitere asymptotische Aussagen herzuleiten, benutzen wir für die Koeffizienten h_{nm} die aus (11) hervorgehende vierte Darstellung

$$h_{nm} = \frac{1}{2\pi i} \int \frac{w^{2m-n-2}}{(1-w+w^2)^{m+1}} dw, \tag{19}$$

wobei über einen geschlossenen Weg in der komplexen w-Ebene zu integrieren ist, der den Nullpunkt genau einmal in der positiven Richtung umläuft, die Pole $w = \xi$ und $w = \bar{\xi}$ aber nicht einschließt, und wenden die Sattelpunktmethode an. Die Sattelpunkte der Betragsfläche des Integranden sind die Nullstellen der Ableitung, d. h., wenn wir $n+4 = s$ und $m+1 = \alpha s$ mit $\alpha \in (0,1/2)$ setzen, die Lösungen der Gleichung

$$w^2 - (1-\alpha)w + 1-2\alpha = 0, \tag{20}$$

und lauten daher

$$w_{1,2} = \frac{1}{2}(1-\alpha) \pm \frac{1}{2}\sqrt{\alpha^2+6\alpha-3}. \tag{21}$$

Somit liefert die Sattelpunktmethode (vgl. /2/) für $s \to \infty$

$$h_{nm} = O\left(\frac{a^s}{\sqrt{s}}\right), \tag{22}$$

wobei 1-1 die Ordnung des Sattelpunktes ist und

$$a = \frac{|w_1|^{2\alpha-1}}{|1-w_1+w_1^2|^{\alpha}} . \tag{23}$$

Die asymptotische Abschätzung (22) ist scharf, d. h., das Symbol O kann nicht durch das Symbol o ersetzt werden, wenn w_1 ein Sattelpunkt ist, für den (23) minimal wird, und ließe sich auch zu einer asymptotischen Darstellung bzw. Entwicklung weiterführen.

Aus (21) geht hervor, daß für $2\sqrt{3}-3 < \alpha < 1/2$ zwei reelle Sattelpunkte erster Ordnung vorliegen, und das Verhalten des Integranden von (19) für reelle w zeigt, daß für w_1 die kleinere der Nullstellen (21) mit dem Minuszeichen vor der Wurzel zu wählen ist. Im Fall $\alpha = 2\sqrt{3}-3$ fallen die beiden Sattelpunkte zusammen, und es liegt ein Sattelpunkt zweiter Ordnung vor, wobei a den Wert

$$a = (2+\sqrt{3})\left(\frac{2-\sqrt{3}}{3}\right)^{\alpha} \approx 1{,}21639 \tag{24}$$

annimmt. Im Fall $0 < \alpha < 2\sqrt{3}-3$ sind die beiden Sattelpunkte (21) wieder von erster Ordnung und zueinander konjugiert komplex, folglich sind die zugehörigen Werte für a in (23) einander gleich. Da diese Sattelpunkte auf dem Kreis $|w-2| = \sqrt{3}$ liegen, kann (23) diesmal zu

$$a = \frac{(1-2\alpha)^{\alpha-1/2}}{(\alpha\sqrt{3})^{\alpha}} \tag{25}$$

umgeformt werden. Durch einfache Rechnung läßt sich bestätigen, daß die Werte (25) bez. α monoton wachsend und die Werte (23)

für $2\sqrt{3}-3 < \alpha < 1/2$ monoton fallend sind, so daß (24) das Maximum von a ist, während die beiden Minima gleich 1 sind.

Wie üblich ist die asymptotische Abschätzung (22) in jedem abgeschlossenen Teilintervall der Intervalle $(0, 2\sqrt{3}-3)$ und $(2\sqrt{3}-3, 1/2)$ bez. α gleichmäßig gültig. Weiterhin ist aus der Gestalt des Integranden und der Lage der Sattelpunkte ersichtlich, daß die h_{nm} für hinreichend große s im Fall $2\sqrt{3}-3 \leq \alpha < 1/2$ positiv sind und im Fall $0 < \alpha < 2\sqrt{3}-3$ um den Nullpunkt oszillieren.

Literatur

/1/ Berg, L.: Einführung in die Operatorenrechnung. Berlin 1965

/2/ Berg, L.: Asymtotische Darstellungen und Entwicklungen. Berlin 1968

/3/ Engel, K.: Über zwei Lemmata von Kaplansky. Rostock. Math. Kolloq. 9, 5 - 26 (1978)

eingegangen: 05. 04. 1979

Anschrift des Verfassers:

Prof. Dr. Lothar Berg
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 13 - 16 (1979)

Dietlinde Lau

Automorphismen auf den maximalen Klassen der k-wertigen Logik

In dieser Arbeit wird gezeigt, daß jeder Automorphismus auf einer maximalen Klasse der k-wertigen Logik ein innerer Automorphismus ist.

Von A. I. Mal'cev wurde in /2/ ein Automorphismus α auf einer abgeschlossenen Klasse A aus P_k ein innerer Automorphismus genannt, wenn eine eineindeutige Abbildung φ von E_k auf E_k existiert mit der Eigenschaft, daß für beliebige $n \geq 1$ und f aus A^n gilt:

$$f^{\alpha}(x_1, \ldots, x_n) = \varphi(f(\varphi^{-1}(x_1), \ldots, \varphi^{-1}(x_n)))$$

bzw.

$$f^{\alpha}(\varphi(x_1), \ldots, \varphi(x_n)) = \varphi(f(x_1, \ldots, x_n)).$$

Weitere im folgenden verwendete Begriffe und Bezeichnungen findet man in /1/ erläutert.
I. A. Mal'cev bewies in /3/ folgenden Satz:

Satz 1: Wenn eine abgeschlossene Klasse $A \subseteq P_k$ alle Konstanten von P_k enthält, dann sind sämtliche Automorphismen auf A innere.

Als unmittelbare Folgerung ergibt sich hieraus für $\varrho \in \mathfrak{M}_k \cup \mathfrak{N}_k \cup \mathfrak{L}_k \cup \mathfrak{D}_k$ der folgende Satz 2. Für $\varrho \in \mathfrak{C}_k$ ist Satz 2 ein Spezialfall von Theorem 2 aus /2/.

Satz 2: Sei $\varrho \in \mathfrak{M}_k \cup \mathfrak{N}_k \cup \mathfrak{L}_k \cup \mathfrak{C}_k \cup \mathfrak{D}_k$. Dann ist jeder Automorphismus auf Pol ϱ ein innerer Automorphismus.

Die Grundidee des Beweises von Satz 1 verwendend, läßt sich außerdem zeigen:

Satz 3: Auf Klassen vom Typ $\mathfrak{T}$ existieren nur innere Automorphismen.
Beweis: Sei $\varrho_s \in \mathfrak{T}_k$. Die Funktionen aus Pol ϱ_s werden im fol-

genden wie in /1/, § 3, dargestellt. Weiter sei α ein Automorphismus auf Pol $\mathcal{P}_s$ und N die Menge aller Funktionen

$$f_a(x) := \sum_{r=1}^{l} \sum_{i=0}^{p-1} j_{a_{r,i}}(x) \cdot s^i(a) \bmod k ,$$

wobei $k = l \cdot p$ (p Primzahl, $l \geq 1$),

$E_k = \{a_{1,0}, a_{1,1}, \ldots, a_{1,p-1}, a_{2,0}, a_{2,1}, \ldots, a_{2,p-1}, \ldots, a_{l,0}, \ldots, a_{l,p-1}\}$,

$a \in E_k$ und $s(a_{r,i}) = a_{r,i+1 \bmod p}$ $(r=1,2,\ldots,l;\ i=0,1,\ldots,p-1)$.

Bekanntlich ändert sich bei einer isomorphen Abbildung die Stellenzahl der Funktionen nicht. Somit ist α auch ein Automorphismus auf der Halbgruppe $((\text{Pol } \mathcal{P}_s)^1, *)$.

Wir zeigen zunächst, daß $N^\alpha = N$.

Offensichtlich existiert für jede Funktion $g \in (\text{Pol } \mathcal{P}_s)^1$ und für jede Funktion $f_a \in N$ eine Funktion $h \in (\text{Pol } \mathcal{P}_s)^1$ mit $f_a = h * g$.

Folglich ist

$$f_a^\alpha = h^\alpha * g^\alpha. \qquad (1)$$

Sei speziell g eine einstellige Funktion aus Pol $\mathcal{P}_s$ mit $g^\alpha \in N$. Dann ist $h^\alpha * g^\alpha \in N$ und (wegen (1)) $f_a^\alpha \in N$ für beliebiges $a \in E_k$, d. h. $N^\alpha = N$.

Wir leiten einige Eigenschaften der Funktionen f_a her.

Für alle $r \in \{1, 2, \ldots l\}$ und $i \in E_p$ gilt: $f_{a_{r,0}} * f_{a_{r,i}} = f_{a_{r,i}}$

und folglich

$$f^\alpha_{a_{r,0}} * f^\alpha_{a_{r,i}} = f^\alpha_{a_{r,i}} ,$$

d. h., für jedes r existiert ein r' mit $f^\alpha_{a_{r,0}} = f_{a_{r',0}}$.

Weiter ist

$$f_{a_{r,i}} = \underbrace{(\ldots((f_{a_{r,1}} * f_{a_{r,1}}) * f_{a_{r,1}})\ldots f_{a_{r,1}})}_{i\text{-mal}}.$$

Hieraus ergibt sich für $f^{\alpha}_{a_{r,1}} = f_{a_{r''},b_r}$ $(b_r \in E_p \setminus \{0\})$:

$$f^{\alpha}_{a_{r,i}} = f_{a_{r''},i\cdot b_r \bmod p}\,.$$

Wegen

$$f_{a_{r,0}} * f_{a_{1,1}} = f_{a_{r,1}} \text{ ist}$$

$$f^{\alpha}_{a_{r,0}} * f^{\alpha}_{a_{1,1}} = f^{\alpha}_{a_{r,1}} = f_{a_{r''},b_r}$$

$$= f_{a_{r',0}} * f_{a_{1''},b_1} = f_{a_{r'},b_1},$$

d. h. $r' = r''$ und $b_r = b_1 =: b$ für beliebiges $r \in \{1,2,\ldots,l\}$.
Zusammenfassend erhält man:

$$f^{\alpha}_{a_{r,i}} = f_{a_{r'},i\cdot b \bmod p}\,, \tag{2}$$

wobei $r, r' \in \{1, 2, \ldots, l\}$, $i \in E_p$ und $b \in E_p \setminus \{0\}$.
Aus (2) ergibt sich eine eineindeutige Abbildung φ von E_k auf E_k:

$$\varphi(a_{r,i}) := a_{r',i\cdot b \bmod p}\,.$$

Wir beweisen als nächstes, daß für jedes $a_{u,v} \in E_k$ gilt:

$$f^{\alpha}_{a_{u,v}}(\varphi(x)) = \varphi(f_{a_{u,v}}(x)). \tag{3}$$

Sei $x = a_{r,i}$ $(r \in \{1, 2, \ldots, l\}, i \in E_p)$. Dann ist

$$f^{\alpha}_{a_{u,v}}(\varphi(a_{r,i})) = f_{a_{u',v\cdot b \bmod p}}(a_{r',i\cdot b \bmod p})$$

$$= s^{i\cdot b}(a_{u',v\cdot b \bmod p}) = a_{u',(v+i)\cdot b \bmod p}$$

und

$$\varphi(f_{a_{u,v}}(a_{r,i})) = \varphi(s^{i}(a_{u,v})) = \varphi(a_{u,v+i \bmod p})$$

$$= a_{u',(v+i)\cdot b \bmod p}\,.$$

Die Behauptung (3) ist also richtig.
Die Abbildung φ definiert auf P_k einen inneren Automorphis-

mus α_φ. Anstelle des Automorphismus α betrachten wir den Isomorphismus $\beta = \alpha \cdot \alpha_\varphi^{-1}$ von Pol ϱ_s auf $(\text{Pol } \varrho_s)^\beta$. Unser Satz ist bewiesen, wenn wir zeigen können, daß der Isomorphismus β jede Funktion aus Pol ϱ_s in sich überführt. Aus der Konstruktion von α_φ und (3) ergibt sich $f_a^\beta = f_a$ für alle $a \in E_k$.

Sei jetzt f eine n-stellige Funktion aus Pol ϱ_s, $f^\beta =: g$ und $\tilde{a} = (a_1,\ldots,a_n) \in E_k^n$. Dann gilt $f(f_{a_1}(x),\ldots,f_{a_n}(x)) = f_{f(\tilde{a})}(x)$ und

$$f^\beta(f_{a_1}^\beta(x),\ldots,f_{a_n}^\beta(x)) = f_{f(\tilde{a})}^\beta(x)$$

$$= f^\beta(f_{a_1}(x),\ldots,f_{a_n}(x)) = f_{f(\tilde{a})}(x)$$

$$= g(f_{a_1}(x),\ldots,f_{a_n}(x)) = f_{g(\tilde{a})}(x).$$

Folglich ist $f(\tilde{a}) = g(\tilde{a})$ für alle $\tilde{a} \in E_k^n$, d. h. $f = g$.
Unser Satz ist damit bewiesen.

Literatur

/1/ Lau, D.: Bestimmung der Ordnung maximaler Klassen von Funktionen der k-wertigen Logik. Z. Math. Logik Grundlagen Math. 24, 79 - 96 (1978)

/2/ Mal'cev, A. I.: Iterativnye algebry i mnogoobrazija Posta (Russ.). Algebra i Logika 5, 2, 5 - 24 (1966)

/3/ Mal'cev, I. A.: Kongruencii i avtomorfizmy na kletkach algebr Posta (Russ.). Algebra i Logika 11, 6, 666 - 672 (1972)

eingegangen: 20. 03. 1979

Anschrift des Verfassers:

Dr. rer. nat. Dietlinde Lau
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 17 - 30 (1979)

Günther Steffen

Eigenschaften von Matrixgruppen über dem Restklassenring modulo p^k

In dieser Arbeit werden zunächst einige Resultate über total irreduzible zyklische Matrixgruppen in Anlehnung an Steffen /3/ gebracht. Zu jeder total irreduziblen zyklischen $n \times n$-Matrixgruppe von der Ordnung p^n-1 über R_k gibt es eine äquivalente Matrixgruppe, die sich aus der Begleitmatrix eines total irreduziblen Polynoms erzeugen läßt. Weiterhin wird ein Kriterium angegeben, wann zwei Darstellungen einer Gruppe äquivalent sind. Schließlich zeigen wir, daß eine Darstellung einer Gruppe H über R_k mit $p \nmid |H|$ genau dann total irreduzibel ist, wenn sie irreduzibel ist.

Es sei darauf verwiesen, daß die Ergebnisse dieser Arbeit für Strukturuntersuchungen gewisser metabelscher Gruppen nützlich sind. In späteren Arbeiten wird hierauf genauer eingegangen. Es werden dort metabelsche Gruppen mit abelscher Fittinggruppe untersucht.

Bezeichnungen

$\mathbb{Z}$: Ring der ganzen rationalen Zahlen; Aut G : Automorphismengruppe der Gruppe G; R_k : Restklassenring modulo p^k (dabei geht die Primzahl p aus dem jeweiligen Zusammenhang hervor, $k \geq 1$); für eine Matrix C über R_k bezeichne C^* die modulo p reduzierte Matrix über GF(p) (s. (2.6) aus /3/); $\tau \sim \varrho$: die Darstellungen τ und ϱ sind äquivalent; $(f(x))$: das aus dem Polynom f(x) erzeugte Hauptideal in $R_k[x]$; $R_k[x]/(f(x))$: Faktorstruktur von $R_k[x]$ nach $(f(x))$; $[g(x)]_{f(x)}$: Klasse von $R_k[x]$ nach $(f(x))$, in der g(x) liegt; $R_k[C]$: lineare Hülle aller Potenzen der quadratischen Matrizen C über R_k; $H \leqq G$ (bzw. $H < G$) : H ist Untergruppe (bzw. echte Untergruppe) der Gruppe G; $\langle K \rangle$: Erzeugnis des

Komplexes $K \subseteq G$; falls $\mathfrak{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ ein Spaltenvektor von Elementen einer additiv geschriebenen abelschen Gruppe F ist, dann bedeutet $\langle \mathfrak{y} \rangle$ das Erzeugnis aller Elemente von $\mathfrak{y}$, $a\mathfrak{y} := \begin{pmatrix} ay_1 \\ \vdots \\ ay_n \end{pmatrix}$ $(a \in \mathbb{Z})$, $\mathfrak{y}h := \begin{pmatrix} y_1 h \\ \vdots \\ y_n h \end{pmatrix}$ $(h \in \operatorname{Aut} f)$, $\mathfrak{y} + N := \begin{pmatrix} y_1 + N \\ \vdots \\ y_n + N \end{pmatrix}$ für $N \leq F$; $\mathfrak{U}_n(F) := \{p^n x \mid x \in F\}$ mit $n \in \mathbb{Z}$ und F eine endliche additiv geschriebene abelsche p-Gruppe; für eine quadratische Matrix C über einem Ring mit Einselement $\underline{1}$ bedeute $a \mid C$ oder $C \equiv 0 (\mathrm{mod}\ a)$ $(a \in \mathbb{Z})$, daß $a\underline{1}$ jeden Koeffizienten von C teilt.

1. Sei F eine beliebige additiv geschriebene endliche abelsche Gruppe. Bezeichne $\mathfrak{w} = (v_1, \ldots, v_n)$ eine Basis von F, bei der die Ordnungen der Basiselemente Primzahlpotenzen sind, und alle Basiselemente, deren Ordnungen Primzahlpotenzen ein und derselben Primzahl sind, mögen nebeneinander stehen. Ansonsten sollen die Basiselemente nach der Größe ihrer Ordnungen geordnet sein. Eine solche Basis nennen wir auch geordnete Basis von F. Für eine $n \times n$-Matrix C über $\mathbb{Z}$ bezeichne $\overline{C}$ diejenige Matrix, die aus C entsteht, wenn man jeden Koeffizienten der i-ten Spalte von C $(i=1,\ldots,n)$ durch seine Restklasse modulo der Ordnung des i-ten Basiselementes ersetzt. Jedem Automorphismus α von F ist in natürlicher Weise eine Matrix $\overline{\| x_{ij}^{(\alpha)} \|}$ durch

$$v_i \alpha = \sum_{j=1}^{n} x_{ij}^{(\alpha)} v_j \qquad (i=1,\ldots,n;\ x_{ij}^{(\alpha)} \in \mathbb{Z})$$

zugeordnet. Dabei geht $\overline{\| x_{ij}^{(\alpha)} \|}$ aus der Koeffizientenmatrix $\| x_{ij}^{(\alpha)} \|$ hervor. Bezeichne Δ_F die folgende Menge:

$$\Delta_F := \left\{ \overline{\| x_{ij}^{(\alpha)} \|} \,\middle|\, \alpha \in \operatorname{Aut} F \right\}.$$

Die Abbildung $\alpha \to \overline{\| x_{ij}^{(\alpha)} \|}$ ist eineindeutig. Dem Produkt $\alpha\beta$ zweier Automorphismen α, β von F ist offensichtlich die Matrix

$\overline{\|x_{ij}^{(\alpha)}\| \, \|x_{ij}^{(\beta)}\|} \in \Delta_F$ zugeordnet. Also ist Δ_F mit der Verknüpfung $\overline{\|x_{ij}^{(\alpha)}\| \, \|x_{ij}^{(\beta)}\|} := \overline{\|x_{ij}^{(\alpha)}\| \, \|x_{ij}^{(\beta)}\|}$ eine Gruppe, und zwar eine zu Aut F isomorphe Matrixgruppe.

Für eine Matrix $L \in \Delta_F$ wollen wir mit int L diejenige Matrix über $\mathfrak{k}$ bezeichnen, für die $\overline{\text{int } L} = L$ gilt und in der die Koeffizienten der i-ten Spalte (i=1,...,n) natürliche Zahlen kleiner als die Ordnung des i-tem Basiselementes von $\mathfrak{w}$ sind.

Ein Homomorphismus einer Gruppe H in Δ_F soll auch Darstellung genannt werden. Falls eine Darstellung von H in Δ_F gegeben ist, so läßt sich H bei vorgegebener Basis von F in bekannter Weise zu einem Rechtsoperatorenbereich von F machen. Umgekehrt gehört zu einem Rechtsoperatorenbereich H von F bei vorgegebener Basis eine Darstellung von H in Δ_F. In diesem Sinne soll F auch Darstellungsmodul von H bez. der Basis $\mathfrak{w}$ genannt werden.

Zwei Darstellungen σ und τ von H in Δ_F sollen äquivalent heißen, falls eine Matrix $T \in \Delta_F$ mit $T\sigma(h)T^{-1} = \tau(h)$ für alle $h \in H$ existiert.
Bezeichnung: $\sigma \sim \tau$.

Durch elementare Rechnungen überzeugt man sich davon, daß die Äquivalenz von Darstellungen mit der Basisänderung des Darstellungsmoduls gleichbedeutend ist, daß also folgendes gilt:

Sei eine Darstellung τ von H in Δ_F gegeben, und es gelte
$\mathfrak{w}\, h = (\text{int } \tau(h))\mathfrak{w} \quad (h \in H)$.

a) Für eine Matrix $T \in \Delta_F$ sei $\tilde{\tau}(h)$ definiert durch

$$\tilde{\tau}(h) = T\tau(h)T^{-1} .$$

Dann ist $\tilde{\mathfrak{w}} = (\text{int } T)\mathfrak{w}$ eine geordnete Basis von F, und es gilt $\tilde{\mathfrak{w}}\, h = (\text{int } \tilde{\tau}(h))\tilde{\mathfrak{w}}$. (1)

b) Sei umgekehrt $\tilde{\mathfrak{w}} := (\text{int } T)\mathfrak{w}$ eine geordnete Basis von F, und es gelte $\tilde{\mathfrak{w}} h = (\text{int } \tilde{\tau}(h))\tilde{\mathfrak{w}}$. Dann liegt T in Δ_F, und es gilt

$$\tilde{\tau}(h) = T\tau(h)T^{-1} .$$

Falls F vom Typ $(\underbrace{p^k, p^k, \ldots, p^k}_{n\text{-mal}})$ ist, so ist Δ_F eine Matrixgruppe des Grades n über R_k. Wir können dann wie folgt die Irreduzibilität bzw. die Reduzibilität von Darstellungen in Δ_F erklären.

Eine Matrixdarstellung $\mathfrak{G}$ einer Gruppe H des Grades n über R_k heißt reduzibel, wenn eine zu $\mathfrak{G}$ äquivalente Darstellung μ existiert mit

$$\mu(h) = \begin{Vmatrix} \mu_1(h) & \mu_2(h) \\ 0 & \mu_4(h) \end{Vmatrix} ,$$

wobei $\mu_1(h)$ für alle $h \in H$ eine $n_1 \times n_1$-Matrix mit $1 \leq n_1 < n$ ist. Andernfalls heißt $\mathfrak{G}$ irreduzibel.

Offensichtlich ist eine Darstellung $\mathfrak{G}$ einer Gruppe H in Δ_F genau dann reduzibel, wenn F eine H-zulässige Untergruppe U enthält, die eine Basis besitzt, welche sich zu einer Basis von F ergänzen läßt. U muß also vom Typ $(p^k,\ldots,p^k)$ sein, und falls U von solchem Typ ist, läßt sich jede Basis von U zu einer Basis von F ergänzen (s. Hilfssatz 1).

Sei C eine quadratische Matrix über R_k. Bezüglich der Begriffe wie reguläre Darstellung von $\langle C \rangle$ über R_k, R_k-$R_k[C]$-Modul, das zu C gehörige Minimalpolynom und totale Irreduzibilität eines Polynoms f(x) über R_k verweisen wir auf /3/. Analog zur totalen Irreduzibilität von Polynomen möge die totale Irreduzibilität von $\langle C \rangle$ dadurch definiert sein, daß $\langle C^* \rangle$ ebenfalls irreduzibel über GF(p) ist.

2. Seien die natürlichen Zahlen n und k und die Primzahl p vorgegeben. Nach Steffen /3/, S. 25, gibt es dann eine über R_k total irreduzible zyklische $n \times n$-Matrixgruppe $\langle C \rangle$ der Ordnung p^n-1. Wie wir im Satz 4 sehen werden, gilt außerdem auch $\operatorname{ord} C^* = p^n-1$. Bezeichne $f(x)=x^n+a_{n-1}x^{n-1}+\ldots+a_0$ das zu C gehörige Minimalpolynom. Dann ist f(x) total irreduzibel über R_k und

$f(x) \mid x^{p^n-1} - 1$. Die reguläre Darstellung von $\langle C \rangle$ hat bezüglich der Basis $C^0, C^1, \ldots, C^{n-1}$ des $R_k - R_k[C]$-Moduls $R_k[C]$ die Gestalt

$$C^i \longrightarrow A^i \qquad (i = 0, 1, \ldots, p^n-1)$$

mit

$$A = \begin{Vmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \ldots & 1 \\ -a_0 & & \ldots\ldots & & -a_{n-1} \end{Vmatrix},$$

d. h., A ist die zu dem Polynom f(x) gehörende Begleitmatrix. Da außer C^0 kein Element aus $\langle C \rangle$ alle Basiselemente bei Multiplikationen von rechts festläßt, ist der Kern dieser Darstellung trivial, so daß ord A = p^n-1 gilt. Dem Beweis des Satzes aus /3/, S. 25, ist unter anderem zu entnehmen, daß in $R_k[x]/(f(x))$ die Voraussetzung von Lemma 1.4 aus /3/ erfüllt ist. Die Anwendung dieses Lemmas führt uns auf $R_k[A] \sim R_k[C]$, denn beide Matrixringe sind zur regulären Darstellung von $R_k[x]/(f(x))$ äquivalent. Daraus folgt nun, daß $\langle A \rangle$ und $\langle A^* \rangle$ irreduzibel sind. Zusammengefaßt läßt sich sagen:

Satz 1: Zu vorgegebenen p,n,k gibt es ein total irreduzibles Polynom f(x) des Grades n über R_k, so daß die zu f(x) gehörende Begleitmatrix eine total irreduzible zyklische n × n-Matrixgruppe der Ordnung p^n-1 erzeugt.

Satz 2: Sei $\langle C \rangle$ eine total irreduzible zyklische n × n-Matrixgruppe der Ordnung r über R_k (p vorgegebene Primzahl; $r,n \geq 1$), so daß C^* ebenfalls die Ordnung r hat. Des weiteren sei $\langle a \rangle$ eine abstrakte zyklische Gruppe der Ordnung r. Zwei Isomorphismen τ und ϱ von $\langle a \rangle$ auf $\langle C \rangle$ mit

$$\tau : a^i \longrightarrow C^i,$$

$$\varrho : a^i \longrightarrow (C^x)^i, \qquad i = 1, 2, \ldots, r,$$

wobei x eine natürliche Zahl mit (x,r) = 1 bezeichne, sind genau dann äquivalente Darstellungen, wenn $x = p^j$ mit $1 \leq j \leq n$ ist.

Beweis: Falls $\tau \sim \varrho$ ist, so existiert eine Matrix T über R_k mit $TCT^{-1} = C^x$. Daraus folgt mittels (2.5) aus /3/ $T^*C^*T^{*-1} = (C^*)^x$. Nach (2.4) aus /3/ ist $R_1[C^*] \cong GF(p^n)$. Durch $C^* \longrightarrow T^*C^*T^{*-1}$ wird offensichtlich ein Automorphismus von $R_i[C^*]$ induziert, so daß zwangsläufig $x = p^j$ mit $1 \leq j \leq n$ gelten muß. Sei nun umgekehrt $x = p^j$ mit $1 \leq j \leq n$. Bezeichne $f(x)$ das zu C und $g(x)$ das zu C^{p^j} gehörende Minimalpolynom über R_k. Dann sind $f(x)^*$ und $g(x)^*$ die Minimalpolynome von C^* bzw. $(C^*)^{p^j}$. Durch $C^* \longrightarrow (C^*)^{p^j}$ wird aber ein Automorphismus des Körpers $R_1[C^*]$ induziert, so daß die Minimalpolynome von C^* und $(C^*)^{p^j}$ übereinstimmen müssen, d. h. $f(x)^* = g(x)^*$. Mittels Lemma 2.2 und Lemma 2.4 aus /3/ folgt dann $f(x) = g(x)$. Am Ende des Beweises zu Lemma 2.2 aus /3/ konnten wir sehen, daß $R_k[x]/(f(x)) \cong R_k[C]$ bezüglich

$$\sigma : [a_o x^o + \ldots\ldots]_{f(x)} \longrightarrow a_o C^o + \ldots\ldots$$

gilt. Daher muß der Isomorphismus σ von $R_k[x]/(f(x))$ auf $R_k[C]$ eine irreduzible treue Darstellung sein. Analog ist der Isomorphismus

$$\pi : [b_o x^o + \ldots\ldots]_{g(x)} \longrightarrow b_o (C^{p^j})^o + \ldots\ldots$$

von $R_k[x]/(g(x)) = R_k[x]/(f(x))$ auf $R_k[C]$ eine irreduzible treue Darstellung. Nach Lemma 1.4 aus /3/ sind dann σ und π zur regulären Matrixdarstellung von $R_k[x]/(f(x))$ über R_k äquivalent, d. h. $\sigma \sim \pi$. Es existiert also eine Matrix U über R_k mit $UCU^{-1} = C^{p^j}$, was $\tau \sim \varrho$ bedeutet.

Folgerung: Seien $\langle C \rangle$ und $\langle a \rangle$ wie im Satz 2 gegeben. Darüber hinaus mögen τ und ϱ zwei Isomorphismen von $\langle a \rangle$ auf $\langle C \rangle$ bezeichnen, und τ^* bzw. ϱ^* seien definiert durch

$$\tau^* : b \longrightarrow (\tau(b))^*, \qquad (b \in \langle a \rangle) .$$

$$\varrho^* : b \longrightarrow (\varrho(b))^*$$

Dann sind τ und ϱ über R_k äquivalent, wenn τ^* und ϱ^* über R_1 äquivalent sind, und umgekehrt.

Beweis: Falls τ^* und ϱ^* äquivalent sind, so ist $\tau^*(a)$ nach Satz 2 eine gewisse p-Potenz von $\varrho^*(a)$. Wegen ord C = ord C^* ist dann $\tau(a)$ die gleiche p-Potenz von $\varrho(a)$. Die nochmalige Anwendung des Satzes ergibt die Äquivalenz von τ und ϱ. Die Umkehrung ist trivial.

3. Die beiden folgenden Hilfssätze werden für den Beweis des Satzes 3 benötigt. Dieser wird dann ein wichtiges Hilfsmittel für die Charakterisierung von metabelschen Gruppen mit abelscher Fittinggruppe sein.

Hilfssatz 1: Sei F eine additiv geschriebene abelsche Gruppe vom Typ $(p^k, .., p^k)$ $(k \geq 1)$ und U eine Untergruppe vom Typ $(p^k, .., p^k)$. Dann läßt sich eine Basis von U zu einer Basis von F ergänzen.

Beweis: Nach Szele /4/ ist U eine reine Untergruppe von F (s. a. Fuchs /1/, S. 79). Dann ist U sogar ein direkter Summand in F (/1/, S. 80). Daraus folgt die Behauptung.

Hilfssatz 2: Sei F eine additiv geschriebene abelsche Gruppe vom Typ $(p^k, .., p^k)$ $(k \geq 1)$ mit n Basiselementen und U eine Untergruppe von F vom Typ $(p^{k_1}, .., p^{k_1})$ $(0 < k_1 \leq k)$ mit m Basiselementen $(m \leq n)$. Bezeichne $u_1, .., u_m$ eine Basis von U, dann existiert eine Basis $v_1, .., v_n$ von F mit $u_i \in \langle v_i \rangle$ $(i = 1, .., m)$.

Beweis: Bezeichne $w_1, ..., w_n$ zunächst irgendeine Basis von F. Dann läßt sich jedes u_i $(i = 1, .., m)$ in der Form

$u_i = \sum_{j=1}^{n} x_j^{(i)} w_j$ $(x_j^{(i)} \in \mathbb{Z})$ schreiben. Da ord $u_i = p^{k_1}$ ist, muß $x_j^{(i)} \equiv 0 (\mathrm{mod}\ p^{k-k_1})$ $(j = 1, .., n)$ gelten. Seien nun $\bar{x}_j^{(i)}$ $(j = 1, .., n)$ ganze Zahlen mit $x_j^{(i)} = \bar{x}_j^{(i)} p^{k-k_1}$. Es existiert mindestens ein j $(1 \leq j \leq n)$ mit $p \nmid \bar{x}_j^{(i)}$, da ord $u_i = p^{k_1}$ war.

Dann haben alle v_i, die definiert sind durch $v_i = \sum_{j=1}^{n} \overline{x}_j^{(i)} w_j$ ($i = 1, .., m$), die Ordnung p^k, und es gilt $p^{k-k_1} v_i = u_i$. Es ist nun alles gezeigt, wenn wir die Basiseigenschaft von $v_1, .., v_m$ innerhalb seines Erzeugnisses nachweisen können, denn nach Hilfssatz 1 läßt sich $v_1, .., v_m$ zu einer Basis von F ergänzen. Wenn $v_1, ..., v_m$ keine Basis innerhalb seines Erzeugnisses wäre, dann müßte ein i mit $1 \leq i \leq m - 1$ existieren, so daß $\langle v_1, .., v_i \rangle \cap \langle v_{i+1} \rangle \neq \{0\}$ ist. Demzufolge gäbe es eine Untergruppe V der Ordnung p in $\langle v_1, .., v_i \rangle \cap \langle v_{i+1} \rangle$. Einerseits gälte dann $V \leq \langle v_{i+1} \rangle$ und damit

$$V \leq \left\langle p^{k-k_1} v_{i+1} \right\rangle = \langle u_{i+1} \rangle , \qquad (2)$$

da $k_1 > 0$ vorausgesetzt war. Andererseits hätten wir $V \leq \langle v_1, .., v_i \rangle$ und darum $V = \langle y_1 v_1 + ... + y_i v_i \rangle$ mit gewissen y_i aus Z wegen der Zyklizität von V. Da ord V = p ist, müßte $y_j \equiv 0 (\mathrm{mod}\ p^{k-1})$ sein ($j = 1, .., i$), und es würde

$$V = \left\langle \overline{y}_1 (p^{k-k_1} v_1) + ... + \overline{y}_i (p^{k-k_1} v_i) \right\rangle = \left\langle \overline{y}_1 u_1 + ... + \overline{y}_i u_i \right\rangle$$

wegen $k_1 \geq 1$ folgen, wobei $y_j = \overline{y}_j p^{k-k_1}$ ($j=1,...,i$) ist. Diese letzte Beziehung ist aber ein Widerspruch zu (2), da $u_1, .., u_m$ eine Basis von U ist.

<u>Satz 3:</u> a) Sei F eine additiv geschriebene abelsche Gruppe vom Typ $(p^k, .., p^k)$ und $\mathfrak{M} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ eine Basis von F. Des weiteren sei H eine Gruppe aus Automorphismen von F mit $p \nmid |H|$ und μ die zu F gehörende Matrixdarstellung des Grades n in $\Delta_{\mathfrak{M}}$ bezüglich der Basis $\mathfrak{M}$. Die Matrixdarstellung μ ist genau dann reduzibel, wenn die Matrixdarstellung μ^*, die definiert ist

durch

$$\mu^* : h \longrightarrow \mu^*(h) := (\mu(h))^* \qquad (h \in H) ,$$

über GF(p) reduzibel ist.

b) Die Voraussetzung $p \nmid |H|$ darf in a) nicht fallengelassen werden.

Beweis: a) Falls μ reduzibel ist, so muß offensichtlich auch μ^* reduzibel sein. Sei nun μ^* reduzibel. Dann gibt es eine zu μ^* äquivalente Darstellung π über GF(p) der Gestalt

$$\pi(h) = \begin{Vmatrix} \pi_1(h) & \pi_2(h) \\ \pi_3(h) & \pi_4(h) \end{Vmatrix} \qquad (h \in H) ,$$

so daß $\pi_3(h)$ die Nullmatrix eines gewissen Formates $n_2 \times n_1$ ($n_1, n_2 > 0$, $n_1 + n_2 = n$) ist. Mit anderen Worten, es existiert eine $n \times n$-Matrix U über GF(p) mit $U\mu^*(h)U^{-1} = \pi(h)$ für alle $h \in H$. Sicher gibt es eine Matrix T über R_k mit $T^* = U$. Für die Matrixdarstellung ν, die definiert ist durch

$$\nu(h) := T\mu(h)T^{-1} \qquad (h \in H) ,$$

gilt dann $\nu^* = \pi$, wobei ν^* die Darstellung $h \longrightarrow \nu^*(h) := (\nu(h))^*$ bezeichne, denn es besteht die Beziehung

$$\nu^*(h) = (\nu(h))^* = (T\mu(h)T^{-1})^* = U\mu^*(h)U^{-1} = \pi(h)$$

(hierbei ist (2.6) aus Steffen /3/, S. 16, zu berücksichtigen). Habe $\nu(h)$ die Gestalt

$$\nu(h) = \begin{Vmatrix} \nu_1(h) & \nu_2(h) \\ \nu_3(h) & \nu_4(h) \end{Vmatrix} \qquad (h \in H) ,$$

wobei $\nu_3(h)$ eine $n_2 \times n_1$-Matrix ist. Mit $\pi_3(h)$ ist dann auch $(\nu_3(h))^*$ die Nullmatrix über GF(p).
Wenn es uns gelingt nachzuweisen, daß eine zu ν äquivalente Darstellung ϱ existiert, so daß $\varrho_3(h)$ eine $n_2 \times n_1$-Nullmatrix über R_k ist, wobei $\varrho_3(h)$ die aus

$$\varrho(h) = \begin{Vmatrix} \varrho_1(h) & \varrho_2(h) \\ \varrho_3(h) & \varrho_4(h) \end{Vmatrix} \qquad (h \in H)$$

ersichtliche Teilmatrix bedeute, dann ist a) gezeigt.
Wie wir soeben sahen, gibt es sicher eine natürliche Zahl $k_1 \geq 1$ mit $p^{k_1} \mid \nu_3(h)$ $(h \in H)$. Falls $k_1 = k$ gilt, so ist nichts mehr zu zeigen. Sei nun $k_1 < k$. Zufolge (1) ist $\mathfrak{y} := (\text{int } T)\mathfrak{w}$ eine Basis von F, bezüglich der H die Darstellung ν in Δ_F erfährt. Dabei möge $\mathfrak{y}_1$ den Spaltenvektor der ersten n_1 und $\mathfrak{y}_2$ den Spaltenvektor der letzten n_2 Elemente von $\mathfrak{y}$ bezeichnen. Gewiß ist $\mathfrak{x} = \begin{pmatrix} \mathfrak{x}_1 \\ \mathfrak{x}_2 \end{pmatrix}$ mit $\mathfrak{x}_1 := p^{k_1}\mathfrak{y}_1$ und $\mathfrak{x}_2 := \mathfrak{y}_2$ eine Basis seines Erzeugnisses V. Aus

$$\mathfrak{x}_1 h = p^{k_1}(\mathfrak{y}_1 h) = p^{k_1}((\text{int } \nu_1(h))\mathfrak{y}_1 + (\text{int } \nu_2(h))\mathfrak{y}_2)$$

$$= (\text{int } \nu_1(h))\mathfrak{x}_1 + p^{k_1}(\text{int } \nu_2(h))\mathfrak{x}_2$$

und

$$\mathfrak{x}_2 h = \mathfrak{y}_2 h = (\text{int } \nu_3(h))\mathfrak{y}_1 + (\text{int } \nu_4(h))\mathfrak{y}_2$$

ist wegen $p^{k_1} \mid \nu_3(h)$ die Zulässigkeit von V bezüglich H zu ersehen. V hat den Typ $(p^{k-k_1}, .., p^{k-k_1}, p^k, .., p^k)$ mit $0 < k-k_1 < k$. Aus /2/, S. 125, Aufg. 69 folgt nun die Existenz einer geordneten Basis $\mathfrak{x}' = \begin{pmatrix} \mathfrak{x}'_1 \\ \mathfrak{x}'_2 \end{pmatrix}$ von V, wobei $\mathfrak{x}'_1$ die n_1 Basiselemente der Ordnung p^{k-k_1} und $\mathfrak{x}'_2$ die restlichen n_2 Basiselemente der Ordnung p^k seien, bezüglich der $p \mid \nu'_3(h)$ gilt. Dabei ist $\nu'_3(h)$ die $n_2 \times n_1$-Teilmatrix aus

$$\nu'(h) = \begin{Vmatrix} \nu'_1(h) & \nu'_2(h) \\ \nu'_3(h) & \nu'_4(h) \end{Vmatrix} \qquad (h \in H),$$

die definiert ist durch $\mathfrak{x}'h = (\text{int } \nu'(h))\mathfrak{x}'$ $(h \in H)$. Aus Hilfssatz 2 folgt, daß eine Basis $\mathfrak{y}' = \begin{pmatrix} \mathfrak{y}'_1 \\ \mathfrak{y}'_2 \end{pmatrix}$ von F mit

$\mathfrak{y}'_1 = p^{k_1}\mathfrak{y}'_1$ existiert. Man überlegt sich, daß wir $\mathfrak{y}'_2 = \mathfrak{y}'_2$ annehmen können. Nun gilt

$$\mathfrak{y}'_2 h = \mathfrak{y}'_2 h = (\text{int } \nu'_3(h))\mathfrak{y}'_1 + (\text{int } \nu'_4(h))\,\mathfrak{y}'_2$$

$$= (\text{int } \nu'_3(h))p^{k_1}\mathfrak{y}'_1 + (\text{int } \nu'_4(h))\mathfrak{y}_2 ,$$

wobei $p^{k_1+1} \mid p^{k_1}\nu'_3(h)$ $(h \in H$, beliebig) ist. Nach endlichmaligem Anwenden dieses Schlusses erhält man schließlich eine Basis $\mathfrak{A}$ von F, so daß $p^k \mid \varrho_3(h)$ $(h \in H$, beliebig) gilt. $\varrho_3(h)$ sei dabei die entsprechende $n_2 \times n_1$-Teilmatrix aus

$$\varrho(h) = \begin{Vmatrix} \varrho_1(h) & \varrho_2(h) \\ \varrho_3(h) & \varrho_4(h) \end{Vmatrix} \qquad (h \in H),$$

die durch $\mathfrak{A}h = (\text{int}\,\varrho(h))\,\mathfrak{A}$ definiert ist. Das heißt, $\varrho_3(h)$ $(h \in H)$ ist die $n_2 \times n_1$-Nullmatrix über R_k. Da nach (1) eine Basisänderung des Darstellungsmoduls Δ_F mit der Äquivalenz der entsprechenden Darstellungen gleichbedeutend ist, muß ϱ zu μ äquivalent und demzufolge μ reduzibel sein.

b) Sei $\langle h \rangle$ eine zyklische Gruppe der Ordnung 3. Dann wird durch

$$\mu(h) := \begin{Vmatrix} \underline{4} & \underline{6} \\ \underline{6} & \underline{1} \end{Vmatrix}$$

eine treue Darstellung μ von $\langle h \rangle$ über dem Restklassenring modulo 9 erzeugt. Sicher ist μ^* reduzibel, denn es gilt

$$\mu^*(h) = (\mu(h))^* = \begin{Vmatrix} \underline{1} & \underline{0} \\ \underline{0} & \underline{1} \end{Vmatrix} .$$

Wir wollen nun nachweisen, daß μ irreduzibel ist. Damit wäre b) gezeigt. Falls μ reduzibel wäre, gäbe es einen eindimensionalen zulässigen Teilmodul $x_1 \mathfrak{w}_1 + x_2 \mathfrak{w}_2$ mit $\begin{pmatrix} \mathfrak{w}_1 \\ \mathfrak{w}_2 \end{pmatrix}$ als Basis des Darstellungsmoduls F vom Typ (9,9), und es gälte

$$(x_1 \mathfrak{w}_1 + x_2 \mathfrak{w}_2)h = f \cdot (x_1 \mathfrak{w}_1 + x_2 \mathfrak{w}_2) \quad (f, x_1, x_2 \in \mathbb{Z}) ,$$

wobei ord $(x_1 \omega_1 + x_2 \omega_2) = 9$ ist. Das hieße $x_i \not\equiv 0 (\mathrm{mod}\ 3)$ für mindestens ein x_i $(i=1,2)$. Sicher gilt

$$(4x_1 + 6x_2 - fx_1)\omega_1 + (6x_1 + x_2 - fx_2)\omega_2 = 0$$

und damit

$$4x_1 + 6x_2 - fx_1 \equiv 0 (\mathrm{mod}\ 9)$$

$$6x_1 + x_2 - fx_2 \equiv 0 (\mathrm{mod}\ 9) .$$

Zusammengefaßt ergibt sich

$$(4 - f)x_1 + 6x_2 \equiv 0 (\mathrm{mod}\ 9)$$

und

$$6x_1 + (1 - f)x_2 \equiv 0 (\mathrm{mod}\ 9) .$$

Falls $x_2 \not\equiv 0 (\mathrm{mod}\ 3)$ ist, müßte $3|(1-f)$ und somit $3|(4-f)$ gelten. Das gleiche würde der Fall sein, wenn $x_1 \not\equiv 0 (\mathrm{mod}\ 3)$ wäre. Daraus könnten wir schließen

$$\frac{4-f}{3} x_1 + 2x_2 \equiv 0 (\mathrm{mod}\ 3)$$

und

$$2x_1 + \frac{1-f}{3} x_2 \equiv 0 (\mathrm{mod}\ 3) ,$$

wodurch

$$(\frac{4-f}{3} \frac{1-f}{3} - 4)x_1 \equiv 0 (\mathrm{mod}\ 3)$$

und

$$(\frac{4-f}{3} \frac{1-f}{3} - 4)x_2 \equiv 0 (\mathrm{mod}\ 3)$$

folgen würde. Dieses ergäbe zwangsläufig

$$\frac{4-f}{3} \frac{1-f}{3} - 4 \equiv 0 (\mathrm{mod}\ 3) .$$

Die Zahl f hätte die Gestalt $f = 1+3g$ mit $g \in \mathbb{Z}$, und es ergäbe sich

$$\frac{4-1-3g}{3} \frac{-3g}{3} - 1 \equiv 0 (\mathrm{mod}\ 3) ,$$

d. h. $-(1 - g)g \equiv 1 (\mathrm{mod}\ 3)$.

Diese letzte Beziehung ist aber für kein $g \in \mathbb{Z}$ lösbar.

Satz 4: a) Sei F eine additiv geschriebene abelsche Gruppe vom Typ $(p^k, .., p^k)$ und ω der Spaltenvektor einer Basis von F. Des weiteren sei H eine Gruppe aus Automorphismen von F, und μ

sei die zu dem Darstellungsmodul F gehörende Matrixdarstellung von H in Δ_F bezüglich der Basis ω. Dann ist

$\mu^*: h \to \mu^*(h) := (\mu(h))^*$ die Darstellung von H über GF(p) auf jeder Faktorgruppe $\mathfrak{H}_1(F)/\mathfrak{H}_{i+1}(F)$ bezüglich der Basis

$(p^i \omega + \mathfrak{H}_{i+1}(F))$ $(i=0,..,k-1)$.

b) Wenn wir unter a) zusätzlich $p \nmid |H|$ voraussetzen, so ist $\mu(h)$ $(h \in H)$ genau dann die Einheitsmatrix über R_k, falls $\mu^*(h)$ die Einheitsmatrix über GF(p) ist.

Beweis: a) Es folgt unter Berücksichtigung der Tatsache, daß $\mathfrak{H}_{i+1}(F)$ $(0 \le i \le k-1)$ eine charakteristische Untergruppe von F ist,

$$(p^i \omega + \mathfrak{H}_{i+1}(F))h = p^i(\omega h) + \mathfrak{H}_{i+1}(F)$$
$$= p^i(\operatorname{int} \mu(h))\omega + \mathfrak{H}_{i+1}(F)$$
$$= (\operatorname{int} \mu(h))(p^i \omega + \mathfrak{H}_{i+1}(F))$$

für alle $h \in H$. Da $\mathfrak{H}_i(F)/\mathfrak{H}_{i+1}(F)$ elementarabelsch ist, erhalten wir, daß H auf $\mathfrak{H}_i(F)/\mathfrak{H}_{i+1}(F)$ die Darstellung μ^* bezüglich der Basis $(p^i \omega + \mathfrak{H}_{i+1}(F))$ bewirkt.

b) Falls $\mu(h) = E_n^k$ ist (n = Anzahl der Elemente einer Basis von F; E_n^k = Einheitsmatrix über R_k; E_n^1 = Einheitsmatrix über R_1 vom Grade n), dann gilt trivialerweise $\mu^*(h) = E_n^1$ mit $h \in H$. Wenn umgekehrt $\mu^*(h) = E_n^1$ für ein $h \in H$ ist, dann folgt, daß $\mu(h) = E_n^k + [p]_k B$ gilt (B gewisse $n \times n$-Matrix über R_k). Analog wie im Beweis zu Lemma 2.3 aus /3/, S. 22, erhält man $\mu(h)^{p^k} = E_n^k$. Da $p \nmid |H|$ ist, folgt daraus $\mu(h) = E_n^k$.

Literatur

/1/ Fuchs, L.: Abelian Groups. Budapest 1966

/2/ Huppert, B.: Endliche Gruppen. Bd. I. Berlin 1967

/3/ Steffen, G.: Irreduzible zyklische Matrixgruppen über dem Restklassenring mod p^k.
Rostock. Math. Kolloq. 8, 5 - 29 (1978)

/4/ Szele, T.: On direct decompositions of abelian groups.
J. London Math. Soc. 28, 247 - 250 (1953)

eingegangen: 28. 05. 1979

Anschrift des Verfassers:

Dr. Günther Steffen
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 31 - 47 (1979)

Dieter Schott

Zur Lösung linearer Gleichungen mit nichtvertauschbaren Operatoren

Zusammenfassung

Seit einiger Zeit bemüht man sich verstärkt um die Entwicklung von Operatorkalkülen zur Lösung von linearen Gleichungen mit nichtvertauschbaren Operatoren (siehe u. a. Ditkin, Prudnikov, Meller, Berz, Berg, Dimovski, Przeworska-Rolewicz, Tasche). Die vorliegende Arbeit ist im Anschluß an meine Dissertation /5/ entstanden und stellt einen Beitrag in dieser Richtung dar. Ausgangspunkt sind lineare Operatorgleichungen der speziellen Bauart

$$Q(S+d) = \sum_{i=0}^{n} q_i (S+d)^i y = f$$

mit invertierbarem $Q(S)$. Wie aus /5/ hervorgeht, besitzen sie die Lösung

$$y = r^{-1} Q^{-1}(S) r f,$$

falls r den Beziehungen

$$Sr - rS = rd, \quad rq_i = q_i r$$

genügt und invertierbar ist. Dieser Sachverhalt wird ausgenutzt, um bestimmte Typen von partiellen und rekursiven Differentialgleichungen mit variablen Koeffizienten zu lösen.

I. Allgemeine Vorbemerkungen

1. Gegeben seien die linearen Operatoren $S, U, q_0, q_1, \ldots, q_n$, die alle den (i. a. komplexen) Vektorraum $\underline{M}$ in sich abbilden. Wir betrachten die beiden Operatorgleichungen

$$Q(S)x = \sum_{i=0}^{n} q_i S^i x = f^* \quad (f^* \in \underline{M}), \tag{1}$$

$$Q(U)y = \sum_{i=0}^{n} q_i U^i y = f \qquad (f \in \underline{M}) \tag{2}$$

mit $n \geq 1$ und invertierbarem $Q(S)$. Uns interessiert die Frage, wann die Lösung y von (2) mit Hilfe der Lösung $x = Q^{-1}(S)f^*$ von (1) berechnet werden kann. Es liegt nahe, nach einem linearen Ähnlichkeitsoperator $r : \underline{M} \to \underline{M}$ mit

$$rU = Sr \tag{3}$$

zu suchen (vgl. /3/). Führt man die Differenz $d = U - S$ ein, so ist (3) äquivalent zu

$$r_S^{[1]} := Sr - rS = rd. \tag{3'}$$

Lemma 1: Hat die algebraische Differentialgleichung (3') eine invertierbare Lösung r, die mit allen q_i $(i=0,\ldots,n)$ vertauschbar ist, dann gilt

$$y = Q^{-1}(U)f = Q^{-1}(S+d)f = r^{-1}Q^{-1}(S)rf. \tag{4}$$

Beweis: Aus $r(S+d) = Sr$ folgt durch Induktion

$$r(S+d)^n = S^n r$$

und schließlich

$$rQ(S+d) = Q(S)r.$$

Da r und $Q(S)$ nach Voraussetzung invertierbar sind, existiert auch $Q^{-1}(S+d) = r^{-1}Q^{-1}(s)r$.

Lemma 2: Erfüllt r außer den in Lemma 1 geforderten Bedingungen noch die Beziehung $rd = dr$ und besitzt die algebraische Differentialgleichung

$$P_S^{[1]} := SP - PS = d \tag{5}$$

eine Lösung P, so ergibt sich für alle ganzen n

$$(r^n)_S^{[1]} = nP_S^{[1]} r^n = nr^n P_S^{[1]}, \tag{6}$$

d. h., r hat den Charakter eines algebraischen Exponentialelementes.

Beweis: Aus $rd = dr$ resultiert $Sr - rS = dr = rd$ bzw.

$Sr^{-1}-r^{-1}S = -dr = -rd$. Durch Induktion entsteht

$$Sr^n - r^nS = ndr^n = nr^nd \qquad (n \text{ ganz}).$$

Unter Beachtung von (5) erhält man daraus (6).

Definition 1: Jeden linearen Operator r, der invertierbar ist, mit q_i kommutiert und (für n=1) einer Gleichung (6) genügt, bezeichnen wir im weiteren mit e(P,S).
Offensichtlich ist e(P,S) nicht eindeutig bestimmt.
Die unter den genannten Voraussetzungen vorliegenden Exponentialeigenschaften von r erleichtern in vielen Fällen die Suche nach einem geeigneten Ähnlichkeitsoperator.

2. Für Umformungen von Operatorausdrücken benötigen wir noch eine allgemeine Vertauschungsbeziehung.

Lemma 3: Es sei v ein linearer Operator von $\underline{M}$ in $\underline{M}$ und $v_S^{[i]}$ die i-te algebraische Ableitung von v nach S, d. h.

$$v_S^{[i]} := Sv_S^{[i-1]} - v_S^{[i-1]}S, \quad v_S^{[o]} := v \qquad (i=1,2,\ldots).$$

Weiterhin sei q(S) ein Polynom in S(mit komplexen Koeffizienten) und $q^{(i)}(S)$ die i-te (formale) Ableitung von q(S) nach S mit $q^{(o)}(S) := q(S)$. Dann ist die Gleichung

$$vq(S) = \sum_{i\geq o} \frac{(-1)^i}{i!}\, q^{(i)}(S)v_S^{[i]} \tag{7}$$

erfüllt.

Beweis: Durch Induktion gewinnt man

$$vS^k = \sum_{i=o}^{k}(-1)^i\binom{k}{i}S^{k-i}v_S^{[i]} = \sum_{i\geq o}\frac{(-1)^i}{i!}\,(S^k)_{(i)}v_S^{[i]}$$

(vgl. /1: § 21/). Daraus folgt sofort die Behauptung.

Bemerkung: Ist T ein linearer Operator von $\underline{M}$ in $\underline{M}$ mit $ST - TS = E$ (E: Einheitsoperator) und v=v(T) ein Polynom in T (mit komplexen Koeffizienten), so gilt

$$v_S^{[i]} = v^{(i)}(T). \tag{7'}$$

Das ergibt sich aus (7), wenn man dort v durch S, S durch T und q(S) durch v(T) ersetzt. Man beachte dabei

$S_T^{[1]} = -\mathfrak{E}, \quad S_T^{[i]} = 0 \quad (i \geq 2).$

In konkreten Fällen trifft (7') oft auch für eine größere Funktionenklasse $\{v(T)\}$ zu.

II. Partielle Differentialgleichungen mit variablen Koeffizienten

1. Wir betrachten über einem geeigneten Raum $\underline{F}$ von (hinreichend oft total differenzierbaren, komplexwertigen) Funktionen mit nichtnegativen Argumenten t und x Gleichungen der Form

$$Q(\frac{\partial}{\partial t} + P_t(\frac{\partial}{\partial x},x,t))z(t,x) = f(t,x), \tag{8}$$

$$P(\frac{\partial}{\partial x},x,t) = \sum_{j=0}^{m} b_j(t,x)\frac{\partial^j}{\partial x^j}, \quad P_t(\frac{\partial}{\partial x},x,t) = \sum_{j=0}^{m}(\frac{\partial}{\partial t}b_j(t,x))\frac{\partial^j}{\partial x^j},$$

$$Q(\frac{\partial}{\partial t} + P_t) = \sum_{i=0}^{n} a_i(\frac{\partial}{\partial t} + P_t)^i, \quad b_j(t,x) \in \underline{F}, \quad a_i \text{ komplex.}$$

Unter Einbeziehung vorgegebener Anfangsbedingungen geht man in der Operatorenrechnung von (8) zu einer Gleichung

$$Q(s+P_t(w,x,t))z(t,x) = f(t,x)+g \tag{8'}$$

auf einem Distributionenbereich $\underline{DF}$ über. Dabei ist

$$g = g(s,w,z_{j0}(t,0),z_{0i}(0,x),z_{ji}(0,0)) \quad (i,j \leq n-1),$$

$$s^{-1}f(t,x) := \int_0^t f(\tau,x)d\tau, \quad w^{-1}f(t,x) := \int_0^x f(t,\xi)d\xi.$$

Die Umwandlung erfolgt mit Hilfe der Beziehungen

$$s^k w^l z(t,x) = z_{lk}(t,x) + s^k\sum_{j=0}^{l-1} w^{l-j}z_{j0}(t,0) + w^l\sum_{i=0}^{k-1} s^{k-i}z_{0i}(0,x)$$

$$-\sum_{j=0}^{l-1}\sum_{i=0}^{k-1} w^{l-j}s^{k-i}z_{ji}(0,0),$$

$$z_{ji}(t,x) := \frac{\partial^{i+j}}{\partial t^i \partial x^j} z(t,x) \qquad \text{(siehe /1: § 25/)},$$

$$sq(t,x)f = q(t,x)sf + q_t(t,x)f,$$

$$wq(t,x)f = q(t,x)wf + q_x(t,x)f,\ q(t,x) \in \underline{F},\ f(t,x) \in \underline{F}.$$

Existiert über $\underline{DF}$ oder über einer geeigneten Erweiterung von $\underline{DF}$ ein Element $e(P(w,x,t),s)$ (siehe Definition 1), so bekommt man wegen $P_t = P_s^{[1]}$ nach (4) die (eventuell schwache) Lösung

$$z(t,x) = Q^{-1}(s+P_t)(f+g) = e^{-1}(P,s)Q^{-1}(s)e(P,s)(f+g). \qquad (9)$$

Die Formeln für $Q^{-1}(s)$ sind bekannt (siehe etwa /1: § 9/).

2. An einem Spezialfall von (8) soll die prinzipielle Vorgehensweise bei der Lösung von (8) erläutert werden. Gegeben sei ein Anfangswertproblem der Gestalt

$$Q(\frac{\partial}{\partial t} + p'(t)\frac{\partial}{\partial x} + q'(t))z(t,x) = f(t,x) \qquad (10)$$

$(p(t),q(t)$ hinreichend oft differenzierbar, o.B.d.A. $p(0)=0)$,

$z_{j0}(t,0) = z_{1j}(t),\ z_{0i}(0,x) = z_{2i}(x) \quad (i,j=0,\ldots,n-1).$

Die Gleichung (8') lautet hier

$$Q(s+p'(t)w+q'(t))z(t,x) = f(t,x)+g. \qquad (10')$$

Aus (9) erhält man

$$z(t,x) = e^{-1}(pw+q,s)Q^{-1}(s)e(pw+q,s)(f+g). \qquad (11)$$

Es liegt nahe, ein Element $e(pw+q,s)$ in dem Operator e^{pw+q} zu vermuten, der über einem Funktionenbereich durch

$$e^{p(t)w+q(t)}f(t,x) = e^{q(t)}e^{p(t)w}f(t,x) := e^{q(t)}f(t,x+p(t))$$

erklärt wird. Da die Anwendung dieses Operators i. a. aber aus $\underline{F}$ hinausführt, muß zunächst eine geeignete Erweiterung von $\underline{F}$ vorgenommen werden. Hierzu legen wir fest:

$\underline{F}_+ := \{f(t,x) : f(t,x) \in \underline{F}$ für $x \geq 0,\ f(t,x) \equiv 0$ für $x < 0\}$,

$\underline{R} := \{p(t) : t \geq 0,\ p$ hinreichend oft differenzierbar, $p(0)=0\}$,

$\underline{F}_* := \{f(t,x) : t \geq 0,\ -\infty<x<\infty,\ \exists p(t)\in\underline{R}$ mit $f(t,x-p(t))\in \underline{F}_+\}$.

Der zu $\underline{F}$ gehörende Distributionenbereich ist

$$\underline{DF}_* := \{s^k w^l f(t,x) : f(t,x) \in \underline{F}_*\}$$

mit $s^{-1}f(t,x) := \int_0^t f(\tau,x)d\tau$, $w^{-1}f(t,x) := \int_{-\infty}^x f(t,\xi)d\xi$.

Jetzt definieren wir

$$e^{p(t)w}f(t,x) := f(t,x+p(t)) \quad (f(t,x) \in \underline{F}_*,\ p(t) \in \underline{R}), \tag{12}$$

$$e^{p(t)w}s^k w^l f(t,x) := \sum_{i\geq 0}\frac{(-1)^i}{i!}(s^k)^{(i)}w^l\left(\frac{\partial^i}{\partial t^i}e^{p(t)w}\right)f(t,x)$$

$$= \sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k-i}w^l\left(\frac{\partial^i}{\partial t^i}e^{p(t)w}\right)f(t,x), \tag{12'}$$

wobei $\frac{\partial^i}{\partial t^i}e^{p(t)w}$ die i-te formale Ableitung von $e^{p(t)w}$ nach t bedeutet. Die Festsetzung (12') ist mit (12) verträglich und wird durch (7), (7') nahegelegt. Auf eine Summe von Distributionen aus $\underline{DF}_*$ wird $e^{p(t)w}$ additiv fortgesetzt. Dann gilt

$$we^{pw} = e^{pw}w,\ e^{pw}e^{-pw} = e^{-pw}e^{pw} = E \text{ (E: Einheitsoperator)},$$

$$(e^{p(t)w}e^{q(t)})_s^{[1]} = (p(t)w + q(t))_s^{[1]}e^{p(t)w}e^{q(t)}$$

$$= e^{p(t)w}e^{q(t)}(p(t)w + q(t))_s^{[1]}$$

$$= (p'(t)w + q'(t))e^{p(t)w}e^{q(t)}.$$

Damit läßt sich (11) unter Beachtung der Definition 1 in der Form

$$z(t,x) = Q^{-1}(s+p'w+q')(f+g)$$

$$= e^{-q(t)}e^{-p(t)w}Q^{-1}(s)e^{p(t)w}e^{q(t)}(f+g) \tag{11'}$$

schreiben. Den Ausdruck f+g kann man in eine Summe von Ausdrükken der Gestalt $w^l h(t,x)s^k \cdot 1$ $(k,l \leq n;\ h(t,x) \in \underline{F}_+;$ $1 = 1(t,x) \in \underline{F}_+)$ zerlegen. Für $Q^{-1}(s+p'w+q')w^l h(t,x)s^k \cdot 1$ ist die Angabe konkreter Formeln leicht möglich. Wir wollen uns dabei der Einfachheit halber auf

$p'(t) > 0$ bzw. $p'(t) \equiv 0$, $Q(s) = s^n$

beschränken. Die im Falle $p'(t) \neq 0$ existierende Umkehrfunktion von $x = p(t)$ werde mit $t = \overline{p}(x)$ bezeichnet. Es erweist sich als günstig, $\overline{p}(x)$ durch die Vereinbarung $\overline{p}(x) \equiv 0$ für $x < 0$ auch in den negativen Argumentbereich fortzusetzen.

Satz 1: Mit $h(t,x) \in \underline{F}_+$ ist

$$(s+p'w+q')^{-n}h(t,x)=e^{-q(t)}\int_0^t \frac{(t-\tau)^{n-1}}{(n-1)!}\, e^{q(\tau)}h(\tau,x+p(\tau)-p(t))d\tau. \tag{13}$$

Beweis: Man berechnet

$$\begin{aligned}(s+p'w+q')^{-n}h(t,x)&=e^{-q(t)}e^{-p(t)w}s^{-n}e^{p(t)w}e^{q(t)}h(t,x)\\ &=e^{-q(t)}e^{-p(t)w}\left(\frac{t^{n-1}}{(n-1)!}*e^{q(t)}h(t,x+p(t))\right)\\ &=e^{-q(t)}\int_0^t\frac{(t-\tau)^{n-1}}{(n-1)!}e^{q(\tau)}h(\tau,x+p(\tau)-p(t))d\tau.\end{aligned}$$

Bemerkung: Für $p'(t) = a$ und $q'(t) = b$ ergibt sich speziell

$$(s+aw+b)^{-n}h = \frac{1}{(n-1)!}\int_0^t (t-\tau)^{n-1}e^{-b(t-\tau)}h(\tau,x-a(t-\tau))d\tau$$

$$= \frac{1}{(n-1)!}\int_0^t \tau^{n-1}e^{-b\tau}h(t-\tau,x-a\tau)d\tau \text{ (siehe /1:§26/).}$$

Da $h(t,x)$ für negative x verschwindet, kann man bei $p'(t)>0$ in (13) die untere Integrationsgrenze 0 durch $\overline{t}=\overline{p}(p(t)-x)$ ersetzen. Zur Veranschaulichung der folgenden Aussagen sollte man gelegentlich die δ-Schreibweise der Distributionen $s^{k+1}\cdot 1$ und $w^{l+1}\cdot 1$ heranziehen: $\delta^{(k)}(t)=s^{k+1}\cdot 1$; $\delta^{(l)}(x)=w^{l+1}\cdot 1$.

Satz 2: Mit $h(x)\in \underline{F}_+$ ist

$$e^{\pm pw}s^{k+1}w^l h(x)=\sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k+1-i}w^l\left(\frac{\partial^i}{\partial t^i}e^{\pm pw}\right)h(x)\Big|_{t=0}. \tag{14}$$

Beweis: Wir zeigen nur, daß (14) für das positive Vorzeichen vor p gilt. Weiterhin nehmen wir o.B.d.A. $l=0$ an. Aus

$$w^{-1}h(x+p) - s^{-1}p'h(x+p) = w^{-1}h(x)$$

folgt

$$sh(x+p) - p'wh(x+p) = se^{pw}h(x) - p'we^{pw}h(x) = sh(x).$$

Andererseits ist nach Definition

$$se^{pw}h(x) - p'we^{pw}h(x) = e^{pw}sh(x).$$

Damit stimmt (14) für k = 0. Mit Hilfe von

$$g(t)s^{k+1}\cdot 1 = \sum_{i=0}^{k}(-1)^i\binom{k}{i}g^{(i)}(0)s^{k+1-i}\cdot 1 \qquad (14')$$

($1 = 1(t,x) \in \underline{F}_+$, $g(t)$ k-mal differenzierbar)
verifiziert man

$$p'w\Big(\sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k+1-i}\Big(\frac{\partial^i}{\partial t^i}\,e^{pw}\Big)h(x)\Big)_{t=0}$$

$$= \sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k+1-i}\Big(\frac{\partial^{i+1}}{\partial t^{i+1}}\,e^{pw}\Big)h(x)\Big|_{t=0}.$$

Aus

$$e^{pw}s^{k+2}h(x) = e^{pw}s\cdot s^{k+1}h(x) = (se^{pw}-p'we^{pw})s^{k+1}h(x)$$

$$= se^{pw}s^{k+1}h(x) - p'we^{pw}s^{k+1}h(x)$$

resultiert dann durch Anwendung der vollständigen Induktion die Behauptung.

Satz 3: Für $h(t,x) \in \underline{F}_+$, $1(t,x) \in \underline{F}_+$ ist

$$e^{\pm pw}h(t,x)s^{k+1}\cdot 1 = \sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k+1-i}\Big(\frac{\partial}{\partial t} \pm p'w\Big)^i h(t,x)\Big|_{t=0}. \qquad (15)$$

Beweis: Zunächst gilt (vgl. (14'))

$$A = e^{\pm pw}h(t,x)s^{k+1}\cdot 1 = e^{\pm pw}\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}s^{k+1-j}\frac{\partial^j}{\partial t^j}h(t,x)\Big)_{t=0}.$$

Unter Benutzung von (14) gewinnt man daraus nach entsprechender Reihenumordnung

$$A = \sum_{i=0}^{k}(-1)^i\binom{k}{i}s^{k+1-i}\sum_{j=0}^{i}\binom{i}{j}\Big(\frac{\partial^{i-j}}{\partial t^{i-j}}\,e^{\pm pw}\Big)\frac{\partial^j}{\partial t^j}\,h(t,x)\Big|_{t=0}.$$

Wegen

$$\Big(\frac{\partial}{\partial t} \pm p'w\Big)^i h(t,x)\Big|_{t=0} = \sum_{j=0}^{i}\binom{i}{j}\Big(\frac{\partial^{i-j}}{\partial t^{i-j}}\,e^{\pm pw}\Big)\frac{\partial^j}{\partial t^j}\,h(t,x)\Big|_{t=0}$$

folgt die Behauptung.

Satz 4: Für $h(t,x) \in \underline{F}_+$, $1(t,x) \in \underline{F}_+$ und $n \geq k+1$ ist

$$(s+p'w+q')^{-n} h(t,x) s^{k+1} \cdot 1 \tag{16}$$

$$= e^{-q(t)} \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left(\frac{\partial}{\partial t} + p'(\tau)w\right)^i e^{q(\tau)} h(\tau, x-p(t)) \Big|_{\tau=0} \frac{t^{n+i-k-1}}{(n+i-k-1)!}.$$

Beweis: Mit Hilfe von (15) ergibt sich

$$(s+p'w+q')^{-n} h(t,x) s^{k+1} \cdot 1$$

$$= e^{-q(t)} e^{-p(t)w} s^{-n} e^{p(t)w} e^{q(t)} h(t,x) s^{k+1} \cdot 1$$

$$= e^{-q(t)} e^{-p(t)w} s^{-n} \left(\sum_{i=0}^{k} (-1)^i \binom{k}{i} s^{k+1-i} \left(\frac{\partial}{\partial t} + p'w\right)^i e^{q(t)} h(t,x) \right)_{t=0}$$

$$= e^{-q(t)} \sum_{i=0}^{k} (-1)^i \binom{k}{i} e^{-pw} \left(\left(\frac{\partial}{\partial t} + p'w\right)^i e^{q(t)} h(t,x) \right)_{t=0} s^{-(n+i-k-1)} \cdot 1$$

$$= e^{-q(t)} \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left(\frac{\partial}{\partial \tau} + p'(\tau)w\right)^i e^{q(\tau)} h(\tau, x-p(t)) \Big|_{\tau=0} \frac{t^{n+i-k-1}}{(n+i-k-1)!}.$$

Bemerkung: Verschwinden für $x = 0$ alle (gemischten) Ableitungen von $h(t,x)$ bis zur Ordnung $k-1$, so kann man in (15) und (16) w durch $\frac{\partial}{\partial x}$ ersetzen und erhält

$$\left(\frac{\partial}{\partial t} \pm p'w\right)^i h(t,x) \Big|_{t=0} = \left(\frac{\partial}{\partial t} \pm p' \frac{\partial}{\partial x}\right)^i h(t,x) \Big|_{t=0} = \frac{\partial^i}{\partial t^i} h(t, x \pm p(t)) \Big|_{t=0},$$

$$\left(\frac{\partial}{\partial \tau} + p'(\tau)w\right)^i e^{q(\tau)} h(\tau, x-p(t)) \Big|_{\tau=0} = \frac{\partial^i}{\partial \tau^i} \left(e^{q(\tau)} h(\tau, x+p(\tau)-p(t)) \right) \Big|_{\tau=0}.$$

Alle im weiteren vorkommenden Funktionen sollen für negative Argumente verschwinden.

Satz 5: Für $h(t,x) \in \underline{F}_+$ und $p'(t) > 0$ ist

$$(s+p'w+q')^{-n} w^{l+1} h(t,x)$$

$$= e^{-q(t)} w^l e^{q(\overline{p}(p-x))} h(\overline{p}(p-x), 0) \overline{p}'(p-x) \frac{(t-\overline{p}(p-x))^{n-1}}{(n-1)!}. \tag{17}$$

Beweis: Man berechnet

$$(s+p'w+q')^{-n}w^{l+1}h(t,x) = w^{l+1}(s+p'w+q')^{-n}h(t,x)$$

$$= w^{l+1}e^{-q(t)}\int_{\overline{p}(p-x)}^{t}\frac{(t-\tau)^{n-1}}{(n-1)!}e^{q(\tau)}h(\tau,x+p(\tau)-p(t))d\tau$$

$$= w^{l}e^{-q(t)}\frac{\partial}{\partial x}\int_{\overline{p}(p-x)}^{t}\frac{(t-\tau)^{n-1}}{(n-1)!}e^{q(\tau)}h(\tau,x+p(\tau)-p(t))d\tau$$

$$= e^{-q(t)}w^{l}e^{q(\overline{p}(p-x))}h(\overline{p}(p-x),0)\overline{p}'(p-x)\frac{(t-\overline{p}(p-x))^{n-1}}{(n-1)!}.$$

Wir kommen nun zu Anwendungen der aufgestellten Formeln. Wie schon erwähnt, führt der angegebene Kalkül i. a. auf schwache Lösungen. Interessiert man sich nur für klassische Lösungen, sind in der Regel an die rechten Seiten und an die Anfangswerte noch zusätzliche Bedingungen (wie Differenzierbarkeits- und Stetigkeitsforderungen, Verträglichkeitsbedingungen) zu stellen, die man der Lösungsdarstellung entnehmen kann.

2.1. Wir betrachten das Anfangswertproblem

$$z''(t)+2q'(t)z'(t)+(q''(t)+q'^{2}(t))z(t)=(\frac{\partial}{\partial t}+q'(t))^{2}z(t)=f(t),$$
$$z(0) = a,\ z'(0) = b. \qquad (18)$$

Es führt auf die Gleichung

$$(s^{2}+2q's+q''+q'^{2})z = (s+q')^{2}z = f+as^{2}\cdot 1+(2aq'+b)s\cdot 1. \qquad (18')$$

Mit Hilfe von (13) und (16) gelangt man zu der Lösung

$$z(t) = e^{-q(t)}\int_{0}^{t}(t-\tau)e^{q(\tau)}f(\tau)d\tau+e^{q(0)-q(t)}(a+q'(0)at+bt). \qquad (19)$$

Man kann darüber hinaus versuchen, auch Ausdrücke

$$s^{n} + a_{n-1}(t)s^{n-1} + \ldots + a_{1}(t)s + a_{0}(t)$$

höherer Ordnung in ein Produkt von Linearfaktoren

$$(s+b_{n-1}(t))(s+b_{n-2}(t))\ldots(s+b_{0}(t))$$

aufzuspalten, um auf der Grundlage der angeführten Formeln ent-

sprechende lineare Differentialgleichungen mit variablen Koeffizienten zu lösen (siehe dazu etwa /2/, /1: § 20/, /4/).

2.2. Gegeben sei das Anfangswertproblem

$$z_t(t,x) + p'(t)z_x(t,x) + q'(t)z(t,x)$$

$$= (\frac{\partial}{\partial t} + p'(t)\frac{\partial}{\partial x} + q'(t))z(t,x) = f(t,x), \qquad (20)$$

$$p'(t) > 0,\ z(t,0) = \varphi(t),\ z(0,x) = \psi(x),\ \varphi(0) = \psi(0).$$

Weiterhin mögen $f_x(t,x)$, $\varphi'(t)$ und $\psi'(x)$ existieren und stetig sein. Aus (20) folgt die Gleichung

$$(s+p'w+q')z(t,x) = f(t,x) + p'(t)\,\varphi(t)w\cdot 1 + \psi(x)s\cdot 1. \qquad (20')$$

Nach (13), (16) und (17) schließt man auf

$$z(t,x)=e^{-q(t)}\int_0^t e^{q(\tau)}f(\tau,x+p(\tau)-p(t))d\tau$$

$$+e^{-q(t)}e^{q(\overline{p}(p(t)-x))}\varphi(\overline{p}(p(t)-x))+e^{q(0)-q(t)}\psi(x-p(t)). \qquad (21)$$

Denn (16) geht für $n = 1$, $k = 0$ und $h(t,x) = \psi(x)$ unter Beachtung von $p(0) = 0$ in

$$(s+p'w+q')^{-1}\psi(x)s\cdot 1 = e^{q(0)-q(t)}\psi(x-p(t))$$

über. Formel (17) liefert schließlich für $n = 1$, $l = 0$ und $h(t,x) = \varphi(t)p'(t)$

$$(s+p'w+q')^{-1}p'(t)\,\varphi(t)w\cdot 1 = e^{-q(t)}e^{q(\overline{p}(p(t)-x))}\varphi(\overline{p}(p(t)-x)),$$

da (wegen $\frac{d}{d\lambda}p(\overline{p}(\lambda)) = p'(\overline{p}(\lambda))\overline{p}'(\lambda) \equiv 1$)

$$p'(\overline{p}(p(t)-x))\overline{p}'(p(t)-x) \equiv 1$$

gilt. Beachtet man, daß alle Funktionen für negative Argumente verschwinden sollen, werden durch (21) auch die geforderten Anfangswerte angenommen. Für

$$p'(t) \equiv 1,\ q'(t) = t,\ f(t,x) \equiv 0,\ \varphi(t) \equiv 0,\ \psi(0) = 0$$

lautet (21)

$$z(t,x) = e^{-t^2/2}\psi(x-t). \qquad (22)$$

Für

$$p'(t) \equiv 1,\ q'(t) = t,\ f(t,x) \equiv 0,$$

$$\varphi(t) = e^{-t^2/2}\,\hat{\varphi}(t),\quad \hat{\varphi}(0) = 0,\quad \psi(x) = 0$$

ergibt sich aus (21) speziell

$$z(t,x) = e^{-t^2/2}\,\hat{\varphi}(t-x). \tag{22'}$$

2.3. Wir wenden uns nun dem Anfangswertproblem 2. Ordnung

$$\left(\frac{\partial}{\partial t} + p'(t)\frac{\partial}{\partial x} + q'(t)\right)^2 z(t,x) = f(t,x), \tag{23}$$

$$p'(t) > 0,\ z(t,0) = \varphi_1(t),\ z(0,x) = \psi_1(x),\ \varphi_1(0) = \psi_1(0) = 0,$$

$$z_x(t,0) = \varphi_2(t),\ z_t(0,x) = \psi_2(x),\ \varphi_2'(0) = \psi_2'(0)$$

zu. Dabei sollen $f_{xx}(t,x)$, $\varphi_1''(t)$, $\varphi_2''(t)$, $\psi_1''(x)$, $\psi_2''(x)$, $p'''(t)$ und $q'''(t)$ existieren und stetig sein.
Zunächst gilt

$$\left(\frac{\partial}{\partial t}+p'\frac{\partial}{\partial x}+q'\right)^2 z = z_{tt}+2p'z_{tx}+p'^2 z_{xx}+2q'z_t+(p''+2p'q')z_x+(q''+q'^2)z = f.$$

Damit führt (23) auf die Gleichung

$$\begin{aligned}(s+p'w+q')^2 z = f &+ s^2\varphi_1(x)+s\psi_2(x)+w^2 p'^2(t)\varphi_1(t)+wp'^2(t)\varphi_2(t)\\ &+ 2p'(t)sw(\varphi_1(t)+\psi_1(x)-\varphi_1(0))\\ &+ 2q'(t)s\,\psi_1(x) + w(p''(t)+2p'(t)q'(t))\,\varphi_1(t).\end{aligned} \tag{23'}$$

Formt man den Ausdruck $2p'(t)sw(\varphi_1(t)+\psi_1(x)-\varphi_1(0))$ noch in $2w(p'(t)(\varphi_1(t)+\psi_1(x)-\varphi_1(0))s\cdot 1 + p'(t)\,\varphi_1'(t))$ um, gelingt mit Hilfe von (13), (16) und (17) die Lösungsdarstellung

$$z(t,x) = e^{-q(t)}\int_0^t e^{q(\tau)}(t-\tau)f(\tau,x+p(\tau)-p(t))\,d\tau \tag{24}$$

$$+ e^{q(0)-q(t)}(\psi_1(x-p)+q'(0)\varphi_1(x-p)t+\psi_2(x-p)t+p'(0)\frac{\partial}{\partial x}\psi_1(x-p)t)$$

$$+ e^{-q(t)}e^{q(\bar{p}(p-x))}(\varphi_1(\bar{p}(p-x))+(p'\varphi_2+q'\varphi_1+\varphi_1')(\bar{p}(p-x))(t-\bar{p}(p-x))).$$

Ist speziell $p' = a > 0$, $q' = b$, $\varphi_1(t) = t$, $\varphi_2(t) \equiv 1$,

$\varphi_1(x) = x$, $\varphi_2(x) \equiv 1$, erhält man daraus

$$z(t,x) = e^{-bt}\int_0^t (t-\tau)e^{b\tau}f(\tau,x-a(t-\tau))d\tau$$
$$+ e^{-bt}(t+x+btx-abt^2)1(x-at) \qquad (24')$$
$$+ e^{-bx/a}(t+x+(btx/a)-(bx^2/a^2))1(t-(x/a)).$$

(Die untere Integrationsgrenze 0 kann durch max (0,t-(x/a)) ersetzt werden.)
Die Funktionen (21) und (24) sind i. a. schwache Lösungen, die selbst dann existieren, wenn die Verträglichkeitsbedingungen für die Anfangswerte nicht gefordert werden. Allerdings sind dann alle (t,x) mit x = p(t) Unstetigkeitsstellen von z(t,x). Unter den angegebenen Bedingungen bekommt man in (21) und (24) stetige Lösungen, wenn man die Werte für x = p(t) nicht durch Einsetzen, sondern durch Grenzübergang bestimmt.

2.4. Das Anfangswertproblem (10) hat bei $Q(\lambda) = \lambda^n$ und verschwindenden Anfangswerten nach (13) die Lösung

$$z(t,x) = e^{-q(t)}\int_0^t \frac{(t-\tau)^{n-1}}{(n-1)!}\, e^{q(\tau)}f(\tau,x+p(\tau)-p(t))d\tau, \qquad (25)$$

wenn man f als n-mal (partiell) nach x stetig differenzierbare Funktion voraussetzt.

3. Unter Benutzung von (4) lassen sich außer (10) natürlich noch eine Reihe anderer Typen von partiellen Differentialgleichungen (mit variablen Koeffizienten) erfolgreich behandeln. Sei z. B.

$$Q(\tfrac{\partial}{\partial t} + p'(t)\tfrac{\partial}{\partial x} + q'(t)+R_t(t,x)+p'(t)R_x(t,x))z(t,x)=f(t,x) \qquad (26)$$

(p(t),q(t),R(t,x) hinreichend oft differenzierbar;
o.B.d.A. p(0)=0),
$z_{jo}(t,x) = z_{1j}(t)$, $z_{oi}(0,x) = z_{2i}(x)$ (i,j=0,...,n-1).

Die entsprechende Gleichung über dem Distributionenbereich hat die Gestalt

$$Q(s+p'w+q'+R_t+p'R_x)z = f+g. \tag{26'}$$

Aufgrund von $R^{[1]}_{s+p'w+q'} = R_t+p'R_x$ gelangt man zu

$$z(t,x) = e^{-1}(R,s+p'w+q')Q^{-1}(s+p'w+q')e(R,s+p'w+q')(f+g) \tag{27}$$

mit $Q^{-1}(s+p'w+q') = e^{-pw}e^{-q}Q^{-1}(s)e^{pw}e^{q}$. Dabei kann man $e(R,s+p'w+q')$ durch $e^{R(t,x)}$ realisieren, denn es gilt

$$(e^R)^{[1]}_{s+p'w+q'} = R^{[1]}_{s+p'w+q'}e^R = e^R R^{[1]}_{s+p'w+q'} = (R_t+p'R_x)e^R.$$

Wir untersuchen hierzu das Anfangswertproblem

$$z_t(t,x)+az_x(t,x)+(R_t(t,x)+aR_x(t,x))z(t,x)=f(t,x), \tag{28}$$

$$z(t,0)=\varphi(t),\ z(0,x)=\psi(x),\ \varphi(0)=\psi(0),\ a>0.$$

Wir wollen annehmen, daß f_t, f_x, $\varphi'(t)$ und $\psi'(x)$ als stetige Funktionen existieren. Aus (28) resultiert die Gleichung

$$(s+aw+R_t+aR_x)z = s\psi+aw\varphi+f \tag{28'}$$

und somit die Lösung

$$z(t,x) = e^{-R}e^{-aw}s^{-1}e^{aw}e^{R}(s\psi+aw\varphi+f). \tag{29}$$

Nach (13), (16) und (17) ermittelt man

$$z(t,x)=e^{-R(t,x)}\int_0^t e^{R(t-\tau,x-a\tau)}f(t-\tau,x-a\tau)d\tau \tag{29'}$$

$$+e^{-R(t,x)}(e^{R(0,x-at)}\psi(x-at)+e^{R(t-(x/a),0)}\varphi(t-(x/a))).$$

III. Rekursive Differentialgleichungen mit variablen Koeffizienten

In (8) dürfen durchaus auch innerhalb von P_t noch Verschiebungsoperatoren in t- und x-Richtung sowie Ableitungen $\frac{\partial}{\partial t}$ auftreten, ohne daß sich an der Lösungsformel (9) etwas ändert.

Daher ist etwa die Untersuchung von rekursiven Differentialgleichungen der Art

$$Q(\tfrac{\partial}{\partial t} + p'(t)u + q'(t))z_m(t) = f_m(t) \qquad (30)$$

$(m := [x];\ m = 0,1,2,\ldots;\ z_m(t) := z(t,[x]);\ f_m(t) := f(t,[x]);$

$uh_m(t) := h_{m-1}(t)$ mit $h_{-1}(t) \equiv 0;$

$p(t), q(t)$ hinreichend oft differenzierbar)

unter Einbeziehung entsprechender Anfangsbedingungen möglich. Über

$$Q(s+p'u+q')z_m(t) = f_m(t)+g \qquad (30')$$

gewinnt man

$$z_m(t) = e^{-1}(pu+q,s)Q^{-1}(s)e(pu+q,s)(f_m+g). \qquad (31)$$

Dabei sind etwa $e^{q(t)}e^{p(t)u}$ und $e^{-q(t)}e^{-p(t)u}$ mit

$$e^{\pm p(t)u} := \sum_{i=0}^{\infty} \frac{(\pm 1)^i}{i!} p^i(t)u^i$$

Operatoren der Gestalt $e(pu+q,s)$ bzw. $e^{-1}(pu+q,s)$.
Die prinzipielle Vorgehensweise ist der unter II. analog, so daß wir uns im weiteren auf Andeutungen beschränken.
Als einfaches Beispiel zu (30) wählen wir

$$z'_{m+1}(t)+p'(t)z_m(t)+q'(t)z_{m+1}(t) = f_{m+1}(t), \qquad (32)$$

$$z_m(0) = \omega_m,\ z_o(t) = \varphi(t),\ \omega_o = \varphi(0) = 0.$$

Die Funktion $\varphi(t)$ sei stetig differenzierbar. Nach Multiplikation der Gleichung (32) mit u und Anwendung einiger aus der Operatorenrechnung bekannter Formeln (siehe etwa /1: § 24/) erhält man daraus

$$(s+p'u+q')z_m = f_m(t)+s\,\omega_m+(\varphi'(t)+q'(t)\,\varphi(t)-f_o(t))d_m, \qquad (32')$$

$$d_m := (1,0,\ldots,0,\ldots).$$

Setzt man wieder o.B.d.A. $p(0) = 0$, entsteht die Lösung

$$z_m(t) = e^{-q(t)} \sum_{i+j=0}^{m} \frac{(-1)^i}{i!j!} p^i(t) \int_0^t e^{q(\tau)} p^j(\tau) f_{m-i-j}(\tau) d\tau \qquad (33)$$

$$+e^{-q(t)} \sum_{i+j=m} \frac{(-1)^i}{i!j!} p^i(t) \int_0^t e^{q(\tau)} p^j(\tau)(\varphi'(\tau)+q'(\tau)\varphi(\tau)-f_0(\tau))d\tau$$

$$+e^{q(0)-q(t)} \sum_{i=0}^{m} \frac{(-1)^i}{i!} p^i(t) \psi_{m-i}.$$

Auch die Erfassung allgemeinerer Gleichungstypen der Gestalt

$$Q(\frac{\partial}{\partial t} + p'(t)u + R_t(t,x))z(t,x) = f(t,x) \qquad (34)$$

mit $uh(t,x) := h(t,x-1)$ und $h(t,x) \equiv 0$ für $x < 0$ gelingt, falls $R(t,x)$ eine in x-Richtung 1-periodische Funktion darstellt. Hier gilt

$$Q(s+p'u+R_t) = e^{-R} Q(s+p'u) e^{R} = e^{-R} e^{-pu} Q(s) e^{pu} e^{R}. \qquad (35)$$

Literatur

/1/ Berg, L.: Operatorenrechnung I. Algebraische Methoden. Berlin 1972

/2/ Berz, E.: Lösung der linearen Differentialgleichung durch Transformation in einen Schiefkörper. Math. Z. 76, 174 - 198 (1961)

/3/ Meller, N. A.: Über einige Anwendungen der Operatorenrechnung auf Probleme der Analysis. Ž. Vyčisl. Mat. i Mat. Fiz. 3, 71 - 78 [russ.] (1963)

/4/ Schott, D.: Faltungs-Multiplikations-Operatoren. Diplomarbeit, Wilhelm-Pieck-Universität, Rostock 1972

/5/ Schott, D.: Identitäten in nichtkommutativen Ringen mit Anwendungen in der Operatorenrechnung. Diss. A, Wilhelm-Pieck-Universität, Rostock 1975

eingegangen: 14. 03. 1979

Anschrift des Verfassers:

Dr. rer. nat. Dieter Schott
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 49 - 58 (1979)

Lothar Berg

Zur stabilen Auflösung großer linearer Gleichungssysteme [1]

In den Arbeiten /2 - 5/ wurden große Gleichungssysteme mit einer Bandmatrix auf numerische Stabilität untersucht und in gewissen instabilen Fällen regularisierte Lösungen konstruiert. Für lineare Rekursionsformeln wurden entsprechende Untersuchungen bereits von F. W. J. Olver durchgeführt (vgl. /6/). Im vorliegenden Beitrag soll diskutiert werden, inwiefern die in den zitierten Arbeiten vorgeschlagenen Verfahren allgemein anwendbar sind, doch bleiben bis zur Aufstellung eines automatisch gesteuerten Rechenprogramms noch verschiedene praktische Probleme zu lösen.

1. Regularisierte Lösungen

Zur stabilen Auflösung gewisser Klassen von Gleichungssystemen wurden in /2 - 4/ Verfahren vorgeschlagen, deren Grundprinzipien hier zunächst in verallgemeinerter Form wiedergegeben werden. Gegeben sei ein großes Gleichungssystem

$$Ax = b \tag{1}$$

mit einer im allgemeinen rechteckigen Koeffizientenmatrix $A = (a_{ik})$ und dazu passenden Vektoren $x = (x_k)$, $b = (b_i)$. Weiterhin seien zwei Indexmengen Z und S ausgewählt, die Teilmengen der bei der Matrix A vorkommenden Zeilen- bzw. Spaltenindexmengen sind. Auf die Festlegung von Z und S werden wir weiter unten eingehen. Dann lassen sich die Zerlegungen

$$A = B + C + D + E, \quad x = y + z, \quad b = c + d$$

durchführen, die nach folgender Vorschrift zu bilden sind:

[1] Teil des Vortrags "Numerische Stabilität" vom 20. 4. 1979 anläßlich der 100. Wiederkehr des Gründungstages des mathematisch-physikalischen Seminars an der Wilhelm-Pieck-Universität Rostock.

B enthält die a_{ik} mit $i \notin Z$, $k \notin S$,
C enthält die a_{ik} mit $i \notin Z$, $k \in S$,
D enthält die a_{ik} mit $i \in Z$, $k \notin S$,
E enthält die a_{ik} mit $i \in Z$, $k \in S$,
y enthält die x_k mit $k \notin S$,
z enthält die x_k mit $k \in S$,
c enthält die b_i mit $i \notin Z$,
d enthält die b_i mit $i \in Z$,

während die nicht angeführten Elemente durch Nullen zu ersetzen sind. Wegen

$$Bz = 0, \quad Cy = 0, \quad Dz = 0, \quad Ey = 0$$

läßt sich das System (1) in der Form

$$By + Cz + Dy + Ez = c + d$$

schreiben und zerfällt offenbar in die beiden Systeme

$$By + Cz = c, \quad Dy + Ez = d. \tag{2}$$

Von jetzt ab denken wir uns die zuvor hinzugefügten Nullzeilen und -spalten wieder gestrichen und kennzeichnen den Übergang zu den neuen Matrizen und Vektoren jeweils durch ein Dach, wobei dieser Übergang durch Multiplikation mit geeigneten Auswahlmatrizen [2] verwirklicht werden kann. Unter der Voraussetzung, daß die Matrix $\hat{B}$ quadratisch und nichtsingulär ist, erhalten wir aus (2) nach Elimination von

$$\hat{y} = \hat{B}^{-1}\hat{c} - \hat{B}^{-1}\hat{C}\hat{z} \tag{3}$$

für $\hat{z}$ das Gleichungssystem

$$G\hat{z} = f \tag{4}$$

mit

$$G = \hat{E} - \hat{D}\hat{B}^{-1}\hat{C}, \quad f = \hat{d} - \hat{D}\hat{B}^{-1}\hat{c}.$$

---

[2] Eine Zeilenauswahlmatrix Q enthält in jeder Zeile genau und in jeder Spalte höchstens eine 1, während sie sonst nur aus Nullen besteht, sie wird von links multipliziert. Die zugehörige transponierte Matrix Q^T ist eine Spaltenauswahlmatrix und wird von rechts multipliziert.

Bei Verwendung von drei Normen mit der üblichen Eigenschaft

$$\|G\hat{z}\| \leq \|G\|\|\hat{z}\|$$

betrachten wir als nächstes das Funktional

$$F(\hat{z}) = \|G\hat{z} - f\| + \alpha\|\hat{z}\| \tag{5}$$

mit einem positiven Parameter α. Ist das System (4) instabil, so ist es in Anlehnung an A. N. Tichonov (vgl. /8/) zweckmäßig, die dann nur schwer numerisch bestimmbare (evtl. verallgemeinerte) Lösung von (4) durch denjenigen Vektor $\hat{z}$ zu ersetzen, der (5) minimiert und regularisierte Lösung von (4) genannt wird.

Hilfssatz 1 [3]: Unter der Voraussetzung

$$\|G\| < \alpha \tag{6}$$

nimmt das Funktional (5) sein Minimum für $\hat{z} = 0$ an.

Beweis: Aus $f = (f - G\hat{z}) + G\hat{z}$ folgt unmittelbar

$$\|f\| \leq \|f - G\hat{z}\| + \|G\|\|\hat{z}\|,$$

und hieraus ergibt sich wegen (5) und (6)

$$F(0) = \|f\| \leq F(\hat{z})$$

für alle $\hat{z}$, womit die Behauptung bewiesen ist.

Die große Bedeutung dieses Hilfssatzes besteht darin, daß die Minimalstelle $\hat{z} = 0$ des Funktionals (5) sowohl vom Regularisierungsparameter α als auch von den speziell gewählten Normen unabhängig ist, sofern (6) erfüllt ist, $\|G\|$ also hinreichend klein ist. Da die Kleinheit von $\|G\|$ ein gewisses Maß für die Instabilität des Systems (4) ist, kommt es beim Übergang von einem gegebenen instabilen System (1) zu der Zerlegung (3), (4) darauf an, die Instabilität von (1) in G zu lokalisieren. In diesem Fall ist es sinnvoll, den aus der regularisierten Lösung von (4) und aus (3), d. h. nach dem Hilfssatz aus

$$\hat{y} = \hat{B}^{-1}\hat{c}, \quad \hat{z} = 0, \tag{7}$$

sich ergebenden Vektor $x = y$ regularisierte Lösung von (1) zu nennen, falls $\hat{B}^{-1}$ "nicht zu große" Elemente enthält.

[3] In dieser allgemeinen Form wurde Hilfssatz 1 mit Beweis dem Verf. von Herrn K. Beyer mitgeteilt.

2. Ordnung der Störlösungen

Im folgenden sei die Matrix $A = (a_{ik})$ von (1) eine reelle, quadratische, nichtsinguläre Matrix mit $i,k = 1,\ldots,n$. Wir erweitern A zu einer rechteckigen Matrix $\underline{A}$, indem wir an der linken Seite p und an der rechten Seite q linear unabhängige Spalten hinzufügen mit $0 \leq p,q \leq n$. Die Zahlen p und q wurden eingeführt, damit sie speziell bei Bandmatrizen der Lage und der Breite des Bandes angepaßt werden können.

Definition: Jede Lösung $x = (x_k)$, $-q < k \leq n+p$, von $\underline{A}x = 0$ mit

$$\sum_{k=-q+1}^{0} x_k^2 + \sum_{k=n+1}^{n+p} x_k^2 = 1 \qquad (8)$$

heißt eine S t ö r l ö s u n g des zu (1) gehörenden homogenen Systems.

Wir legen jetzt p+q linear unabhängige Störlösungen $x_{.j} = (x_{kj})$, $1 \leq j \leq p+q$, durch die Forderungen

$$x_{k,j+q} = \delta_{kj} \text{ für } -q < k,j \leq 0 \text{ bzw. } 1 \leq k-n \leq p,$$

$$x_{k+n,j+q} = \delta_{kj} \text{ für } 1 \leq k,j \leq p \text{ bzw. } -q < k+n \leq 0$$

fest, was wegen $\det A \neq 0$ in eindeutiger Weise möglich ist. Dann läßt sich jede andere Störlösung s in der Form $s = Xu$ mit

$$X = (x_{kj}), \quad u = (u_j), \quad -q < k \leq n+p, \quad 1 \leq j \leq p+q$$

und $u^T u = 1$ darstellen (vgl. (8)).

Um die Störlösungen s der Größe nach zu ordnen, bestimmen wir die stationären Werte des Funktionals $s^T s$. Wegen der Nebenbedingung $u^T u = 1$ bedeutet dies nach der Methode von Lagrange, die stationären Werte von

$$uX^T Xu - \lambda u^T u$$

zu bestimmen. Durch Differentiation nach den Koordinaten von u gelangen wir zu dem Eigenwertproblem

$$X^T Xu = \lambda u$$

mit einer symmetrischen, positiv definiten Matrix. Letztere

besitzt bekanntlich p+q positive Eigenwerte λ_l, $1 \leq l \leq p+q$, die sich der Größe nach ordnen lassen:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p+q} > 0, \tag{9}$$

wobei die zugehörigen Eigenvektoren $u_{.l}$ ein Orthogonalsystem bilden. Wegen

$$u_{.j}^T X^T X u_{.l} = \lambda_l u_{.j}^T u_{.l} = \lambda_l \delta_{jl}$$

bilden dann auch die p+q Störlösungen

$$s_{.l} = X u_{.l} \tag{10}$$

ein Orthogonalsystem mit

$$s_{.l}^T s_{.l} = \lambda_l. \tag{11}$$

3. Stabilitätsdiskussion

Das Gleichungssystem (1) verhält sich bei einer numerischen Auflösung instabil, wenn das zugehörige homogene System "zu große" Näherungslösungen besitzt. Bis auf einen Faktor, der von der Größe der bei den Rechnungen vorkommenden Rundungsfehler abhängt, können wir die zuvor definierten Störlösungen als Modell für diese Näherungslösungen ansehen. In diesem Modell ist das maximale Normquadrat λ_1 der Störlösungen (vgl. (9) und (11)) ein sinnvolles Maß für die Größe der Instabilität. Will man eine vorhandene Instabilität beseitigen, so muß man durch geeignete Maßnahmen die größten Eigenwerte in (9) hinreichend verkleinern.

Hilfssatz 2: Es seien X und Y reelle Matrizen vom Format (t,m) bzw. (r,m) mit $r \leq m \leq t$ und jeweils maximalem Rang. Die stationären Werte des Funktionals $u^T X^T X u$ bezeichnen wir unter der Nebenbedingung $u^T u = 1$ mit λ_k und unter der zusätzlichen Nebenbedingung

$$Yu = 0 \tag{12}$$

mit μ_k, wobei diese Werte der Größe nach geordnet sein sollen:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0, \quad \mu_1 \geq \mu_2 \geq \ldots \geq \mu_{m-r} > 0.$$

Dann gilt für $k = 1,\ldots,m-r$

$$\mu_k \leq \lambda_k,$$

und im Spezialfall

$$Y^T = (u_{\cdot 1},\ldots,u_{\cdot r}), \tag{13}$$

wobei $u_{\cdot 1},\ldots,u_{\cdot r}$ die r ersten Eigenvektoren der Matrix X^TX sind, gilt sogar für dieselben k

$$\mu_k = \lambda_{k+r}.$$

Der Beweis dieses Hilfssatzes soll hier nicht weiter ausgeführt werden, da man ähnlich wie im vorhergehenden Punkt nur zu den Eigenwertproblemen $(X^TX-\lambda I)u = 0$ und

$$\begin{pmatrix} X^TX-\mu I & Y^T \\ Y & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = 0$$

überzugehen braucht, wobei $I = (\delta_{ij})$ mit $i,j = 1,\ldots,m$ eine Einheitsmatrix ist und $-2v$ der Vektor der zugehörigen Lagrange-Multiplikatoren. Die Behauptungen ergeben sich dann unmittelbar aus dem bekannten Courantschen Maximum-Minimum-Prinzip (vgl. etwa /7/).

Wir wenden jetzt Hilfssatz 2 auf die Matrix X von Abschnitt 2 mit $t = n+p+q$ und $m = p+q$ an. Zu diesem Zweck denken wir uns in Abhängigkeit von der Matrix A aus (1) eine kritische Zahl ß festgelegt, so daß wir für $\lambda_1 \geq ß$ von Instabilität und für $\lambda_1 < ß$ von Stabilität sprechen können. Im Fall

$$\lambda_1 \geq \ldots \geq \lambda_r \geq ß > \lambda_{r+1} \geq \ldots \geq \lambda_m \tag{14}$$

besteht unsere Aufgabe darin, die r größten Eigenwerte unter die Schranke ß hinabzudrücken. Nach Hilfssatz 2 erreichen wir dies, indem wir Y gemäß (13) wählen.

Ist der Abstand zwischen ß und λ_{r+1} in (14) groß, so benötigen wir nach dem Courantschen Maximum-Minimum-Prinzip bei der Wahl von Y nicht die genauen Eigenvektoren $u_{\cdot 1}$, sondern es genügen Näherungen dafür. Wie wir gleich sehen werden, ist es unter gewissen Voraussetzungen möglich, an Stelle von (13),

$$Y = \begin{pmatrix} s_{k_1 1} & \cdots & s_{k_1 m} \\ \vdots & & \vdots \\ s_{k_r 1} & \cdots & s_{k_r m} \end{pmatrix} \tag{15}$$

mit

$$\left|s_{k_j j}\right| = \operatorname*{Max'}_{1 \le k \le n} \left|s_{kj}\right| \tag{16}$$

für $j = 1,\ldots,r$ zu wählen, wo bei mehreren Maximalstellen jeweils ein geeignetes k_j auszuwählen ist und der Strich bei Max bedeutet, daß im Fall $j > 1$ die Werte $k = k_1,\ldots,k_{j-1}$ auszuschließen sind. Die Bedingung (12) geht dann, wenn wir für u die Entwicklung

$$u = \sum_{l=1}^{m} c_l u_{.l} \tag{17}$$

mit $c_l = u_{.l}^T u$ verwenden und (10) beachten, in das Gleichungssystem

$$\sum_{l=1}^{m} c_l s_{k_j l} = 0 \tag{18}$$

mit $j = 1,\ldots,r$ über. Aus der Umformung

$$c_j = -\frac{1}{s_{k_j j}} \left(\sum_{l=1}^{j-1} c_l s_{k_j l} + \sum_{l=j+1}^{m} c_l s_{k_j l}\right)$$

dieses Systems erkennen wir jetzt unter Beachtung von

$$\sum_{l=1}^{m} c_l^2 = u^T u = 1$$

schrittweise für $j = 1,\ldots,r$, daß die ersten Koeffizienten $c_1,\ldots,c_r$ von (17) klein sind, also die Orthogonalitätsrelationen (12) mit (13) näherungsweise erfüllt sind, wenn

$$\left|s_{k_j j}\right| \gg \left|s_{k_j l}\right| \tag{19}$$

für $1 \le j < l \le r$ und

$$\left|s_{k_j j}\right| \gg \left|c_l s_{k_j l}\right| \tag{20}$$

für $1 \le l < j \le r$ gilt.

In den Anwendungen kennt man weder die $u_{\cdot 1}$ noch die $s_{\cdot 1}$, so daß die vorhergehenden Überlegungen lediglich den Charakter eines Existenzbeweises haben. Tritt bei der numerischen Berechnung einer Lösung $x = (x_k)$ des Gleichungssystems (1) mit Hilfe irgendeiner Methode eine Instabilität auf, so ist es sinnvoll, k_1 an Stelle von (16) aus

$$\left|x_{k_1}\right| = \max_k \left|x_k\right|$$

zu bestimmen, wobei vielfach sogar ein Näherungswert genügen wird. Jetzt ist die Berechnung von x zu wiederholen, aber diesmal mit Hilfe der Methode von Abschnitt 1 mit $S = \{k_1\}$. Ist das Ergebnis immer noch instabil, so bestimme man analog wie zuvor k_2, wiederhole die Rechnung nach Abschnitt 1 mit $S = \{k_1, k_2\}$ usw., bis man befriedigende Werte erhält.

Hat man routinemäßig große Gleichungssysteme aufzulösen, so ist das soeben angedeutete Vorgehen vom Aufwand her vertretbar, falls Instabilitäten relativ selten vorkommen und falls man im instabilen Fall mit möglichst wenig Wiederholungen auskommt. Besonders angenehm sind diejenigen Fälle, wo man die gesuchten Werte k_j bereits nach Ausführung eines Teils des jeweiligen Rechenganges erkennt bzw. wo bereits bei einem einzigen Rechengang ersichtlich ist, wie man gleichzeitig mehrere k_j in sinnvoller Weise festlegen kann.

Abschließend bleibt nur noch etwas über die Bestimmung der Indexmenge Z aus Abschnitt 1 zu sagen. Bei einer schlecht konditionierten Matrix A hat die inverse Matrix A^{-1} "zu große" Elemente. Bei den in /1/ und /5/ untersuchten Bandmatrizen setzen sich die Elemente von A^{-1} aus den Lösungen der homogenen Gleichung $\underline{A}x = 0$ und den Lösungen der dazu adjungierten Gleichung zusammen. Hiernach ist zu erwarten, daß sich die Menge Z ganz analog wie zuvor bestimmen läßt, wenn man die vorhergehenden Rechnungen noch einmal mit A^T an Stelle von A durchführt. Für symmetrische Matrizen bedeutet dies einfach $Z = S$. Im allgemeinen Fall wird sich jedoch bei einer solchen Bestimmung von Z der Rechenaufwand verdoppeln, so daß die Erarbeitung einer einfacheren Methode zur Festlegung von Z als Problem offen bleibt. Möglicherweise besteht die Menge Z im Fall (14) aus

denjenigen r Indizes, für die in Analogie zu (16) die Koordinaten von $x_{\cdot m-r+1},\ldots,x_{\cdot m}$ betragsmäßig minimal werden. Vielleicht läßt sich bei der Bestimmung von Z auch der in /5/ hergestellte Zusammenhang zwischen den Lösungen von $\underline{A}x = 0$ und der zugehörigen adjungierten Gleichung ausnutzen.

4. Beispiele

Im Fall, daß $\underline{A}$ eine Bandmatrix mit konstanten Elementen in den Parallelen zur Hauptdiagonale ist bzw. die Voraussetzungen des Satzes von Poincaré und Perron erfüllt sind, findet man Beispiele zu den vorhergehenden Ausführungen in /3/ und /4/. Hierbei besteht entweder Z aus den r ersten und S aus den r letzten Indizes, oder es liegt der umgekehrte Fall vor, so daß die Instabilitäten jeweils am Rande auftreten.

In /2/ und /4/ wurde das instabile Beispiel

$$(6-7.2^{-i})x_{i-1} - (15-63.2^{-1-i})x_i + (6-7.2^{1-i})x_{i+1} = 0,$$

$x_0 = x_{n+1} = -1$, behandelt, bei dem die homogene Gleichung die allgemeine Lösung

$$x_i = c_1(2^{-i}-2^{-2i}) + c_2 2^i$$

besitzt. Hier ist nach den zitierten Arbeiten $Z = S = \{1\}$ zu wählen. Da die zugehörige adjungierte Gleichung

$$(6-7.2^{2-i})y_{i-1} - (15-63.2^{-1-i})y_i + (6-7.2^{-1-i})y_{i+1} = 0$$

lautet und die allgemeine Lösung

$$y_i = c_3(2^i + \tfrac{7}{2}) + c_4 \frac{2^i-1}{(3.2^i-7)(3.2^{i+1}-7)}$$

besitzt, bestätigt auch dieses Beispiel die vorhergehenden Ausführungen.

Es lassen sich auch leicht Beispiele für Bandmatrizen konstruieren, bei denen die Instabilitäten nicht am Rande, sondern im Innern auftreten. Man braucht nämlich nur homogene Differenzengleichungen aufzustellen, die Lösungen vom Typ

$$2^{-|i-n/2|},\ 2^{-|i-n/3|} + 2^{-|i-2n/3|}$$

usw. besitzen, und sie unter geeigneten Randbedingungen als Gleichungssysteme aufzufassen.

Literatur

/1/ Berg, L.: Auflösung von Gleichungssystmen mit einer Bandmatrix. Z. Angew. Math. Mech. 57, 373 - 380 (1977)

/2/ Berg, L.: Zur numerischen Stabilität des Gaußschen Algorithmus. Beiträge Numer. Math. 5, 19 - 25 (1976)

/3/ Berg, L.: Stable solution of instable systems of linear equations with a band matrix. Computing 20, 127 - 137 (1978)

/4/ Berg, L.: Regularization of instable boundary value problems for linear difference equations. Beiträge Numer. Math. 8 (im Druck)

/5/ Berg, L.: Stable right inverses of linear difference equations. Resultate der Mathematik (im Druck)

/6/ Cash, J. R.: An extension of Olver's method for the numerical solution of linear recurrence relations. Math. Comp. 32 No. 142, 497 - 510 (1978)

/7/ Collatz, L.: Eigenwertaufgaben mit technischen Anwendungen. Leipzig 1949

/8/ Tikhonov, A., et Arsénine, V.: Méthodes de résolution de problèmes mal posés (Übers. a. d. Russ.). Moscou 1976

eingegangen: 09. 03. 1979

Anschrift des Verfassers:

Prof. Dr. Lothar Berg
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 59 - 63 (1979)

Wolfgang Moldenhauer
Helmut Thielcke

Zur Pivotisierung bei der Auflösung linearer Gleichungssysteme

In der vorstehenden Arbeit /1/ werden regularisierte Lösungen großer instabiler linearer Gleichungssysteme

$$Ax = b \tag{1}$$

durch geeignete Zerlegungen in Teilsysteme

$$\begin{aligned} By + Cz &= c \\ Dy + Ez &= d \end{aligned} \tag{2}$$

gewonnen, wobei zwei Indexmengen Z - bestehend aus Zeilenindizes - und S - bestehend aus Spaltenindizes - die Zerlegung näher kennzeichnen. In /1/ wurde die Bedeutung dieser Indexmengen herausgearbeitet und gleichzeitig die Aufgabe gestellt, einfache Methoden zu ihrer Bestimmung zu entwickeln. Dieses Problem ist Gegenstand der vorliegenden Arbeit.

Bei der Verwendung direkter Verfahren zur Auflösung linearer Gleichungssysteme stellt sich die Situation als Reihenfolgeproblem der Auflösungsschritte dar (vgl. /2/, S. 38 - Austauschstrategie - bzw. Aufgabe 2.9., S. 43).

Wie wir am Beispiel demonstrieren, führt das dort angegebene Austauschverfahren mit Pivotisierung zu den gewünschten Indexmengen, wobei allerdings offen bleibt, ob es noch andere, günstigere Möglichkeiten gibt.

Es sei $A = (a_{ik})$ eine $n \times m$ Matrix, Z zunächst die Menge aller Zeilenindizes, S die Menge aller Spaltenindizes und β eine vorgegebene positive reelle Zahl.

Der Austauschalgorithmus besteht (in leicht abgeänderter Form) darin, daß bei jedem Schritt ein dem Betrag nach größtes Element a_{pq} der aktuellen Restmatrix G (beim ersten Schritt ist $G = A$) ausgewählt und die Ersetzungen

$$a_{ik} := a_{ik} - \frac{a_{iq}a_{pk}}{a_{pq}}, \quad b_i := -\frac{a_{iq}b_p}{a_{pq}}$$

für alle $i \neq p$, $k \in Z$,

$$Z := Z \setminus \{p\}, \quad S := S \setminus \{q\}$$

und

$$G := (a_{ik})$$

für $i \in Z$, $k \in S$ vorgenommen werden. Der Algorithmus wird abgebrochen, falls

$$|a_{ik}| < \beta \tag{3}$$

gilt für alle Elemente von G, bzw., falls eine der Mengen Z, S leer wird.

Für quadratische Matrizen A (n=m) vermerken wir die Relation

$$\det A = \pm \prod a_{pq} \det G, \tag{4}$$

wobei das Produkt über alle ausgewählten Pivot-Elemente a_{pq} zu bilden ist. Wegen der Maximalbedingung bei der Wahl der a_{pq} treten in vielen Fällen die Ungleichungen

$$|\det A| \ll |\prod a_{pq}|$$

und damit nach (4)

$$|\det G| \ll 1 \tag{5}$$

auf. Hieraus folgt unter Beachtung von (3)

$$\|G\| < K\beta, \tag{6}$$

wobei K eine positive Konstante ist, die von der speziell gewählten Matrixnorm abhängt. Interpretiert man die rechte Seite der Ungleichung (6) als Regulasisierungsparameter $\alpha = K\beta$, so ist (6) nichts anderes als die Voraussetzung von Hilfssatz 1 in /1/, und dieser Hilfssatz kann angewendet werden. Als Ergebnis liefert der Austauschalgorithmus nicht nur die Indexmengen Z und S, sondern nach einer Rückrechnung auch eine regularisierte Lösung

$$x_q = \begin{cases} b_p/a_{pq} \\ 0, q \in S \end{cases} \tag{7}$$

des Gleichungssystems (1) im Sinne der Arbeit /1/.

Da α klein ist, wird die Regularisierung in der angegebenen Form in der Regel erst für große Gleichungssysteme praktisch sinnvoll. Um den durch die Regularisierung bewirkten Effekt zu studieren, wenden wir den Austauschalgorithmus auf ein bekanntes Beispiel an (vgl. /3/ und /1/), bei dem bereits für kleine n (= 5, 9 bzw. 20) die numerischen Effekte deutlich werden. Während das erste Beispiel noch einen zu kleinen Umfang hat, erkennt man beim zweiten und erst recht beim dritten Beispiel in den ersten Komponenten bereits eine gute Annäherung an die exakte Lösung $x_i = 1$.

Das Gleichungssystem

$$\begin{pmatrix} 6 & 1 & 0 & 0 & 0 \\ 8 & 6 & 1 & 0 & 0 \\ 0 & 8 & 6 & 1 & 0 \\ 0 & 0 & 8 & 6 & 1 \\ 0 & 0 & 0 & 8 & 6 \end{pmatrix} \underline{x} = \begin{pmatrix} 7 \\ 15 \\ 15 \\ 15 \\ 14 \end{pmatrix}$$

mit $\underline{x} = (x_1 \ \dots \ x_5)^T$ wird durch den Austauschalgorithmus in die aufgelöste Form

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0.49219 \\ 8 & 0 & 0 & 0 & -0.96875 \\ 0 & 8 & 0 & 0 & 1.875 \\ 0 & 0 & 8 & 0 & -3.5 \\ 0 & 0 & 0 & 8 & 6 \end{pmatrix} \underline{x} = \begin{pmatrix} 0.49219 \\ 7.03125 \\ 9.875 \\ 4.5 \\ 14 \end{pmatrix}$$

übergeführt, d. h., mit den Indexmengen $Z = \{1\}$ und $S = \{5\}$ ist die regularisierte Lösung

$x^T = (0.8789 \ \ 1.2344 \ \ 0.5625 \ \ 1.75 \ \ 0)$ verbunden.

Zum Vergleich erhalten wir im Fall n = 9 das Ergebnis

$$\begin{pmatrix} 0 & & & & & & & & 0 & 0.03122 \\ 8 & \ddots & & & & & & & \ddots & -0.06238 \\ 0 & \ddots & \ddots & & & & & & \ddots & 0.12451 \\ & \ddots & \ddots & \ddots & & & & & \ddots & -0.24805 \\ & & \ddots & \ddots & \ddots & & & & \ddots & 0.49219 \\ & & & \ddots & \ddots & \ddots & & & \ddots & -0.96875 \\ & & & & \ddots & \ddots & \ddots & & \ddots & 1.875 \\ & & & & & \ddots & \ddots & 8 & 0 & -3.5 \\ 0 & & & \dots & & & 0 & 8 & 6 & \end{pmatrix} \underline{x} = \begin{pmatrix} 0.03122 \\ 7.93762 \\ 8.12451 \\ 7.75195 \\ 8.49219 \\ 7.03125 \\ 9.875 \\ 4.5 \\ 14 \end{pmatrix}$$

mit den Indexmengen Z = {1}, S = {9}, und die regularisierte Lösung lautet x^T = (0.99220 1.01556 0.96899 1.06152 0.87891 1.23437 0.5625 1.75 0).
Für n = 20 erhält man die Indexmengen Z = {1}, S = {20} und die regularisierte Lösung

$\underline{x}^T$ = (1 0.99999 1.00002 0.99997 1.00006 0.99988 1.00024 0.99951 1.00098 0.99805 1.00390 0.99220 1.01556 0.96899 1.06152 0.87891 1.23438 0.56250 1.75 0).

Normiert man im obigen Beispiel (n = 5) die Gleichungen, indem man etwa durch die Summe der Koeffizienten teilt, ändert man nicht die Lösungen, wohl aber das Stabilitätsverhalten und damit die regularisierte Lösung.

Das Gleichungssystem

$$\begin{pmatrix} \frac{6}{7} & \frac{1}{7} & 0 & 0 & 0 \\ \frac{8}{15} & \frac{6}{15} & \frac{1}{15} & 0 & 0 \\ 0 & \frac{8}{15} & \frac{6}{15} & \frac{1}{15} & 0 \\ 0 & 0 & \frac{8}{15} & \frac{6}{15} & \frac{1}{15} \\ 0 & 0 & 0 & \frac{8}{14} & \frac{6}{14} \end{pmatrix} \underline{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

wird durch den Austauschalgorithmus in die aufgelöste Form

$$\begin{pmatrix} \frac{6}{7} & 0 & 0 & 0 & -0.03348 \\ 0 & 0 & 0 & 0 & -0.04375 \\ 0 & \frac{8}{15} & 0 & 0 & 0.125 \\ 0 & 0 & \frac{8}{15} & 0 & -0.23333 \\ 0 & 0 & 0 & \frac{4}{7} & 0.42857 \end{pmatrix} \underline{x} = \begin{pmatrix} 0.82366 \\ -0.04375 \\ 0.65833 \\ 0.3 \\ 1 \end{pmatrix}$$

übergeführt. Den Indexmengen Z = {2} und S = {5} ist die regularisierte Lösung

x^T = (0.96094 1.2344 0.5625 1.75 0) zugeordnet.
Behält man dagegen im normierten Beispiel die ursprünglichen

Indexmengen $Z = \{1\}$ und $S = \{5\}$ bei, wird $G = 0.07031 > 0.04375$ und daher eine im Sinne von /1/ weniger geeignete Lösung ermittelt.

Folgendes Beispiel zeigt jedoch, daß der Austauschalgorithmus nicht in jedem Fall zur Abspaltung einer Restmatrix minimaler Norm führen muß:

$$\begin{pmatrix} 1+\varepsilon & 1 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1+\varepsilon & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & \frac{2\varepsilon}{1+\varepsilon} \end{pmatrix}, \quad \begin{pmatrix} \varepsilon & 0 & 0 \\ 0.5 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Die erste Matrix ist gegeben, die zweite ist die nach dem Austauschalgorithmus umgeformte Matrix und die dritte ist nach einer anderen Reihenfolge der Auflösungsschritte entstanden. Für alle $0 < \varepsilon < 1$ ist aber $\varepsilon < \frac{2\varepsilon}{1+\varepsilon}$.

Falls man allerdings die Zeilen wie beim vorhergehenden Beispiel normiert, ist die Pivotisierung auch hier wiederum erfolgreich durchführbar.

Literatur

/1/ Berg, L.: Zur stabilen Auflösung großer linearer Gleichungssysteme. Rostock. Math. Kolloq. 12, 49 - 58 (1979)

/2/ Kiesewetter, H., und Maeß, G.: Elementare Methoden der numerischen Mathematik. Berlin 1974

/3/ Tanabe, K.: Projection method for solving a singular system of linear equations and its applications. Numer. Math. 17, 203 - 214 (1971)

eingegangen: 02. 05. 1979

Anschrift der Verfasser:

Dr. Wolfgang Moldenhauer
DDR-25 Rostock 9
Landreiterstr. 8

Dr. Helmut Thielcke
DDR-252 Rostock
Dr. J.-Dieckmann-Str. 1

Rostock. Math. Kolloq. 12, 65 - 76 (1979)

Hans-Jürgen Albrand

l_1-Approximation und angepaßte Iterationsverfahren

Da die meisten Iterationsverfahren zur Lösung linearer Gleichungssysteme (Jacobi-, Gauß-Seidel-, u. a. Verfahren) von der formalen Vorschrift her zu speziell ausgerichtet sind und eine brauchbare Bestimmung eines Relaxationsfaktors in der Regel mit großem Aufwand verbunden ist und unter Umständen keine oder nur geringe Verbesserungen mit sich bringt, soll in dieser Arbeit ein Vorschlag zur Konstruktion angepaßter Iterationsverfahren gemacht werden.
Diese Anpassung erfolgt so, daß zu einem gegebenen linearen Gleichungssystem die in der Iterationsmatrix auftretenden Elemente mit Hilfe von geeigneten l_1-Approximationsaufgaben bestimmt werden.
In der Darstellung folgen wir der Arbeit /1/.

1. Das verallgemeinerte stationäre Richardson-Verfahren

Soll ein vorgelegtes lineares Gleichungssystem $A\,x = b$, A vom Format $n \times n$, $\det A \neq 0$, $A = (a_{ij})$, a_{ij} reell, iterativ in der Form

$$x^{s+1} := T\,x^s + v \tag{1}$$

gelöst werden, so entsteht die Frage nach einer geeigneten Iterationsmatrix T. Dabei sind verschiedene Grade der "Anpassung" von T an das gegebene System $A\,x = b$ möglich. Die Iterationsmatrizen des Jacobi- und des Gauß-Seidel-Verfahrens stellen bekanntlich nur in ganz speziellen Fällen (z. B. wenn A diagonaldominant ist) eine genügend gute Anpassung dar. Manchmal kann die Anpassung mit Hilfe eines Relaxationsfaktors weiter verbessert werden.
In dieser Arbeit wird mit einer nichtsingulären Parametermatrix $G = (g_{ij})$, g_{ij} reell, der folgende, z. B. auch in /2/, /3/ auf-

tretende, allgemeine Ansatz für T gewählt:

$$T := I - G A.$$

Hierbei bedeutet I die Einheitsmatrix. Durch geeignete Wahl von G kann man jedes konkrete Iterationsverfahren der Form (1) erhalten. Aus $A x = b$ folgt

$$x = x - G A x + G b = (I - G A) x + G b$$

und hieraus die Iterationsvorschrift

$$x^{s+1} = (I - G A) x^s + G b. \tag{2}$$

Die Vorschrift (2) geht in (1) über für $G := (I - T) A^{-1}$.
Der Ansatz (2), in /4/ als GRF-Methode bezeichnet, ist eine Verallgemeinerung der stationären Richardson-Methode, bei der mit einer festen Zahl p

$$G = p I$$

gilt. Die stationäre Richardson-Methode ist ein Spezialfall der Methode von Richardson /5/, bei der p von Iterationsschritt zu Iterationsschritt variieren kann.
Es bezeichne T_L die linke, untere Dreiecksmatrix von T.
Zum Gesamtschrittverfahren (2) kann man z. B. das folgende Einzelschrittverfahren definieren:

$$x^{s+1} = (I - T_L)^{-1} (T - T_L) x^s + (I - T_L)^{-1} G b. \tag{3}$$

Zur Abkürzung setzen wir

$$S := (I - T_L)^{-1} (T - T_L),$$

$$M := (I - T_L)^{-1} G.$$

Mit $G_k := (g_{ij}^{(k)})$, $g_{kj}^{(k)} := g_{kj}$ für $j = 1,\ldots,n$, $g_{ij}^{(k)} = 0$ für $i \neq k$, erhalten wir für S und M auch die folgenden Darstellungen

$$\begin{aligned} S &= (I - G_n A)(I - G_{n-1} A) \ldots (I - G_1 A), \\ M &= G_n + (I - G_n A) G_{n-1} + \ldots + (I - G_n A)\ldots(I - G_2 A) G_1. \end{aligned} \tag{4}$$

Für die Restvektoren $r^s := b - A x^s$ ergibt sich aus (2) die

Beziehung

$$r^{s+1} = (I - A\,G)\,r^s,\quad s = 1,2,\ldots, \tag{5}$$

für das Gesamtschrittverfahren und aus (3) die entsprechende Beziehung für das Einzelschrittverfahren

$$r^{s+1} = R\,r^s,\quad s = 1,2,\ldots, \tag{6}$$

mit

$$R = (I - A\,G_n)\,(I - A\,G_{n-1})\ \ldots\ (I - A\,G_1). \tag{7}$$

Wählen wir $g_{ii} := a_{ii}^{-1}$, $i = 1,\ldots,n$, $g_{ij} = 0$ für $i \neq j$, so geht (2) in die Punkt-Jacobi- und (3) in die Punkt-Gauß-Seidel-Vorschrift über.
Hinreichend für die Konvergenz von (2) bzw. (3) sind die Ungleichungen

$$\|I - G\,A\| < 1 \tag{8}$$

oder

$$\|I - A\,G\| < 1 \tag{9}$$

bzw.

$$\|S\| < 1 \tag{10}$$

oder

$$\|R\| < 1. \tag{11}$$

Aus (8) bzw. (9) folgt, daß dann G A regulär sein muß, d. h., A und G müssen regulär sein.
Nach Wahl einer Matrixnorm kann man versuchen, die g_{ij} so zu wählen, daß wenigstens eine der Matrixnormen (8) - (11) möglichst klein wird, jedenfalls kleiner als Eins.
Für $G = A^{-1}$ führen (2) und (3) trivialerweise in einem Schritt zum Ziel. Im allgemeinen Fall stellt G in gewissem Sinne eine Approximation für A^{-1} dar, ist also eine Art Ersatzinverse.
Die Parametermatrix G ist der Koeffizientenmatrix A so anzupassen, daß sich auch praktisch verwertbare Konvergenzaussagen für (2) oder (3) ergeben. Für ein spezielles Gleichungssystem stellt die Wahl der g_{ij} ein Anfangsproblem dar, das zunächst

gelöst werden muß, bevor mit der iterativen Lösung nach (2) bzw. (3) begonnen werden kann.

2. l_1-Strategien zur Wahl der Parametermatrix G

Wir betrachten das Verfahren (2) und die monotone Abnahme des Restvektors r^s in der l_1-Norm, d. h., mit (5) wird die Ungleichung

$$\|I - A\,G\|_S < 1 \tag{12}$$

angestrebt, wobei $\|\cdot\|_S$ die Spaltensummennorm bezeichnet. Die Ungleichung (12) führt auf die Bestimmung der g_{ij} so, daß

$$\max_k \left(\left|1 - \sum_{j=1}^{n} a_{kj}\,g_{jk}\right| + \sum_{\substack{j=1\\ j\neq k}}^{n} \left|\sum_{s=1}^{n} a_{js}\,g_{sk}\right|\right) < 1 \tag{13}$$

gilt. Der i-te Spaltenvektor der Matrix A sei a^i, der von I sei e^i. Dann können wir statt (13) auch

$$\left\|e^k - \sum_{j=1}^{n} g_{jk}\,a^j\right\|_1 < 1,\ k = 1,\ldots,n, \tag{14}$$

schreiben, wobei $\|\cdot\|_1$ die l_1-Norm bezeichnet.

Wenn die g_{ij} als Koeffizienten der besten Approximation gewählt werden, so ergibt sich $G = A^{-1}$. Der Aufwand zur Bestimmung der besten l_1-Approximationen ist jedoch zu groß. Man kann aber einen Algorithmus zur Lösung der l_1-Approximationen

$$\min_{g_{ij}} \left\|e^j - \sum_{i=1}^{n} g_{ij}\,a^i\right\|_1 \tag{15}$$

bereits dann abbrechen, wenn (14) eingetreten ist.
Das Iterationsverfahren (2) wird um so schneller konvergieren, je kleiner die links in den Ungleichungen (14) stehenden Größen ausfallen.

2.1. Die einfache l_1-Strategie

Die einfachste Wahl der g_{ij} besteht darin, in jeder Spalte von G jeweils nur ein $g_{ij} \neq 0$ zu wählen. Diese Strategie soll zunächst verfolgt werden.
Es sei $g_{jk} = 0$ für $j \neq k_1$; $k = 1,\ldots,n$. Man versucht jetzt also, e^k durch $g_{k_1 k}\, a^{k_1}$ zu approximieren. Dabei ist

$$\min_{g_{k_1 k}} \|e^k - g_{k_1 k}\, a^{k_1}\|_1 = \min\,(1, \|e^k - a_{kk_1}^{-1}\, a^{k_1}\|_1),$$

sofern $a_{kk_1} \neq 0$ ist. Wegen

$$\|e^k - a_{kk_1}^{-1}\, a^{k_1}\|_1 = \sum_{\substack{j=1 \\ j\neq k}}^{n} \left|\frac{a_{jk_1}}{a_{kk_1}}\right| \quad .$$

ist dieser Ausdruck kleiner als Eins, wenn

$$\sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{jk_1}| < |a_{kk_1}| \tag{16}$$

gilt. Falls (16) für $k = 1,2,\ldots,n$ erfüllt ist, haben wir mit dieser einfachen Strategie Erfolg, und (16) stellt das übliche Spaltensummenkriterium dar. Das entstandene Iterationsverfahren ist, nach eventueller Umnummerierung der Variablen, Vertauschung von Zeilen und Spalten, das gewöhnliche Jacobi-Verfahren. Die einfache l_1-Strategie führt also auf das Jacobi-Verfahren, wenn (16) gilt.

2.2. Die zweifache l_1-Strategie

Die nächsteinfache Strategie ist die Zweier-Strategie:
Man versucht die l_1-Approximation von e^k mit jeweils zwei Spaltenvektoren der Matrix A, etwa mit a^{k_1}, a^{k_2}.
Dann sind also die folgenden Approximationsaufgaben zu betrachten:

$$E_k := \min \left\| e^k - g_{k_1 k}\, a^{k_1} - g_{k_2 k}\, a^{k_2} \right\|_1, \quad k = 1,2,\ldots,n. \tag{17}$$

Im Fall $E_k < 1$ ergeben sich die Koeffizienten der besten Approximation $g_{k_1 k}$, $g_{k_2 k}$ aus einem der $n - 1$ linearen Gleichungssysteme

$$\begin{aligned} g_{k_1 k}^{(s)}\, a_{k k_1} + g_{k_2 k}^{(s)}\, a_{k k_2} &= 1, \\ g_{k_1 k}^{(s)}\, a_{s k_1} + g_{k_2 k}^{(s)}\, a_{s k_2} &= 0, \end{aligned} \tag{18}$$

$s = 1,\ldots,n;\ s \neq k$. Aus (18) folgt

$$g_{k_1 k}^{(s)} = \frac{a_{s k_2}}{\begin{vmatrix} a_{k k_1} & a_{k k_2} \\ a_{s k_1} & a_{s k_2} \end{vmatrix}}, \quad g_{k_2 k}^{(s)} = \frac{-a_{s k_1}}{\begin{vmatrix} a_{k k_1} & a_{k k_2} \\ a_{s k_1} & a_{s k_2} \end{vmatrix}}, \tag{19}$$

falls die auftretenden Determinanten nicht Null sind. Mit

$$m_k := \min_{s \neq k} \sum_{\substack{i=1 \\ i \neq k}}^{n} \left| g_{k_1 k}^{(s)}\, a_{i k_1} + g_{k_2 k}^{(s)}\, a_{i k_2} \right| \tag{20}$$

wird

$$E_k = \min\,(1,\ m_k),\ k = 1,\ldots,n.$$

Die hinreichende Konvergenzbedingung lautet jetzt

$$m := \max_k m_k < 1. \tag{21}$$

Wird das Minimum in (20) bei s_k, $k = 1,\ldots,n$, angenommen, so lautet (21) ausführlich:

$$\sum_{\substack{i=1 \\ i \neq k}}^{n} \left| \det \begin{pmatrix} a_{i k_1} & a_{i k_2} \\ a_{s_k k_1} & a_{s_k k_2} \end{pmatrix} \right| < \left| \det \begin{pmatrix} a_{k k_1} & a_{k k_2} \\ a_{s_k k_1} & a_{s_k k_2} \end{pmatrix} \right|, \tag{22}$$

$k = 1,2,\ldots,n$. Die Koeffizienten der besten Approximation sind

$$g_{k_1 k} = g_{k_1 k}^{(s_k)}, \quad g_{k_2 k} = g_{k_2 k}^{(s_k)}, \quad k = 1,\ldots,n. \tag{23}$$

Praktisch wird man so vorgehen, daß man sich zwei Spalten a^{k_1}, a^{k_2} von A sucht mit der Eigenschaft, daß in der n x 2 Matrix $(a^{k_1}\ a^{k_2})$ eine 2 x 2 Matrix

$$\begin{pmatrix} a_{kk_1} & a_{kk_2} \\ a_{s_k k_1} & a_{s_k k_2} \end{pmatrix}$$

existiert, deren Determinante einen so großen Betrag hat, daß (22) gilt. Auf diese Weise sind dann die Parameter $g_{k_1 k}$, $g_{k_2 k}$ gemäß (23) und (19) festgelegt.

Bei dem folgenden Beispiel kann eine Bestimmung der Parameter g_{ij} auf diesem Wege erfolgen.

Es sei $a_{ik} := (i + k - 1) \bmod n$; $i,k = 1,2,\ldots,n$, d. h.

$$A = \begin{pmatrix} 1 & 2 & \ldots & n-1 & n \\ 2 & 3 & \ldots & n & 1 \\ \ldots & & & & \\ n & 1 & \ldots & n-2 & n-1 \end{pmatrix}.$$

Es ist klar, daß die einfache l_1-Strategie hier nicht den erwünschten Erfolg bringen kann. Ganz anders sieht es aber schon mit der Zweier-Strategie aus. Durch die Strategie der Zweier-Anpassung hat man eine gute Möglichkeit, das Iterationsverfahren auf die Struktur des Gleichungssystems zuzuschneiden.

Zur Approximation von e^1 nehmen wir die Spalten a^1, a^n. Mit $s_1 := n$ wird

$$\begin{vmatrix} a_{11} & a_{1n} \\ a_{n1} & a_{nn} \end{vmatrix} = n - 1 - n^2$$

und

$$\sum_{i=2}^{n} \left| \det \begin{pmatrix} a_{i1} & a_{in} \\ a_{n1} & a_{nn} \end{pmatrix} \right| = \sum_{i=2}^{n-1} (n-i) = (n-1)(n-2)/2.$$

Zur Approximation von e^2 verwenden wir die n-te und die (n-1)-te Spalte von A. Mit $s_2 := 1$ erhalten wir

$$\begin{vmatrix} a_{2n} & a_{2,n-1} \\ a_{1n} & a_{1,n-1} \end{vmatrix} = n - 1 - n^2$$

und

$$\sum_{\substack{i=1 \\ i\neq 2}}^{n} \left| \det \begin{pmatrix} a_{in} & a_{i,n-1} \\ a_{1n} & a_{1,n-1} \end{pmatrix} \right| = (n-1)(n-2)/2.$$

Allgemein verwenden wir zur Approximation von e^k, $k \geq 2$, die $(n-k+2)$-te und die $(n-k+1)$-te Spalte von A und setzen $s_k := k-1$. Dann wird analog wie zuvor

$$\begin{vmatrix} a_{k,n-k+2} & a_{k,n-k+1} \\ a_{k-1,n-k+2} & a_{k-1,n-k+1} \end{vmatrix} = n - 1 - n^2$$

und

$$\sum_{\substack{i=1 \\ i\neq k}}^{n} \left| \det \begin{pmatrix} a_{i,n-k+2} & a_{i,n-k+1} \\ a_{k-1,n-k+2} & a_{k-1,n-k+1} \end{pmatrix} \right| = (n-1)(n-2)/2.$$

Wir erhalten also

$$\|I - A\,G\|_S = m = m_k = (2 + \frac{2(2n-1)}{(n-1)(n-2)})^{-1}. \qquad (24)$$

Insbesondere ist damit der Spektralradius $\varrho(I - A\,G)$ der Iterationsmatrix $I - A\,G$ für jedes n kleiner als 1/2:

$$\varrho(I - A\,G) < 1/2.$$

Die Parametermatrix G lautet explizit

$$G = (n-1-n^2)^{-1} \begin{pmatrix} n-1 & & & & -n \\ & \ddots & & \iddots & \\ & & -n & & n-1 \\ & -n & n-1 & & \\ -n & n-1 & & & \end{pmatrix}.$$

3. Ergänzende Bemerkungen

1. Bei der p-fachen l_1-Strategie werden zur Approximation von e^k p linear unabhängige Spaltenvektoren von A verwendet. Aus

(14) wird dann

$$\min_{g_{k_i k}} \| e^k - \sum_{i=1}^{p} g_{k_i k} a^{k_i} \|_1 =: m_k < 1.$$

Auf die allgemeinen Ausführungen zur p-fachen Strategie sei aus Platzgründen verzichtet. Eine solche Darstellung würde analog zur Darstellung für p = 2 verlaufen.
Man kann natürlich auch eine gemischte Strategie verwenden, die zur l_1-Approximation von e^k p_k Spaltenvektoren von A benutzt, $k = 1,2,\ldots,n$. Bei manchen Matrizen wird man gezwungen sein, eine größere Anzahl von Spaltenvektoren a^{k_i}, $i = 1,\ldots,p$, zur Approximation von e^k zu verwenden, um $m_k < 1$ zu erreichen. Hier muß man dann auf Näherungsverfahren zur Lösung der l_1-Approximationsaufgaben zurückgreifen und sich mit Näherungen begnügen. Es gibt eine Reihe von Verfahren zur Berechnung von besten l_1-Approximationen. Häufig überführt man dabei das Approximationsproblem in eine lineare Optimierungsaufgabe (siehe z. B. Barrodale, Young /6/).

2. Die Konstruktion von G, so daß die Ungleichung $\|I - AG\|_Z < 1$ erfüllt wird ($\|\cdot\|_Z$ bezeichne die Zeilensummennorm), führt auf Approximationsaufgaben etwas anderer Art. Es sei $\tilde{a}^k$ der k-te Spaltenvektor von A^T, dann wird

$$\|I - A\,G\|_Z = \max_k \| e^k - G^T \tilde{a}^k \|_1 .$$

Diese Aufgabe kann in gewissem Sinne auch als eine inverse Approximationsaufgabe interpretiert werden: Für eine gegebene Ersatzinverse G sind Matrizen A zu konstruieren, so daß G eine hinreichend gute Rechts-Ersatzinverse bez. $\|I - A\,G\|_Z$ ist.
Hier sind die Koeffizienten der besten Approximation mit A gegeben, und gesucht sind die approximierenden Spaltenvektoren der Matrix G^T.
Analog kann man zur Erreichung von $\|I - G\,A\|_S < 1$ Approximationsaufgaben der Art

$$\min_G \max_k \| e^k - G\,a^k \|_1$$

betrachten.

Die Konstruktion von G mit $\|I - G A\|_Z < 1$ führt auf Approximationsaufgaben der Form

$$\min_G \max_k \|e^k - A^T \tilde{g}^k\|_1 ,$$

wobei $\tilde{g}^k$ der k-te Zeilenvektor von G ist. In diesem Falle wird e^k durch Linearkombinationen der Zeilenvektoren von A approximiert.

3. Der Einfachheit wegen haben wir uns bei der Konstruktion von G auf die Forderungen (8), (9) beschränkt.
Man kann mit der in dieser Weise konstruierten Parametermatrix G zum Verfahren (3) übergehen. Mitunter erhält man dann bessere Konvergenz. Die Konvergenz kann sich aber auch verschlechtern.
Direkt auf das Einzelschrittverfahren (3) zugeschnittene Konstruktionsmöglichkeiten ergeben sich analog zu der hier geschilderten Vorgehensweise aus den Forderungen (10), (11).
Die Verwendung der Zeilensummennorm führt hier aber auf komplizierte Approximationsaufgaben, so daß es in der Regel günstiger erscheint, hinreichend gute Gesamtschrittverfahren zu konstruieren.

4. Durch entsprechende Wahl der Parametermatrix G erhält man die Block-Jacobi- und Block-Gauß-Seidel-Verfahren. Bei diesen Verfahren geht man aus von einem System der Form

$$X_1 = - A_{11}^{-1} A_{12} X_2 - \ldots - A_{11}^{-1} A_{1q} X_q + A_{11}^{-1} B_1 ,$$
$$\ldots\ldots$$
$$X_q = - A_{qq}^{-1} A_{q1} X_1 - \ldots - A_{qq}^{-1} A_{q,q-1} X_{q-1} + A_{qq}^{-1} B_q ,$$

wobei die A_{ii} quadratische Matrizen der Ordnung n_i sind,

$\sum_{i=1}^{q} n_i = n$, X_i aus R^{n_i}, $x^T = (X_1^T, \ldots, X_q^T)$, $A = (A_{ij})$,
$b^T = (B_1^T, \ldots, B_q^T)$. Setzt man

$$G := \operatorname{diag}(A_{11}^{-1}, \ldots, A_{qq}^{-1}), \tag{25}$$

so wird aus (2) das Block-Jacobi-Verfahren und analog aus (3) das Block-Gauß-Seidel-Verfahren.
Die exakte Lösung der l_1-Approximationsaufgaben (m_k gemäß (20)) führt bei entsprechender Strategienwahl zur Approximation der e^k auf (25), falls eine zu (22) analoge Ungleichung erfüllt ist. Insofern motivieren die hier dargestellten l_1-Strategien auch eine Verallgemeinerung der Block-Jacobi- und Block-Gauß-Seidel-Verfahren.

5. Eine universelle Wahl von G:
Es sei A nichtsingulär, $0 < a < 2/\text{Spur } A\,A^T$ und $G := a\,A^T$.
Dann ist $\varrho(I - A\,G) < 1$.
Auf den nicht schwierigen Beweis sei hier verzichtet (man vgl. z. B. auch mit Isaacson, Keller /7/, S. 88).

6. Es sollte ein Ziel weiterer Untersuchungen sein, geeignete Ersatzinversen G "ohne viel Aufwand aus A abzulesen", wenigstens für gewisse Klassen von Matrizen A.

Literatur

/1/ Albrand, H.-J.: Beiträge zur Theorie zyklischer Iterationsverfahren. Dissertationsschrift (B), Wilhelm-Pieck-Universität Rostock 1978

/2/ Wittmeyer, H.: Über die Lösung von linearen Gleichungssystemen durch Iteration. Z. Angew. Math. Mech. 16, 301 - 310 (1936)

/3/ Maeß, G.: Iterative Lösung linearer Gleichungssysteme. Dissertation (B), Wilhelm-Pieck-Universität Rostock 1976

/4/ Young, D. M.: Iterative solution of large linear systems. New York, London 1971

/5/ Richardson, L. F.: The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stresses in a masonry dam. Philos. Trans. Roy. Soc. London Ser. A 210, 307 - 357 (1910)

/6/ Barrodale, I., and Young, A.: Algorithms for best L_1 and L_∞ linear approximations on a discret set. Numer. Math. 8, 295 - 306 (1966)

/7/ Isaacson, E., and Keller, H. B.: Analyse Numerischer Verfahren. Leipzig 1972

eingegangen: 03. 05. 1979

Anschrift des Verfassers

Dr. H.-J. Albrand
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 77 - 85 (1979)

Gerhard Maess

A projection method solving general linear algebraic equations[1)]

The paper is concerned with a total step variant of the single step projection methods investigated by Tanabe /5/, Peters /3/, and others (cf. also Householder and Bauer /1/). The method is potentially useful for solving problems with sparse matrices which are too large to be stored in the internal memory. Furthermore it is suited to be used on multiprocessor computers.

1. Introduction

Let us consider a system of linear algebraic equations

$$A x = b, \quad A : \mathbb{R}^N \to R(A) \subset \mathbb{R}^M \tag{1}$$

with a rectangular matrix A without nullrows, mapping the euclidean space $\mathbb{R}^N$ onto its range R(A), a subspace of $\mathbb{R}^M$. If the equations are consistent, we are interested in a solution of minimal norm. In the inconsistent case we are looking for a so-called pseudo- or generalized solution (least-squares-solution) mimizing the residual vector

$$r = b - A x . \tag{2}$$

If we split $\mathbb{R}^N$ and $\mathbb{R}^M$ into the two orthogonal subspaces $R(A^T)$, N(A) and R(A), $N(A^T)$ respectively, the vector $x \in \mathbb{R}^N$ and the right hand side $b \in \mathbb{R}^M$ have the orthogonal components x_* , x_0 and b_* , b_0, respectively:

$$x = x_* + x_0, \ (x_*, x_0) = 0, \ x_* \in R(A^T), \ x_0 \in N(A), \tag{3}$$

$$b = b_* + b_0, \ (b_*, b_0) = 0, \ b_* \in R(A), \ b_0 \in N(A^T). \tag{4}$$

1) Part of a lecture, held at the conference "Algorithms '79" in Strbske Pleso, ČSSR, April 23 - 27, 1979

Here (.,.) denotes the euclidean scalar product and N(.) the null space. It is a well-known result from the theory of generalized inverses that the solution of

$$A\, x_* = b_* \tag{5}$$

is uniquely determined by

$$x_* = A^+ b_* = A^+ b, \tag{6}$$

A^+ denoting the Moore-Penrose pseudoinverse of A (cf. /2/).

2. Approximation by rows

In this section we consider the consistent case $b_0 = 0$. We split the set $\mathbb{M}$ of natural numbers $1,2,\ldots,M$ into N_0 $(1 \le N_0 \le M)$ not necessarily disjoint subsets L_i such that the transposed row vectors z^j of the matrix A, belonging to the same subset L_i, are linearly independent. By x^n, r^n we denote the n-th iterate of the solution and the corresponding residual vector, respectively. In the n-th step we compute N_0 projection vectors

$$y^i = \sum_{k \in L_i} u_k{}^i\, z^k, \quad i = 1,2,\ldots,N_0 \ , \tag{7}$$

such that

$$(x + y^i - x_*,\ z^j) = 0 \ . \tag{8}$$

That means that the coefficients $u_k{}^i$ are determined by solving (parallel if N_0 processors are available) the small systems

$$\sum_{k \in L_i} (z^j, z^k)\, u_k{}^i = r_j{}^n, \quad j \in L_i \ . \tag{9}$$

Combining the $u_k{}^i$ and the y^i in the form

$$u_k = \sum_i u_k{}^i, \quad y = \sum_{i=1}^{N_0} y^i = \sum_{k=1}^{M} u_k\, z^k = A^T u \tag{10}$$

we obtain the new approximation

$$x^{n+1} = x^n + t_n y \,, \tag{11}$$

$$r^{n+1} = r^n - t_n A y \,, \tag{12}$$

where t_n is computed from the orthogonality condition

$$(x^{n+1} - x_*, y) = 0 \,, \tag{13}$$

i. e.

$$t_n = (u, r^n)/(y, y) \,. \tag{14}$$

Because of the orthogonality condition (13) the procedure is error-reducing:

$$\|x^{n+1} - x_*\|^2 = \|x^n - x_*\|^2 - t_n^2 \|y\|^2 \,. \tag{15}$$

Theorem:

1) The algorithm (9) - (14) converges for arbitrary initial vectors $x^0 \in \mathbb{R}^N$ to

$$x^\infty = (I_N - B A)\, x^0 + B b \,, \tag{16}$$

$$r^\infty = (I_M - A B)\, r^0 \,. \tag{17}$$

2) The matrix

$$I_N - B A = P_{N(A)} \tag{18}$$

is the orthogonal projector onto the nullspace of A.

3) The matrix B is a generalized inverse of the type $A^{1,2,4}$ (a so-called reflexive minimum-norm inverse):

$$A B A = A, \quad B A B = B, \quad (B A)^T = B A \,. \tag{19}$$

4) If the system $A x = b$ is consistent, then x^∞ is a solution for all initial vectors x^0. If $x^0 \in R(A^T)$, then x^∞ is equal to x_*, the solution with minimal euclidean norm.

5) If A has maximal row-rank

$$\text{rank } A = M \,, \tag{20}$$

then B is the Moore-Penrose pseudoinverse A^+ .

Proof: (The argument is similar to the convergence proofs of the PSH and the SPA algorithms (see /3/, /4/, /5/)).
To describe the algorithm in a vectorial form, we use matrices E_i, which consist of those columns of the unity matrix I_M, the numbers of which belong to L_i. Then the matrices $(z^j, z^k)_{j,k \in L_i}$ of the small systems (9) can be written in the form $E_i^T A A^T E_i$. If we use the abbreviations

$$C_i = E_i (E_i^T A A^T E_i)^{-1} E_i^T , \quad C = \sum_{i=1}^{N_o} C_i , \tag{21}$$

we obtain

$$u = C r^n , \quad y = A^T C r^n , \tag{22}$$

and the algorithm reads

$$x^{n+1} = T_n x^n + v_n , \tag{23}$$

$$r^{n+1} = S_n r^n , \tag{24}$$

with

$$T_n = I_N - P_n A , \quad S_n = I_M - A P_n , \quad v_n = P_n b , \tag{25}$$

$$P_n = t_n A^T C , \quad t_n = (r^n, C r^n) / \|A^T C r^n\|^2 . \tag{26}$$

According to the orthogonal splitting $\mathbb{R}^N = R(A^T) \oplus N(A)$ we have $T_n = T_n P_{R(A^T)} + T_n P_{N(A)}$, where $P_{R(A^T)}$, $P_{N(A)}$ denote the orthogonal projectors onto the corresponding subspaces. Since $A P_{N(A)} = 0$, it follows that

$$T_n = \tilde{T}_n + P_{N(A)} , \quad \tilde{T}_n = T_n P_{R(A^T)} . \tag{27}$$

We first prove that the restriction $\tilde{T}_n$ of T_n onto the subspace $R(A^T)$ is contractive.

$$\|\tilde{T}_n\| \leq q < 1 . \tag{28}$$

As a majorant operator we use

$$T := I_N - \frac{1}{N_o} \sum_{i=1}^{N_o} A^T C_i A \,. \tag{29}$$

For all $z \in \mathbb{R}^N$ we have

$$\|T z\| \leq \frac{1}{N_o} \sum_{i=1}^{N_o} \|(I_N - A^T C_i A) z\| \leq \|z\| \,, \tag{30}$$

since the mappings $I_N - A^T C_i A$ are orthogonal projectors and consequently they have the norm 1. From (30) it follows that $\|T\| \leq 1$. We now define $\tilde{T} := T P_{R(A^T)}$ according to (27) and show that

$$q := \|\tilde{T}\| < 1. \tag{31}$$

If $\|\tilde{T}\|$ would be equal to 1, there would exist a nonvanishing vector $z_o \in R(A^T)$ with $\|z_o\| = \|\tilde{T} z_o\| = \frac{1}{N_o} \sum \|(I_N - A^T C_i A) z_o\|$, i. e. a vector with $\|(I_N - A^T C_i A) z_o\| = \|z_o\|$. That means, z_o would be orthogonal to all subspaces, spanned by the columns z^j, $j \in L_i$ of the matrix A^T, i. e. orthogonal to $R(A^T)$, what implies $z_o = 0$ in contradiction to the assumption.
Because of the orthogonality condition (13) the following inequalities hold for all $x \in R(A^T)$:

$$\|\tilde{T}_n x + t_n A^T C b - x_*\| \leq \|\tilde{T} x + \frac{1}{N_o} A^T C b - x_*\| \,,$$

$$\|\tilde{T}_n (x - x_*)\| \leq \|\tilde{T}(x - x_*)\| \,,$$

and consequently $\|\tilde{T}_n\| \leq \|\tilde{T}\|$, which completes the proof of (28).

1) To show the convergence of the sequence $\{x^n\}$, we use the relation

$$P_n S_k = T_k P_n \text{ for all natural } k, n \tag{32}$$

and obtain from (23) - (26) by induction

$$x^{n+1} - x^n = T_n T_{n-1} \cdots T_1 P_{n+1} r^o \,. \tag{33}$$

By consecutive multiplication it follows from (27)

$$T_n\, T_{n-1} \cdots T_1 = \tilde{T}_n \tilde{T}_{n-1} \cdots \tilde{T}_1 + P_{N(A)} \,, \tag{34}$$

so we have

$$x^{n+1} - x^n = \tilde{T}_n\, \tilde{T}_{n-1} \cdots \tilde{T}_1\, P_{n+1}\, r^o \,, \tag{35}$$

since $P_{N(A)}\, P_{n+1}$ vanishes. The matrices P_n are bounded because t_n is bounded:

$$|t_n| = \frac{\|I_N - T_n\|}{\|A^T C A\|} \leq \frac{2}{\|A^T C A\|} \,. \tag{36}$$

With

$$\|P_n\| = |t_n|\, \|A^T C\| \leq \frac{2\|A^T C\|}{\|A^T C A\|} =: K \tag{37}$$

we can estimate

$$\|x^{n+1} - x^n\| = K \|r_o\| q^n$$

and

$$\|x^{n+m} - x^n\| = K\|r_o\| q^n \frac{1 - q^m}{1-q} < K\|r_o\| \frac{q^n}{1-q} \,,$$

which proves the convergence of $\{x^n\}$.
To obtain the explicit expression (16) for the limit vector x^∞ we derive from (23) an explicit formula for x^n by consecutive insertion:

$$x^n = T_n\, T_{n-1} \cdots T_1 x^o + B_n b, \; B_n = \sum_{l=0}^{n-2} T_n \cdots T_{n-l} P_{n-l-1} + P_n \,. \tag{38}$$

According to (28) and (34) we obtain

$$\lim_{n\to\infty} T_n\, T_{n-1} \cdots T_1 = P_{N(A)} \,, \tag{39}$$

and the convergence of the sequence $\{B_n\}$ can be proved by similar arguments as in the case of $\{x^n\}$. We call the limit matrix B:

$$B := \lim_{n\to\infty} B_n \,. \tag{40}$$

2) To prove (18) (and thereby (16), (17)) we use the relations

$$I_N - B_n A = T_n T_{n-1} \cdots T_1, \quad I_M - A B_n = S_n S_{n-1} \cdots S_1, \quad (41)$$

which are easily derived from (25) and (38) by induction.

3) The relations (19) now follow from (18) by multiplication with A from the left and B from the right hand side.

4) By multiplication with A and using $b = A x_*$ (16) implies $A x^\infty = (A - A B A) x^0 + A B A x_* = b$, so x^∞ solves (1). If we choose the initial vector x^0 from $R(A^T)$, it is projected by $P_{N(A)}$ onto zero, hence we have

$$x^\infty = B b = B A x_* = x_*$$

because B·A is equal to the projector $P_{R(A^T)}$.

5) Multiplying (18) with A^T we obtain the equation $A^T = BAA^T$, since $P_{N(A)}A^T$ vanishes. If A has maximal row-rank, then AA^T is nonsingular and for B we have

$$B = A^T (A A^T)^{-1} , \quad (42)$$

so A·B is equal to the unity matrix I_M and the third Moore-Penrose condition $(A B)^T = A B$ is fulfilled.

3. Approximation by columns

In the rankdeficient inconsistent case it is possible to use one of the SPA-algorithms, described in /4/, /6/, /7/. The result is a least-squares-solution

$$x^\infty = (I_N - D A) x^0 + D b \quad (43)$$

with the residual vector

$$r^\infty = (I_M - A D) r^0 , \quad (44)$$

where D is a $A^{1,2,3}$-inverse (a so-called reflexive least-squares-inverse) of the matrix A and $I_M - A D$ is the orthogonal projector onto the nullspace of A^T.
If we are interested in the least-squares-solution with minimal norm $x_* = A^+ b$, we must at first compute D b and then start the above described row-algorithm with the initial vector $x^0 = D b$ and the right hand side b = 0 (cf. /3/). The result is

$$x^\infty = (I_N - B A) D b = x^0 - A^+ b ,$$

since $B A = A^+ A$ and $A D = A A^+$ according to the properties of the $A^{1,2,4}$- and the $A^{1,2,3}$-inverses.

Literature

/1/ Householder, A. S., and Bauer, F. L.: On certain iterative methods for solving linear systems. Numer. Math. 2, 55 - 59 (1960)

/2/ Nashed, M. Z., and Votruba, G. F.: A unified operator theory of generalized inverses. In: Nashed, M. Z. (Ed.): Generalized inverses and applications. New York 1976, pp. 1 - 109

/3/ Peters, W.: Projektionsverfahren und verallgemeinerte Inverse. Dissertation, Wilhelm-Pieck-Universität Rostock 1976

/4/ Maess, G.: Iterative Lösung linearer Gleichungssysteme. Nova Acta Leopoldina, Neue Folge, erscheint 1979

/5/ Tanabe, K.: Projection method for solving a singular system of linear equations and its applications. Numer. Math. 17, 203 - 214 (1971)

/6/ Kiesewetter, H., und Maess, G.: Elementare Methoden der numerischen Mathematik. Berlin - Wien 1974

/7/ Maess, G.: Ein Gesamtschrittverfahren zur Lösung linearer Gleichungssysteme. Acta Polytechnica, Práce ČVZT v Praze 4, 65 - 70 (1973)

received: Mai 7, 1979

Authors address:

Doz. Dr. Gerhard Maess
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 87 - 99 (1979)

Wolfgang Moldenhauer

Ein spezielles Dreieckselement beim Finite-Elemente-Verfahren

1. Bisherige Ergebnisse

Die Untersuchung der Interpolationspolynome n-ten Grades in einem Dreieck wird erst seit 1969 systematisch betrieben. Bis 1968 waren neben dem trivialen linearen Fall nur der Fall n = 2 von Fraeijs de Veubeke /8/ und der Fall n = 3 von Holand und Bergan /11/ untersucht. Bell /3/, Bosshard /5/, Visser /22/, Argyris, Fried, Scharpf /1/ und Zlamal /24/ konstruierten dann Polynome 5-ten Grades. In /1/ wurden darüber hinaus Polynome 6-ten und 7-ten Grades betrachtet. In /24/ und /23/ werden dann Polynome bis einschließlich 14-ten Grades beschrieben. Für ein beliebiges Polynom vom (4n+1)-ten Grad wird in /6/ von Bramble und Zlamal die Theorie aufbereitet und ein Konvergenzkriterium hergeleitet. Koukal /13/ untersucht Polynome von ungeradem Grad und erzielt Konvergenzkriterien. Unabhängig von dieser Arbeit werden hier eigene ähnliche Ergebnisse vorgestellt.

2. Auswahl der Stützwerte

Es ist zweckmäßig, die Bedingungen, die ein Polynom bestimmen, in den Eckpunkten des Dreiecks zu konzentrieren, weil diese Bedingungen gleichzeitig für mehrere Dreiecke herangezogen werden können.

In den eingangs erwähnten Arbeiten werden Bedingungen auf den Seiten des Dreiecks zum Beweis von Abschätzungen benutzt. Diese Bedingungen müssen bei der numerischen Rechnung eliminiert werden, wie das etwa bei /12/ und Zlamal /25/ dargestellt ist, was jedoch einen erheblichen Aufwand nach sich zieht und Berechnungen nicht optimal gestaltet. Außerdem ist unklar, ob sich dies im allgemeinen Fall realisieren läßt. Die hier neu vorgestellte Theorie kommt weitgehend ohne Verwendung solcher Stützwerte aus. Nur in dem Falle, daß der Polynomgrad gerade

ist, wird auf jeder Seite ein Stützwert benötigt.
Seien mit P_i, $i = 1, 2, 3$, die Eckpunkte und mit P_0 der Schwerpunkt eines Dreiecks bezeichnet. Ein Polynom $p(x,y)$ n-ten Grades hat $\binom{n+2}{2}$ Koeffizienten, ist also durch $\binom{n+2}{2}$ unabhängige Bedingungen eindeutig bestimmt. Wir schreiben nun folgende Werte vor:

$D^j p(P_i)$ für alle $j \leq m$, $i = 1, 2, 3$, und $D^k p(P_0)$ für alle $k \leq l$.

Dabei ist $D^j p = \frac{\partial^j p}{\partial x^{j_1} \partial y^{j_2}}$ mit $j_1 + j_2 = j$.
Wir geben also $3\binom{m+2}{2}$ Bedingungen in den Eckpunkten und $\binom{l+2}{2}$ Bedingungen im Schwerpunkt vor. Soll ein Polynom n-ten Grades durch diese Vorgaben bestimmt sein, muß gelten:

$$3\binom{m+2}{2} + \binom{l+2}{2} = \binom{n+2}{2}. \qquad (1)$$

Wir suchen nun Lösungen dieser diophantischen Gleichung in natürlichen Zahlen m und l bei vorgegebenem Grad n derart, daß m möglichst groß und l entsprechend klein ist. Diese letzte Forderung ergibt sich aus den oben durchgeführten Überlegungen. Die vollständige Lösung der Gleichung (1) kann nicht gegeben werden, da es sich bei der Darstellung einer natürlichen Zahl durch eine ternäre quadratische Form um ein ungelöstes Problem handelt /2/, /10/, /18/, /19/. Uns sind folgende partikuläre Lösungen der Gleichung (1) bekannt:

a)	$n = 4r+1$	$m = 2r$	$l = 2r-1$
b)	$n = 4r+2$	$m = 2r$	$l = 2r+1$
c)	$n = 4r+3$	$m = 2r+1$	$l = 2r$
d)	$n = 4r+4$	$m = 2r+1$	$l = 2r+2$,

wobei $r \geq 0$ eine natürliche Zahl ist.
Nun genügen einige dieser partikulären Lösungen nicht der Forderung nach maximalem m. Übereinstimmung zeigt sich jedoch für alle $n \leq 7$. Praktische Erfahrungen, z. B. /7/, haben jedoch gezeigt, daß Polynome bei $n \leq 3$ einen vertretbaren Rechenaufwand mit sich bringen, wobei deren sämtliche Koeffizienten ungleich Null sein können. Will man mit Polynomen vom Grade $n \geq 3$ arbei-

ten, so ist es zweckmäßig, bei Beginn der Rechnung einige Koeffizienten gleich Null zu setzen. In diesem Fall können Lösungen der Gleichung (1), die nicht von unserer Fallunterscheidung erfaßt wurden und der Forderung nach maximalem m genügen, nutzbringend sein.

3. Zwei Hilfssätze

Hilfssatz 1: Sei $g(s) \in C^{(n+1)}[0,1]$, und es gelte
$g^{(\mu)}(0) = y_0^{(\mu)}$, $g^{(\mu)}(1) = y_1^{(\mu)}$,
$\mu = 0, 1, \ldots, \alpha_0 - 1$ ($\alpha_0 = 1$). Ist $|g^{(n+1)}(s)| \leq M_{n+1}$ in $(0,1)$, dann ist in $[0,1]$

$$|g^{(i)}(s)| \leq K_1 \max_{\nu,\mu} |y_\nu^{(\mu)}| + K_2 M_{n+1} 1^{n+1-i}, \qquad \nu = 0,1,$$

für alle $i \leq n-1$ ($n = 2\alpha_0 - 1$).

Hilfssatz 2: Sei $g(s) \in C^{(n+1)}[0,1]$, und es gelte
$g^{(\mu)}(0) = y_0^{(\mu)}$, $g(\frac{1}{2}1) = y_1$, $g^{(\mu)}(1) = y_2^{(\mu)}$,
$\mu = 0, 1, \ldots, \alpha_0 - 1$ ($\alpha_0 \geq 1$). Ist $|g^{(n+1)}(s)| \leq M_{n+1}$ in $(0,1)$, dann ist in $[0,1]$

$$|g^{(i)}(s)| \leq K_1 \max_{\nu,\mu} |y_\nu^{(\mu)}| + K_2 M_{n+1} 1^{n+1-i}, \qquad \nu = 0,1.$$

Der Beweis der beiden Hilfssätze kann mit der von Beresin, Shidkow /4/ beschriebenen Methode erbracht werden. Die Konstanten sind bei Mollenhauer /16/ angegeben.

4. Zwei Approximationssätze

Für die Anwendung der neuen Dreieckselemente ist es erforderlich zu untersuchen, inwieweit man eine Funktion auf einem Element durch das zugehörige Approximationspolynom n-ten Grades annähern kann. Die Beantwortung dieser Fragestellung ist mit

den beiden folgenden Sätzen möglich. Der Approximationscharakter der Sätze wird deutlich, wenn man die Funktion $w(x,y)$ als Fehlerfunktion interpretiert.

Satz 1: Sei $n = 4r + d$, $r \geq 0$, $d \in \{1,3\}$, $r \in \mathbb{N}$.
Sei $w(x,y) \in C^{(n+1)}(\overline{T})$ und $|D^{n+1}w(x,y)| \leq M_{n+1}$.

Ist $D^j w(P_i) = D^k w(P_0) = 0$, $i = 1,2,3$,

$$\forall\, j \leq \begin{cases} 2r \\ 2r+1 \end{cases} \text{ und } \forall\, k \leq \begin{cases} 2r-1 \\ 2r \end{cases}, \text{ falls } d = \begin{cases} 1 \\ 3 \end{cases},$$

und gilt auf jeder Dreiecksseite

$$\max_{s_1} \left| \frac{\partial^{n-1} w}{\partial s_1^{n-1}} \right| \neq 0, \text{ dann gilt in } \overline{T}:$$

$$|D^i w(x,y)| \leq K M_{n+1} \frac{1}{\sin^i \varphi} h^{n+1-i}$$

für alle $i \leq n-1$, wobei φ der kleinste Winkel und h die Länge der größten Seite von T ist.

Bemerkung: Das Koordinatensystem (x,y) wird in ein (s_1,s_2)-Koordinatensystem transformiert, so daß die s_1-Achse mit der jeweiligen Dreiecksseite zusammenfällt.

Der Beweis befindet sich bei Moldenhauer /17/. Er greift auf die Hilfssätze 1 und 2 zurück.
Koukal /13/ hat gezeigt, daß man im Falle $i = 0,1$ die Voraussetzungen abschwächen kann. Insbesondere kann auf die Randvoraussetzung verzichtet werden.
Wir wollen nun die analogen Sätze für gerade n herleiten. Im Falle des Satzes 1 sind auf dem Rande des Dreiecks $n+1$ Bedingungen vorgegeben. Bei den geraden n haben wir auf dem Rande des Dreiecks lediglich n Bedingungen vorgegeben. Es bleibt zu vermuten, daß bei gleichbleibender Struktur der Stützwertvorgabe ein entsprechender Satz bez. der Konvergenzrate eine h-Potenz weniger liefert. Hinzu käme, daß nur die Beschränktheit der n-ten Ableitung im Inneren des Dreiecks benötigt würde.

Aus diesen Gründen werden wir, wie bereits früher angedeutet, unsere Strategie etwas abändern. Wir geben uns dazu in den Seitenmittelpunkten jeder Dreiecksseite genau eine Bedingung vor und verzichten auf drei Bedingungen im Schwerpunkt.

Satz 2: Sei $n = 4r + d$, $r \geq 0$, $d \in \{2,4\}$, $r \in \mathbb{N}$.
Sei $w(x,y) \in C^{(n+1)}(T)$ und $|D^{n+1}w(x,y)| \leq M_{n+1}$.
Ist $D^j w(P_i) = D^k w(P_0) = w(Q_i) = 0$, $i = 1,2,3$, wobei Q_i die Seitenmittelpunkte des Dreiecks sind,

$$\forall\, j \leq \begin{cases} 2r \\ 2r+1 \end{cases} \text{ und } \forall\, 2 \leq k \leq \begin{cases} 2r+1 \\ 2r+1 \end{cases}, \text{ falls } d = \begin{cases} 2 \\ 4 \end{cases},$$

und gilt auf jeder Dreiecksseite

$$\max_{s_1} \left| \frac{\partial^{n-1} w}{\partial s_1^{n-1}} \right| \neq 0, \text{ dann gilt in } \bar{T}:$$

$$|D^i w(x,y)| \leq M_{n+1} \frac{1}{\sin^i \varphi}\, h^{n+1-i}$$

für alle $i \leq n-1$, wobei φ der kleinste Winkel und h die Länge der größten Seite von T ist.

Der Beweis dieses Satzes ist dem des Satzes 1 analog.

5. Zwei Eindeutigkeitssätze

Die beiden nachfolgenden Sätze geben an, durch welche Vorgabe der Bedingungen ein Polynom n-ten Grades in zwei Veränderlichen in den einzelnen Fällen auf einem Dreieck T eindeutig bestimmt ist.

Satz 3: Seien P_1, P_2, P_3 die Eckpunkte und P_0 der Schwerpunkt eines Dreiecks T. Ein Polynom vom Grade $n = 4r + d$, $d \in \{1,3\}$, $p(x,y) = a_1 + a_2 x + a_3 y + \ldots + a_N y^n$, wobei $N = \binom{n+2}{2}$ ist, ist eindeutig auf T bestimmt durch die Vorgabe der Werte $D^j p(P_i)$ und $D^k p(P_0)$, $i = 1,2,3$, für alle j und k mit

$$j \leq \begin{cases} 2r \\ 2r+1 \end{cases}, \quad k \leq \begin{cases} 2r-1 \\ 2r \end{cases}, \text{ falls } d = \begin{cases} 1 \\ 3 \end{cases}.$$

Zum Beweis kann man die bereits in /12/ verwendete Argumentation benutzen. Es wird dabei der Satz 1 benutzt.
Man vergleiche hierbei mit Satz 5 von Koukal /13/, der aussagt: Zu jedem $f \in C^{(n-1)}(T)$ gibt es genau ein Polynom von höchstens (2n-1)-tem Grad, so daß

$$D^j p(P_i) = D^j f(P_i), \quad j \leq n-1,$$

und $D^k p(P_o) = D^k f(P_o), \quad k \leq n-2$, gilt.

Satz 4: Seien P_1, P_2, P_3 die Eckpunkte, Q_1, Q_2 und Q_3 die Seitenmitten und P_o der Schwerpunkt eines Dreiecks T. Ein Polynom vom Grade $n = 4r + d$, $d \in \{2,4\}$,
$p(x,y) = a_1 + a_2 x + a_3 y + \ldots + a_N y^n$, wobei $N = \binom{n+2}{2}$ ist, ist eindeutig auf T bestimmt durch die Vorgabe der Werte $D^j p(P_i)$, $p(Q_i)$ und $D^k p(P_o)$, $i = 1,2,3$, für alle j und k mit

$$j \leq \begin{cases} 2r \\ 2r+1 \end{cases} \text{ und } 2 \leq k \leq \begin{cases} 2r+1 \\ 2r+2 \end{cases}, \text{ falls } d = \begin{cases} 2 \\ 4 \end{cases}.$$

Der Beweis ist dem des Satzes 3 analog, jedoch wird vom Satz 2 Gebrauch gemacht.

6. Spezialfälle für $n \leq 3$

Für die praktisch interessanten Fälle $n \leq 3$ wollen wir der Vollständigkeit halber einige Verschärfungen aus der Literatur angeben. Für $n = 1$ hat Synge /20/ gezeigt:
Ist $w(x,y) \in C^2(T)$, $w(P_i) = 0$, $i = 1,2,3$, und $|D^2 w(x,y)| \leq M_2$, so ist in $\bar{T}$ $|w(x,y)| \leq K M_2 h^2$.
Von Kolar, Kratochvil, Zenisek, Zlamal /12/ wird ferner $|Dw(x,y)| \leq 2 \cos^{-1}\vartheta \, M_2 h$ ohne Beweis angegeben, wobei 2ϑ der größte Winkel des Dreiecks ist.

Von Zlamal /24/ stammen die folgenden beiden Sätze für n = 2 und n = 3.

Ist $w(x,y) \in C^3(T)$, $w(P_i) = w(Q_i) = 0$, $i = 1,2,3$, und $|D^3 w(x,y)| = M_3$, so ist in $\overline{T}$ $|w(x,y)| \leq M_3 h^3$ und $|Dw(x,y)| \leq 2 \sin^{-1}\varphi \, M_3 h^2$.

Ist $w(x,y) \in C^4(T)$, $D^j w(P_i) = w(P_o) = 0$, $i = 1,2,3$, $j = 0,1$, und $|D^4 w(x,y)| \leq M_4$, so ist in $\overline{T}$ $|w(x,y)| \leq 3 \sin^{-1}\varphi \, M_4 h^4$ und $|Dw(x,y)| \leq 5 \sin^{-1}\varphi \, M_4 h^3$.

Diese in Spezialfällen hergeleiteten Abschätzungen stimmen bez. der Konvergenzrate mit unseren Ergebnissen überein. Im Falle n = 3 haben wir sogar für die Funktionswerte eine schärfere Abschätzung erhalten.

7. Zur Gesetzmäßigkeit der Fehlerabschätzung

In den angegebenen Fehlerabschätzungen treten die geometrischen Größen des Dreiecks auf. Dies ist kein Zufall und auch keine Folge der Beweistechnik. Wir beweisen dies durch Angabe von Gegenbeispielen in den praktisch interessanten Fällen n = 1,2,3. Sei h fest vorgegeben. Wir betrachten die Menge der Dreiecke T_g mit $0 < g \leq \frac{1}{2}\sqrt{3}\, h$ und den Eckpunkten $P_1(-\frac{1}{2}h, -\frac{1}{2}g)$, $P_2(\frac{1}{2}h, -\frac{1}{2}g)$ und $P_3(0, \frac{1}{2}g)$. $p(x,y,g)$ sei das entsprechende Hermite-Interpolationspolynom und $G(x,y) = w(x,y) - p(x,y,g)$ der entsprechende Fehler.

Sei n = 1. Wir betrachten die Funktion $w(x,y) = x^2 + y^2$ auf T_g. Es gilt $\max_{T_g} |G(x,y)| \geq \frac{1}{4}h^2$ und $\max_{T_g} |DG(x,y)| \geq h$.

Damit ist auch das Ergebnis von Synge /20/ bez. der Konvergenzrate o(h) gesetzmäßig.

Sei n = 2. In /20/ wird die Funktion $w(x,y) = 4h^3y^2 - 256 \cdot (3h)^{-1}x^2(x^2 - 4^{-1}h^2) - h^3$ betrachtet. Es ergibt sich $\max_{T_g} |w_y - p_y| \geq 4 \sin^{-1}\varphi \, h^2$. Weiter ist

$\max_{T_g} |G(x,y)| \geq \frac{4}{3} h^3$.

Sei $n = 3$. Wir betrachten die Funktion $w(x,y) = 64\,x^4$ auf T_g. Wir erhalten

$$\max_{T_g} |G| \geq 2^7 3^{-5} h^4, \quad \max_{T_g} |DG| \geq 20 \sin^{-1}\varphi\, h^3, \quad \max_{T_g} |D^2 G| \geq 56 \sin^{-2}\varphi\, h^2.$$

Damit haben wir Beispiele der gewünschten Art angegeben. Interessant ist nun folgendes: Behalten wir im Fall $n = 2$ die alte Strategie bei, d. h., geben wir keinen Funktionswert in den Seitenmitten vor, so bekommen wir die gleiche Konvergenzrate, wie folgendes Beispiel zeigt.

Sei $w(x,y) = 2x^3$. Wir finden

$$\max_{T_g} |G(x,y)| \geq 18^{-1}\sqrt{3}\, h^3 \text{ und } \max_{T_g} |DG(x,y)| \geq 3 \cdot 4^{-1} \sin^{-1}\varphi\, h^2.$$

8. Verträglichkeit zweier Elemente

Wir wollen untersuchen, welche Eigenschaften die Funktion $P(x,y)$ im Gesamtgebiet $\bar{\Omega}$ hat, die in jedem Dreieck ein Polynom n-ten Grades darstellt. Es ist sofort klar, daß $P(x,y)$ stetig ist. Beide zu betrachtende Dreiecke mögen die Seite P_1P_2 gemeinsam haben. $P(x,y)$ ist in jedem der beiden Dreiecke ein Polynom n-ten Grades. Wir parametrisieren P_1P_2 durch $x = x_1 + (x_2-x_1)s$, $y = y_1 + (y_2-y_1)s$ mit $0 \leq s \leq 1$, wobei P_1 bzw. P_2 die Koordinaten (x_1,y_1) bzw. (x_2,y_2) haben. $p_1(x,y)$ geht damit auf P_1P_2 in ein Polynom n-ten Grades in der Veränderlichen s über und das entsprechende Polynom $p_2(x,y)$ des benachbarten Dreiecks gleichfalls. Auf P_1P_2 sind aber sowohl p_1 als auch p_2 genau $(n+1)$ Bedingungen in den Eckpunkten bzw. der Seitenmitte unterworfen, die sowohl p_1 als auch p_2 eindeutig bestimmen. Da alle diese Bedingungen in P_1 und P_2 bzw. den Seitenmitten übereinstimmen, ist dann auch $p_1 = p_2$ auf P_1P_2. Die eindeutige Bestimmtheit von p_1 und p_2 läßt sich mit dem Hilfssatz 1 bzw. 2 nachweisen. Sind etwa die vorgeschriebenen Bedingungen homogen, so ist $y_p^{(\mu)} = 0$. Weiter ist $M_{n+1} = 0$, und wenden wir auf p_1 und p_2 die Hilfssätze 1 bzw. 2 an, so folgt $p_1 = p_2 = 0$, und dies belegt die Eindeutigkeit.

Weitere Aussagen über die Verträglichkeit zweier Elemente können nicht gemacht werden. Wir weisen dies an folgendem Beispiel nach:
Die Funktion $w(x,y) = y^2 \cos \pi x$ sei auf den Dreiecken T_1 und T_2 erklärt. Die Eckpunkte seien
für T_1 $P_1(-2,0)$, $P_2(2,0)$ und $P_3(0,2)$,
für T_2 $P_1(-2,0)$, $P_2(2,0)$ und $P_3(0,-2)$.
Wir erhalten in T_1 $P_y(0,0) = -4$
und in T_2 $P_y(0,0) = 4$, womit der Nachweis geführt ist.

9. Eine Anwendung dieses neuen Dreieckselementes

Wir setzen voraus, daß der Rand S des Gebietes $\bar{\Omega}$ ein geschlossener Polygonzug ist und betrachten die Gleichung

$$Af = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial f}{\partial x_j}\right) + cf = g \quad \text{in } \bar{\Omega}.$$

Hierbei seien a_{ij}, c, g stetige Funktionen von (x_1,x_2) in $\bar{\Omega}$, und es gelte:
a) $a_{ij} = a_{ji}$ sei stetig differenzierbar in $\bar{\Omega}$,
b) $c \geq 0$,
c) $\sum_{i,j=1}^{2} a_{ij} b_i b_j \geq B \sum_{i=1}^{2} b_i^2$, $B = \text{const.} > 0$.

Es liegt also der nichtentartete elliptische Fall vor, und wir wollen das Dirichletproblem
$f|_S = 0$ lösen.
Dieses Problem ist positiv definit /15/, und die Lösung minimisiert das Funktional

$$F(w) = \int_{\Omega} \left(\sum_{i,j=1}^{2} a_{ij} \frac{\partial w}{\partial x_i} \frac{\partial w}{\partial x_j} + cw^2 - 2gw\right) dx$$

in $\overset{\circ}{W}{}_2^{(1)}(\Omega)$.

Wir triangulieren $\bar{\Omega}$, d. h., wir zerlegen $\bar{\Omega}$ in eine endliche Menge von Dreiecken T_i mit folgenden Eigenschaften:

a) $\bigcup_i T_i = \bar{\Omega}$, b) $T_i \cap T_j = \emptyset \quad (i \neq j)$.

Jeder dieser Triangulierungen ordnen wir zwei Parameter zu:

1. h als größte auftretende Dreiecksseite,
2. φ als kleinster auftretender Dreieckswinkel.

Sei $\tilde{f}(x_1,x_2)$ die approximierte Lösung des Randwertproblems. Für die Fehlerabschätzung benutzen wir die in /9/, /21/ und /24/ beschriebene Methode. Es ist, wie in /14/ gezeigt wurde, folgendes bekannt: Ist f die exakte Lösung und ist $f_1 \in \overset{o}{W}_2^{(1)}$, so gilt $E(f_1 - f) = F(f_1) - F(f)$, wobei $E(z) = E(z,z)$ mit

$$E(z_1,z_2) = \int_\Omega \Big(\sum_{i,j=1}^{2} a_{ij} \frac{\partial z_1}{\partial x_i} \frac{\partial z_2}{\partial x_j} + c z_1 z_2 \Big)\, dx \text{ ist.}$$

Also haben wir:

$$E(\tilde{f} - f) = F(\tilde{f}) - F(f) = \min_{v \in \mathfrak{P}} F(v) - F(f) = \min_{v \in \mathfrak{P}} E(v - f) \leq E(f - f_2), \qquad (2)$$

wobei $f_2 \in \mathfrak{P}$ diejenige Funktion ist, die in den Stützstellen dieselben Werte wie die Lösung f annimmt, und $\mathfrak{P}$ der Raum der Funktionen $P(x_1,x_2)$ ist.

Wir betrachten die Funktion $w = f - f_2$. Es ist $D^{n+1}w = D^{n+1}(f - f_2) = D^{n+1}f$, und es wird vorausgesetzt, daß f beschränkte Ableitungen (n+1)-ter Ordnung hat und auch die weiteren Bedingungen der Sätze 1 oder 2 erfüllt. Wir erhalten dann durch Anwendung der Sätze 1 oder 2 auf w unter Benutzung von (2) und der Friedrichsungleichung:

$$\| f - \tilde{f} \|_{W_2^{(1)}(\Omega)} \leq C M_{n+1} \frac{1}{\sin\varphi} h^n,$$

wobei die Konstante C nicht von der Triangulierung abhängt.

Literatur

/1/ Argyris, J. H., Fried, I., and Scharpf, D. W.:
The tuba family of plate elements for the matrix displacement method. The Aeronautical J. of the R. Aer. Soc. 72, 618 - 623 (1968)

/2/ Bachmann, P.: Die Arithmetik der quadratischen Formen. Leipzig 1898

/3/ Bell, K.: A refined triangular plate bending finite element. Internat. J. Numer. Meth. Engrg. 1, 101 - 122 (1969)

/4/ Beresin, I. S., und Shidkow, N. P.:
Numerische Methoden 1. Berlin 1970

/5/ Bosshard, W.: Ein neues vollverträgliches endliches Element für Plattenbiegung. Abhandlung der Internationalen Vereinigung für Brückenbau und Hochbau, Zürich 1968

/6/ Bramble, J. H., and Zlamal, M.:
Triangular elements in the finite element method. Math. Comp. 24, 809 - 821 (1970)

/7/ Dankert, J.: Probleme der Programmierung der Methode der finiten Elemente. Kurzvortrag auf der wissenschaftlichen Haupttagung der MGDDR, Halle 1974

/8/ Fraeijs de Veubeke, B.:
Displacement and equilibrium models in the finite element method. In: Zienkiewicz, O.C., and Holister, G. S. (Ed.): Stress analysis, chapter 9. London 1965

/9/ Friedrichs, K. O., and Keller, H. B.:
A finite difference scheme for generalized Neumann problems. In: Bramble, J. H. (Ed.): Numerical solution of partial differential equations. Proceedings of a Symposium held at the University of Maryland. New York 1966

/10/ Gelfond, A. O.: Ganzzahlige Lösungen von Gleichungen. München 1954

/11/ Holand, J., and Bergan, P. G.:
Higher-order finite element for plane stress. Proc. ASCE, EM2, 698 - 702 (1968)

/12/ Kolar, V., Kratochvil, J., Zlamal, A., and Zenisek, A.:
Technical, physical and mathematical principles of the finite element method. Rozpravy Československe Akad. Věd Řada Techn. Věd, Ročnik 81, Sesit 2, Praha 1971

/13/ Koukal, S.: Piecewise polynomial interpolations in the finite element method. Apl. Mat. 18, 146 - 160 (1973)

/14/ Michlin, S. G., and Smolickij, H. L.:
Approximate methods for solution of differential and integral equations. English translation, New York 1967

/15/ Michlin, S. G.: Lehrgang der mathematischen Physik. Berlin 1972

/16/ Moldenhauer, W.: Konvergenzuntersuchungen an finiten Elementen. Diplomarbeit, Wilhelm-Pieck-Universität, Rostock 1973

/17/ Moldenhauer, W.: Ein neues Dreieckselement beim Finiten-Element-Verfahren. Diss. A, Wilhelm-Pieck-Universität, Rostock 1976

/18/ Nagell, T.: Introduction to number theory. Stockholm-New York 1951

/19/ Skolem, T.: Diophantische Gleichungen. Berlin 1938

/20/ Synge, J. L.: The hypercycle in mathematical physics. Cambridge 1957

/21/ Varga, R. S.: Hermite interpolation-type Ritz methods for two-point boundary value problems. In: Bramble, J. H. (Ed.): Numerical solution of partial differential equations. Proceedings of a symposium held at the University of Maryland. New York 1966

/22/ Visser, M.: The finite element method in deformation and head conduction problems. Delft 1968

/23/ Zenisek, A.: Interpolation polynomials on the triangle. Numer. Math. 15, 283 - 296 (1970)

/24/ Zlamal, M.: On the finite element method. Numer. Math. 12, 394 - 409 (1968)

/25/ Zlamal, M.: On some finite element procedures for solving second order boundary value problems. Numer. Math. 14, 42 - 48 (1969)

eingegangen: 16. 10 1978

Anschrift des Verfassers:

Dr. Wolfgang Moldenhauer
DDR-25 Rostock 9
Landreiterstraße 8

Rostock. Math. Kolloq. 12, 101 - 111 (1979)

Wolfgang Moldenhauer und
Raimond Strauß

Zur Lösung der Helmholtz-Gleichung in einem polygonberandeten Gebiet mittels des Finite-Elemente-Verfahrens

1. Einleitung

Die Methode der finiten Elemente gehört zu den gebräuchlichsten Verfahren zur näherungsweisen Lösung von Randwertproblemen. Sie geht auf Courant /2/ zurück, der vorschlägt, ein gegebenes Gebiet in Dreiecke zu zerlegen und Versuchsfunktionen zu benutzen, die stückweise linear sind. In der vorliegenden Arbeit wird ein Verfahren beschrieben, bei dem stückweise quadratische Polynome verwendet werden. Es wird ein Algorithmus zum Aufbau der Elementsteifigkeitsmatrizen angegeben (vgl. Zlamal /9/). Bei den auftretenden Numerierungsverfahren folgen wir zunächst einer Idee von Mittelmann /4/. Die neu definierten Zusammenhangsmatrizen gestatten einen leichten Aufbau der Steifigkeitsmatrix aus den Elementmatrizen. Das Verfahren wird an den Beispielen "Torsionsproblem für das Quadrat und das gleichseitige Dreieck" getestet. Konvergenzaussagen, bei denen die Versuchsfunktionen Polynome beliebigen Grades sind, wurden von Zlamal /10/, Bramble, Zlamal /1/, Koukal /8/ und Moldenhauer /6/ angegeben.

2. Randwertprobleme zweiter Ordnung

Sei $\Omega \subset R^2$ ein beschränktes Gebiet mit dem Polygonrand Γ. Wir betrachten die Gleichung

$$-\Delta u(x,y) + c_0 u(x,y) = f(x,y), \tag{1}$$

wobei $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ ist.

Die Randbedingungen seien

$$u|_{\Gamma^*} = 0 \text{ und } \frac{\partial u}{\partial \nu}|_{\Gamma^{**}} = 0, \tag{2}$$

wobei $\Gamma = \Gamma^* \cup \Gamma^{**}$ und ν die Außennormale ist. Für die Konstante c_0 setzen wir $c_0 \geq 0$ voraus und fordern im Falle des Neumann-Problems ($\Gamma^{**} = \Gamma$) sogar $c_0 > 0$. Unter diesen Voraussetzungen ist das Randwertproblem (1), (2) V-elliptisch. Mit Hilfe der von Michlin /3/ bewiesenen Sätze über ein minimales Funktional erhält man, daß (1), (2) äquivalent dem Minimalproblem für das Energiefunktional

$$F(u) = \int_\Omega \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + c_0 u^2 - 2fu \right] d\Omega \tag{3}$$

in der Klasse $V \subset W_2^{(1)}(\Omega)$, deren Elemente der Dirichlet-Bedingung $u|_{\Gamma^*} = 0$ genügen, ist.

3. Ein Finite-Elemente-Verfahren

Das Gebiet Ω sei in endlich viele disjunkte Dreiecke zerlegt. Zwei benachbarte Dreiecke haben eine Seite gemeinsam, und die Vereinigung aller abgeschlossenen disjunkten Dreiecke sei gleich $\Omega \cup \Gamma$.

Unser Ziel ist es, das Minimum $u_0(x,y)$ von (3) durch quadratische Polynome $p(x,y) = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2$ über jedem Dreieck zu approximieren. $p(x,y)$ wird über dem Dreieck Q der Zerlegung eindeutig durch die Funktionswerte u_1, u_2, u_3 in den Eckpunkten und v_1, v_2, v_3 in den Seitenmittelpunkten bestimmt.

Das Funktional (3) ist eine quadratische Funktion in allen Parametern (u_i, v_i, $i = 1, 2, 3$). Wir erhalten die Werte der approximierenden Lösung $p_0(x,y)$ als die Näherungswerte der Lösung $u_0(x,y)$ in den Ecken und Seitenmitten, indem wir das Minimum von

$$F(p) = \sum_Q \int_Q \left[\left(\frac{\partial p}{\partial x}\right)^2 + \left(\frac{\partial p}{\partial y}\right)^2 + c_0 p^2 - 2fp \right] dxdy := \sum_Q \Phi(p,Q)$$

bestimmen.

Analog zur Arbeit von Zlamal /9/ stellen wir jedes Integral dieser Summe in der Form $w^T Kw - 2w^T \delta$ dar, wobei $w^T = (u_1, v_1, u_2, v_2, u_3, v_3)$ ist. $K + K^T$ nennen wir die Elementsteifigkeitsmatrix. Die Komponenten des Spaltenvektors δ werden durch $f(x,y)$ bestimmt. Die Bestimmung des Minimums von $F(p)$ führt auf das lineare Gleichungssystem

$$(K + K^T)\, w = 2\delta. \tag{4}$$

Im nachfolgenden geben wir ein Verfahren zur Konstruktion von K und δ an. Das Dreieck Q habe die Eckpunkte $P_i(x_i, y_i)$, $i = 1,2,3$. Durch

$$x = x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta, \quad y = y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta,$$
$$r(\xi,\eta) = \alpha_1 + \alpha_2\xi + \alpha_3\eta + \alpha_4\xi^2 + \alpha_5\xi\eta + \alpha_6\eta^2 \tag{5}$$

und

$$g(\xi,\eta) = |J|\, f\left[x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta,\ y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta\right]$$

erhalten wir

$$\Phi(p,Q) = \int_{Q1} \left[a\left(\frac{\partial r}{\partial \xi}\right)^2 + 2b\,\frac{\partial r}{\partial \xi}\frac{\partial r}{\partial \eta} + c\left(\frac{\partial r}{\partial \eta}\right)^2 + dr^2 - 2gr\right] d\xi d\eta. \tag{6}$$

Q1 ist dabei das Einheitsdreieck der (ξ,η)-Ebene, J die Funktionaldeterminante von (5), und es gilt

$$a = |J|^{-1}\left[(x_3 - x_1)^2 + (y_3 - y_1)^2\right], \quad c = |J|^{-1}\left[(x_2 - x_1)^2 + (y_2 - y_1)^2\right],$$
$$b = -|J|^{-1}\left[(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1)\right], \quad d = |J|\, c_0. \tag{7}$$

$r(\xi,\eta)$ ist auf Q1 durch folgende Bedingungen eindeutig bestimmt: $u_1 = r(0;0)$, $v_1 = r(0,5;0)$, $u_2 = r(1;0)$, $v_2 = r(0,5;0,5)$, $u_3 = r(0;1)$, $v_3 = r(0;0,5)$.
Mit $\alpha^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$ gilt dann

$$w = S\alpha \quad \text{und} \quad \alpha = S^{-1}w. \tag{8}$$

Mit $1^T = (0, 1, 0, 2\xi, \eta, 0)$ ist $\int_{Q1} a\left(\frac{\partial r}{\partial \xi}\right)^2 d\xi d\eta = a\alpha^T L \alpha = a w^T A w$,

wobei $L = \left\{ \int_{Q1} l_i l_j d\xi d\eta \right\}$ und $A = (S^{-1})^T L S^{-1}$ ist.

Mit $m^T = (0, 0, 1, 0, \xi, 2\eta)$ folgt

$$\int_{Q1} 2 \frac{\partial r}{\partial \xi} \frac{\partial r}{\partial \eta} d\xi d\eta = \alpha^T (M + M^T) \alpha = w^T (B + B^T) w,$$

wobei $M = \left\{ \int_{Q1} l_i m_j d\xi d\eta \right\}$ und $B = (S^{-1})^T M S^{-1}$ ist.

Setzen wir dies analog fort, so erhalten wir die Summe der ersten vier Terme von $\Phi(p,Q)$ als $w^T K w$ mit

$$K = aA + b(B + B^T) + cC + dD. \qquad (8a)$$

Es gilt:

$$S^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -3 & 4 & -1 & 0 & 0 & 0 \\ -3 & 0 & 0 & 0 & -1 & 4 \\ 2 & -4 & 2 & 0 & 0 & 0 \\ 4 & -4 & 0 & 4 & 0 & -4 \\ 2 & 0 & 0 & 0 & 2 & -4 \end{bmatrix} \qquad A = \frac{1}{6} \begin{bmatrix} 3 & -4 & 1 & 0 & 0 & 0 \\ -4 & 8 & -4 & 0 & 0 & 0 \\ 1 & -4 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & -8 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -8 & 0 & 8 \end{bmatrix}$$

$$B = \frac{1}{6} \begin{bmatrix} 3 & 0 & 0 & 0 & 1 & -4 \\ -4 & 4 & 0 & -4 & 0 & 4 \\ 1 & -4 & 0 & 4 & -1 & 0 \\ 0 & -4 & 0 & 4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & -4 & -4 & 4 \end{bmatrix} \qquad C = \frac{1}{6} \begin{bmatrix} 3 & 0 & 0 & 0 & 1 & -4 \\ 0 & 8 & 0 & -8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -8 & 0 & 8 & 0 & 0 \\ 1 & 0 & 0 & 0 & 3 & -4 \\ -4 & 0 & 0 & 0 & -4 & 8 \end{bmatrix}$$

$$D = \frac{1}{360} \begin{bmatrix} 6 & 0 & -1 & -4 & -1 & 0 \\ 0 & 32 & 0 & 16 & -4 & 16 \\ -1 & 0 & 6 & 0 & -1 & -4 \\ -4 & 16 & 0 & 32 & 0 & 16 \\ -1 & -4 & -1 & 0 & 6 & 0 \\ 0 & 16 & -4 & 16 & 0 & 32 \end{bmatrix}$$

Da A, C und D symmetrische Matrizen sind, ist K symmetrisch vom Format (6 x 6).

Sei $n = (1, \xi, \eta, \xi^2, \xi\eta, \eta^2)^T$, so ist

$$-2 \int_{Q1} \operatorname{grad} \xi d\eta = -2\alpha^T \int_{Q1} n d\xi d\eta = -2\alpha^T \beta = -2w^T (S^{-1})^T \beta = -2w^T \delta \quad \text{mit}$$

$$\delta = (S^{-1})^T \beta, \qquad (8b)$$

wobei

$$B = \left(\int_{Q1} g\,d\xi\,d\eta,\ \int_{Q1} \xi g\,d\xi\,d\eta,\ \int_{Q1} \eta g\,d\xi\,d\eta,\right.$$
$$\left. \int_{Q1} \xi^2 g\,d\xi\,d\eta,\ \int_{Q1} \xi\eta g\,d\xi\,d\eta,\ \int_{Q1} \eta^2 g\,d\xi\,d\eta \right)^T \qquad (9)$$

ist. Die Komponenten von ß müssen im allgemeinen mit Hilfe einer Quadraturformel (Moan /5/) berechnet werden.

Somit ist $\phi(p,q) = w^T K w - 2w^T \delta$.

Die Matrizen A, B, C und D, die keine aufgabenspezifische Daten enthalten, sind nur durch den obigen speziellen Ansatz von $r(\xi,\eta)$ als quadratisches Polynom bestimmt. Über die Konstanten a, b, c, d gehen in K lediglich Formparameter des Dreiecks Q ein, die lageunabhängig sind. Damit können wir durch eine geschickte Zerlegung von Ω und durch günstige Numerierung der Eckpunkte erreichen, daß die Elementsteifigkeitsmatrix für alle Dreiecke konstant bleibt. Dies führt zu einer wesentlichen Vereinfachung des Verfahrens, die bei der rechentechnischen Auswertung von Bedeutung ist. Aus den Elementsteifigkeitsmatrizen (4) erhalten wir durch Überlagerung das Gleichungssystem

$$K^* w^* = \delta^*. \qquad (10)$$

K^* ist die Steifigkeitsmatrix und w^* ist ein Vektor, in dem jede Unbekannte des Gebietes Ω genau einmal auftritt. Die Zusammenhangsmatrizen zwischen den Dreiecken gestatten dabei den Aufbau aus den Elementsteifigkeitsmatrizen und dem Vektor δ.

4. Numerierung und Zusammenhangsmatrizen

Betrachten wir zwei benachbarte Dreiecke, so müssen wir beim Aufbau von K^* bzw. δ^* berücksichtigen, daß gewisse Gitterpunkte in beiden Dreiecken auftreten. Ferner müssen wir wissen, welche Gitterpunkte auf Γ liegen, da hier die Randbedingungen eingehen. Dies Problem kann man durch eine von Mittelmann /4/ vorgeschlagene Doppelnumerierung lösen:

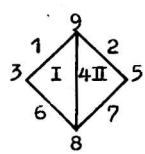
1. Die Gitterpunkte jedes Dreiecks werden von 1 bis 6 durchnumeriert.

2. Die Gitterpunkte, die in Ω liegen, werden von 1 bis k und die, die auf Γ liegen, von k+1 bis n durchnume- (11)
riert.

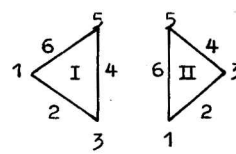
Um aus (4) das Gleichungssystem (10) aufzubauen, müssen wir die Abbildung $w \to w^*$ für alle Dreiecke der Zerlegung kennen. Dies läßt sich z. B. mit den Zusammenhangsmatrizen realisieren.

Definition: Eine Matrix $V = (v_{ij})$ für $j = 1, \ldots, m$ und $i = 1, \ldots, 6$, wobei m die Anzahl der Dreiecke der Zerlegung ist, heißt Zusammenhangsmatrix, wenn durch sie der Zusammenhang zwischen den beiden Numerierungen der Doppelnumerierung (11) in folgender Weise gegeben ist: Aus $v_{ij} = k$ folgt, daß der i-te Gitterpunkt des j-ten Dreiecks in der Gesamtnumerierung der k-te Punkt ist.

Beispiel: Gesamtnumerierung



Doppelnumerierung



Es ergibt sich:

$$V = \begin{bmatrix} 3 & 6 & 8 & 4 & 9 & 1 \\ 8 & 7 & 5 & 2 & 9 & 4 \end{bmatrix}^T .$$

Mit Hilfe von V kann man K^* folgendermaßen aus den K_{Q_i} (K_{Q_i} ... Elementsteifigkeitsmatrix des i-ten Dreiecks Q) der einzelnen Dreiecke aufbauen (Algol-Schreibweise):

$K^* := (0)$ (Nullmatrix).
Für $i = 1, \ldots, m$, $j = 1, \ldots, 6$, $k = 1, \ldots, 6$, (12)
$K^*[v[j,i], v[k,i]] := K^*[v[j,i], v[k,i]] + K_{Q_i}[j,k]$.

Analog erhält man für δ^* (Algol-Schreibweise):

$\sigma^* := (0)$.

Für $i = 1, \ldots, m$, $j = 1, \ldots, 6$, (13)

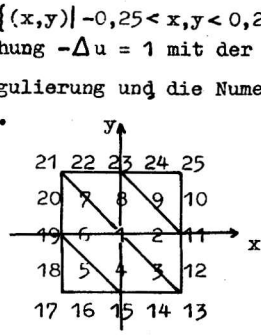
$\sigma^*[v[j,i]] := \sigma^*[v[j,i]] + \sigma_i[j]$.

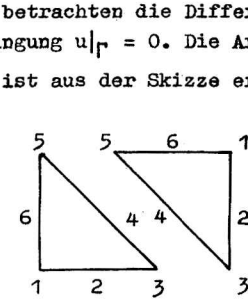
Damit ist das hergeleitete Verfahren aufbereitet. Ist das Gebiet Ω in Dreiecke zerlegt und eine Numerierung nach (11) vorgenommen, ergibt sich folgender Algorithmus:

1. Eingabe der Matrizen A, B, C, D, X, V, S^{-1} und der Konstanten c_0. X ist eine Matrix, die die Koordinaten der Eckpunkte jedes Dreiecks enthält.
2. Für jedes Dreieck Q erhält man K_Q nach (8a).
3. Mit Hilfe eines Quadraturverfahrens wird β_Q nach (9) und σ_Q nach (8b) berechnet.
4. Mit (12) und (13) wird K^* und σ^* erzeugt.
5. Das Gleichungssystem (10) wird mit einem Iterationsverfahren gelöst.

5. Numerische Ergebnisse

Sei $\Omega = \{(x,y) | -0{,}25 < x,y < 0{,}25\}$. Wir betrachten die Differentialgleichung $-\Delta u = 1$ mit der Randbedingung $u|_\Gamma = 0$. Die Art der Triangulierung und die Numerierung ist aus der Skizze ersichtlich.





Da alle Dreiecke kongruent sind, ist $a = c = 1$, $b = d = 0$ und $J = 0{,}0625$. Somit ergibt sich die Elementsteifigkeitsmatrix K zu:

$$K = \frac{1}{6}\begin{bmatrix} 6 & -4 & 1 & 0 & 1 & -4 \\ -4 & 16 & -4 & -8 & 0 & 0 \\ 1 & -4 & 3 & 0 & 0 & 0 \\ 0 & -8 & 0 & 16 & 0 & -8 \\ 1 & 0 & 0 & 0 & 3 & -4 \\ -4 & 0 & 0 & -8 & -4 & 16 \end{bmatrix} \qquad (14)$$

K kann somit eingegeben werden. Dies bedeutet bei der gewählten Triangulierung eine wesentliche Vereinfachung des Verfahrens. Weiter ist $\beta = J\,(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{24}, \frac{1}{12})^T$ und damit $\sigma = \frac{J}{6}\,(0, 1, 0, 1, 0, 1)^T$. Durch Programmierung der Formeln (12) und (13) werden K^* und σ^* aufgebaut. Da die Randbedingungen homogen sind, ist K^* eine (9 x 9) Matrix und σ^* ein Vektor vom Format (9 x 1).

Die Lösung des Gleichungssystems erfolgte nach dem Gauß-Seidel-Verfahren, wobei anschließend die Rechnung mit halber Schrittweite vorgenommen wurde. Die Numerierung begann im Ursprung und wurde im Uhrzeigersinn fortgesetzt. Statt 9 inneren Gitterpunkten lagen 49 vor, so daß der Aufwand sprunghaft anstieg. Aus diesem Grunde nutzen wir die Symmetriebeziehungen der Lösungsfunktion des Torsionsproblems (Membrangleichung) aus. Damit reduziert sich die Anzahl der Unbekannten von 49 auf 16. Nach Szabo /7/ ist die exakte Lösung des Problems:

$$u(x,y) = -0{,}5\left(y^2 - \frac{1}{16}\right)$$

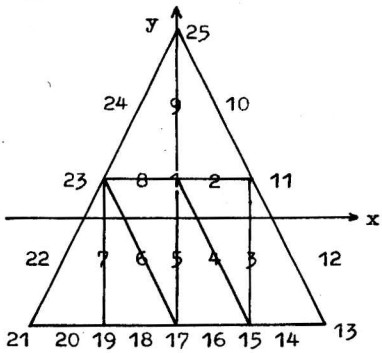
$$+ \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n+1)^3\,\pi^3\,\cosh\left[(2n+1)\frac{\pi}{2}\right]} \cos\left[2(2n+1)\pi y\right] \cosh\left[2(2n+1)\pi x\right].$$

Damit ergibt sich folgende Tabelle:

Nummer des Punktes	exakte Lösung	erste Rechnung	rel. Fehler	zweite Rechnung	rel. Fehler
1	0,0184178	0,0187465	1,78 %	0,0183908	0,15 %
2	0,0143337	0,0140608	1,90 %	0,0143216	0,08 %
3	0,0113215	0,0109364	3,40 %	0,0113322	0,09 %

Betrachten wir nun das Problem $-\Delta u = 1$, $u|_\Gamma = 0$.
Ω sei das gleichseitige Dreieck mit den Eckpunkten (0;0,6),

$(\frac{3}{10}\sqrt{3}; -0,3)$, $(-\frac{3}{10}\sqrt{3}; -0,3)$. Wegen der Kongruenz der Dreiecke der Zerlegung (siehe Skizze) ist die Elementsteifigkeitsmatrix K wiederum für alle Dreiecke gleich (bereits unter (14) angegeben).



Die exakte Lösung ist:

$u(x,y)$

$= \frac{1}{4h}(y^3 - hy^2 - 3x^2y - hx^2 + \frac{4}{27}h^3)$

mit $h = 0,9$.

Wir geben folgende Übersicht:

Nummer des Punktes	exakte Lösung	Näherung	rel. Fehler
1	0,0253125	0,0253099	0,010 %
2	0,0189844	0,0189831	0,007 %
3	0,0158203	0,0158193	0,004 %
4	0,0253125	0,0253196	0,028 %
5	0,0284766	0,0284750	0,006 %
6	0,0253125	0,0253114	0,004 %
7	0,0158203	0,0158199	0,003 %
8	0,0189844	0,0189833	0,006 %
9	0,0094922	0,0094919	0,003 %

Literatur

/1/ Bramble, J. H., and Zlamal, M.:
Triangular elements in the finite element method. Math. Comp. 24, 809 - 821 (1970)

/2/ Courant, R.: Variational methods for the solution of equilibrium and vibrations. Bull. Amer. Math. Soc. 49, 1 - 23 (1943)

/3/ Michlin, S. G.: Variationsmethoden der mathematischen Physik. Berlin 1962

/4/ Mittelmann, A. D.: Numerische Behandlung des Minimalflächenproblems mit finiten Elementen. In: Albrecht, J., Collatz, L. (Ed.): Finite Elemente und Differenzenverfahren. Basel-Stuttgart 1975

/5/ Moan, T.: Experiences with orthogonal polynomials and "best" numerical formulas on a triangle; with particular reference to finite element approximations. Z. Angew. Math. Mech. 54, 501 - 508 (1974)

/6/ Moldenhauer, W.: Ein spezielles Dreieckselement beim Finite-Elemente-Verfahren. Rostock. Math. Kolloq. 12, 87 - 99(1979)

/7/ Szabo, I.: Höhere technische Mechanik. Berlin- Göttingen-Heidelberg 1958

/8/ Koukal, S.: Piecewise polynomial interpolations in the finite element method. Apl. Mat. 18. 146 - 160 (1973)

/9/ Zlamal, M.: On some finite element procedures for solving second order boundary value problems. Numer. Math. 14, 42 - 48 (1969)

/10/ Zlamal, M.: On the finite element method. Numer. Math. 12, 394 - 409 (1968)

eingegangen: 26. 02. 1979

Anschrift der Verfasser:

Dr. Wolfgang Moldenhauer
DDR-25 Rostock 9
Landreiterstraße 8

Stud.-Math. Raimond Strauß
Wilhelm-Pieck-Universität Rostock
Sektion Mathematik
DDR-25 Rostock
Universitätsplatz 1

Rostock. Math. Kolloq. 12, 113 - 120 (1979)

Dieter Pötschke

Über die mittlere Länge und die Anzahl optimaler regulärer Suchcodes [1]

Der Begriff des Suchcodes wurde von Rényi /8/ eingeführt. Suchcodes sind spezielle Präfixcodes, die einer bestimmten Verzweigungsbedingung genügen und denen sich damit spezielle baumartige, kanteninterpretierte Graphen zuordnen lassen. Da bei der Anwendung des Algorithmus von Huffman auf eine Wahrscheinlichkeitsverteilung reguläre Suchcodes (bei denen jeder Verzweigungspunkt die gleiche Verzweigungszahl N, $N \geq 2$, besitzt) entstehen, ist die Konstruktion von optimalen regulären Suchcodes der von optimalen, eindeutig decodierbaren, längenvariablen Codes äquivalent.

Suchcodes sind sowohl in der Codierungstheorie als auch in der Theorie der Suchalgorithmen (vgl. Knuth /4/, Picard /5/) von Interesse.

In der vorliegenden Arbeit soll eine kurze Übersicht über die Resultate zur Berechnung der mittleren Weglänge (als explizite Formel) optimaler regulärer Suchcodes bei vorgegebenen Parametern, wie Anzahl der Endknoten (der den Suchcodes entsprechenden Bäume), Umfang, maximle Länge usw., gegeben werden. Insbesondere erweist sich die Abzählung von Suchcodes als Problem der Abzählung entsprechender baumartiger Graphen. Eine ausführliche Übersicht findet sich in Kap. 1 der Monographie Pötschke/Sobik /7/. Außerdem werden hier erste Resultate zur Konstruktion von Suchcodes für den Fall angegeben, daß anstelle der wahren Verteilung π nur eine Hypothese δ (Schätzung) bekannt ist.

[1] Kurzfassung eines Vortrags auf dem Symposium "Diskrete Mathematik und Anwendungen in der mathematischen Kybernetik", Rostock, April 1978

Sei X ein endliches Alphabet und $[X,\pi]$ eine unabhängige, stationäre Quelle. Sei Y ein weiteres endliches Alphabet, dessen Anzahl N kleiner als die von X ist, und W(Y) die Menge aller endlichen Wörter über Y einschließlich des leeren Wortes e. Dann ist W(Y) eine freie Halbgruppe mit dem freien Erzeugendensystem Y.

Sei $\gamma : X \to W(Y)$ eine injektive Funktion, dann ist die durch

$$\gamma^*(e) = e ,$$
$$\gamma^*(x_1 \ldots x_n) = \gamma(x_1) \ldots \gamma(x_n) , \quad x_1 \ldots x_n \in W(X) ,$$

$n \geq 1$, definierte Fortsetzung von γ auf W(X) ein Homomorphismus von der freien Halbgruppe W(X) in die freie Halbgruppe W(Y). γ^* wird als kombinatorische Codierung von X mittels Y bezeichnet. Im allgemeinen ist aber γ^* kein Isomorphismus auf der Halbgruppe $W(\gamma(X))$ mit dem freien Erzeugendensystem $\gamma(X)$, d. h., $\gamma(X)$ ist kein freies Erzeugendensystem dieser Halbgruppe.

Definition 1: Eine kombinatorische Codierung γ^* heißt genau dann eindeutig decodierbar, wenn $W(\gamma(X))$ eine freie Halbgruppe mit $\gamma(X)$ als freiem Erzeugendensystem ist, d. h., wenn die Elemente von $W(\gamma(X))$ eindeutig durch Elemente von $\gamma(X)$ darstellbar sind.

Selbst wenn γ eine injektive Abbildung ist, muß γ^* nicht eindeutig decodierbar sein. Denn sei für $X = \{x_1, x_2, x_3\}$ γ die folgende injektive Abbildung : $x_1 \to 1$, $x_2 \to 0$, $x_3 \to 01$, dann ist γ^* nicht eindeutig decodierbar. Es genügt auch nicht, für eindeutige Abbildungen γ die eindeutige Decodierbarkeit zu fordern. Bei der Zuordnung $x_1 \to 0$, $x_2 \to 0$, $x_3 \to 1$ ist γ zwar eindeutig, aber nicht eindeutig decodierbar.

Wir bezeichnen

$$CAW(\gamma,\pi) = \sum_{x \in X} \pi(x) |\gamma(x)|$$

als (mittleren) Codieraufwand der eindeutig decodierbaren Codierung γ bezüglich der Wahrscheinlichkeitsverteilung π.

Es ist bekannt, daß die Shannon-Entropie, die eine wichtige Rolle in der klassischen Theorie der Informationsübertragung spielt (vgl. Pötschke /6/), eine untere Schranke des Codieraufwandes darstellt, d. h., es gilt

$$H(X,\pi) \leq CAW(\gamma,\pi),$$

mit

$$H(X,\pi) =_{Df.} -\sum_{x\in X} \pi(x) \log \pi(x) .$$

Der Algorithmus von Huffman /3/ führt, auf π angewendet, bekanntlich zu einem optimalen Code, bei dem also die Differenz zwischen Codieraufwand und Entropie minimal ist. Will man nun die mittlere Länge eines Suchcodes (Def. 3) bzw. den Codieraufwand eines Codes ermitteln, so mußte man bisher mit Hilfe des Algorithmus von Huffman einen optimalen Code konstruieren und dessen Codieraufwand ausrechnen. In einigen Fällen kann aber eine explizite Formel für den Codieraufwand angegeben werden, und diese gestattet es, den minimalen Codieraufwand direkt aus π zu berechnen. Dies soll hier für eine spezielle Klasse von Verteilungen gezeigt werden. Für ein Wahrscheinlichkeitsmaß α über X mit $M = |X| > 2$ formulieren wir die

Bedingung A: Es gilt $\alpha(x_1) \geq \alpha(x_2) \geq \ldots \geq \alpha(x_M)$ für $M > 2$ und

$$\alpha(x_1) < \alpha(x_{M-1}) + \alpha(x_M).$$

Satz 1: Wenn ein Wahrscheinlichkeitsmaß π über X der Bedingung A genügt, dann gilt für den mit Hilfe des Huffmanschen Algorithmus konstruierten Code γ_π

$$CAW(\gamma_\pi,\pi) =]ld\ M[- \sum_{i=1}^{s} \pi(x_i) \qquad (1)$$

mit $s = -M + exp_2]ld\ M[$, wobei $]c[$ die kleinste ganze Zahl, die größer oder gleich c ist, bedeutet.

Beispiel: Sei $\pi(x_1) = 0{,}35$, $\pi(x_2) = 0{,}29$, $\pi(x_3) = 0{,}2$, $\pi(x_4) = 0{,}16$; wegen Satz 1 gilt $CAW(\gamma_\pi,\pi) = 2{,}0$.

Aus Formel (1) kann man auch die Codewortlängen entnehmen, mit deren Hilfe ein bezüglich π optimaler Code konstruiert werden kann.

Andere Fälle für explizite Formeln des minimalen Codieraufwandes findet man bei Pötschke/Sobik /7/, Kap. 1, und Geçkinli /2/. Für den Fall, daß π dem Coder nicht bekannt ist und er nur eine Hypothese σ über die Verteilung π besitzt, gilt der leicht zu beweisende

Satz 2: Wenn π und σ die Bedingung A erfüllen, dann haben die mit Hilfe des Algorithmus von Huffman konstruierten Codes γ_π und γ_σ den gleichen Codieraufwand, der sich durch (1) berechnen läßt.

Trotz der ungenauen Kenntnis der wahren Verteilung ergibt sich also aus σ ein bezüglich π optimaler Code. Das folgende Beispiel zeigt, daß eine gute Approximation σ' (im Sinne des Variationsabstandes) der Hypothese σ an die wahre Verteilung π nicht hinreichend dafür ist, daß sich derselbe Codieraufwand ergibt.

Beispiel: Sei

$\pi(x_1) = 0{,}35$	$\sigma(x_1) = 0{,}30$	$\sigma'(x_1) = 0{,}35$
$\pi(x_2) = 0{,}29$	$\sigma(x_2) = 0{,}29$	$\sigma'(x_2) = 0{,}31$
$\pi(x_3) = 0{,}20$	$\sigma(x_3) = 0{,}28$	$\sigma'(x_3) = 0{,}18$
$\pi(x_4) = 0{,}16$	$\sigma(x_4) = 0{,}13$	$\sigma'(x_4) = 0{,}16$.

Für den Variationsabstand gilt $\|\pi - \sigma\| = 0{,}16$, $\|\pi - \sigma'\| = 0{,}04$. Aus Satz 2 folgt $CAW(\gamma_\pi, \pi) = CAW(\gamma_\sigma, \pi) = 2{,}0$. Durch Anwendung des Huffmanschen Algorithmus ergibt sich aber $CAW(\gamma_{\sigma'}, \pi) = 2{,}01$.

Das Problem einer expliziten Formel für den Codieraufwand optimaler Codes ist bisher nicht vollständig gelöst. Es ist auch nicht bekannt, wie die Anzahl $H(n)$ der nicht längenäquivalenten Codebäume wächst.

Wir wenden uns nun der allgemeineren Frage zu, wie man die überhaupt auftretenden Codebäume abzählen kann, d. h., wir lassen die Längenäquivalenz zu. Dazu müssen wir "kompaktifizierte" Codes einführen, die wir als regulär verzweigt bezeichnen. Für $Y = nz$ heiße jede endliche Menge $C \subseteq W(Y)$ Code. Es sei

$$C_p =_{Df.} \{q \mid q \in W(Y) \wedge pq \in C\} \text{ für } p \in W(Y).$$

Offensichtlich gilt

$C_e = C$ und

$(C_p)_q = C_{pq}$ für $p,q \in W(Y)$.

Wenn C_p für jedes $p \in C$ der triviale Code ist, d. h. nur e enthält, dann heißt C Präfixcode.

Definition 2: Ein Code C heißt verzweigt, wenn folgendes gilt:
1. C ist der leere Code oder
2. C ist der triviale Code oder
3. C enthält nicht das leere Wort, und es existiert eine ganze Zahl v(C) - die Verzweigungszahl von C - mit $v(C) \geq 2$, so daß für $k < v(C)$ der Code C_k nicht leer, für $k \geq v(C)$ dagegen leer ist, k = 0, 1, , ...

Für den leeren Code setzen wir v(C) = 0 und für den trivialen Code v(C) = 1. Damit ist die Verzweigungszahl für jeden verzweigten Code definiert.

Definition 3: Ein Code heißt Suchcode genau dann, wenn C_p für jedes $p \in W(Y)$ verzweigt ist.

$C = \{0, 10, 2\}$ ist ein Pseudo-Suchcode im Sinne von Chorneyko und Mohanty /1/, aber kein Suchcode. Offensichtlich ist jeder Suchcode C ein Präfixcode, aber nicht umgekehrt: So ist jeder Code, der aus einem einzelnen, nichtleeren Codewort besteht, ein Präfixcode, aber kein Suchcode.
Wenn C ein Suchcode ist, so bezeichnen wir alle $p \in W(Y)$ mit $v(C_p) \geq 2$ als Verzweigungspunkte von C. Die Anzahl aller Verzweigungspunkte von C bezeichnen wir mit V(C). Ein Suchcode C heißt regulär vom Grade N, wenn jeder seiner Verzweigungspunkte die Verzweigungszahl N hat, d. h., für jedes $p \in W(Y)$ gilt: Wenn $v(C_p) \geq 2$ ist, dann gilt $v(c_p) = N$.

Wenn C ein regulärer Suchcode vom Grade N ist, dann genügt die Anzahl A(C) des Codes der Relation

$A(C) \equiv 1 \bmod (N-1)$.

Der folgende Satz gibt eine Formel für die Anzahl $C_N(n)$ der regulären Suchcodes vom Grad N und der Anzahl n an, wobei na-

türlich $n \equiv 1 \bmod (N-1)$ ist.

Satz 3: (Rényi /8/): Für jede natürliche Zahl n mit $n = k(N-1) + 1$ $(k=0,1,2,\ldots)$ gilt

$$C_N(n) = \frac{(kN)!}{k!\,n!} \,. \tag{2}$$

Im Falle N = 2, also binärer Suchcodes, gilt A(C) = V(C) + 1; damit existieren reguläre binäre Suchcodes jeder Anzahl $n \geq 1$, und aus (2) ergibt sich für ihre Gesamtanzahl

$$C_2(n) = \frac{\binom{2n-1}{n}}{(2n-1)} = \frac{(2n-3)!!\,2^{n-1}}{n!} \,. \tag{3}$$

Diese Formel war schon Cayley (1857/59) bekannt, der das entsprechende Problem für die den Suchcodes entsprechenden Codebäume löste. Cayley wies auch darauf hin, daß $C_2(n)$ gleich der Anzahl der möglichen Interpretationen von einem "Produkt" von n Faktoren bezüglich einer nichtassoziativen Operation ist, d. h., $C_2(n)$ ist gleich der Anzahl der möglichen Klammerungen eines derartigen Produktes.

Sei D(n,k) die Gesamtanzahl aller Suchcodes des Umfanges n mit k Verzweigungspunkten. Offensichtlich gilt

$$k \leq n - 1,$$

da für eine gegebene Anzahl n der binäre Code die maximale Anzahl von Verzweigungspunkten besitzt. Man sieht leicht, daß $D(1,0) = 1$ und $D(n,0) = 0$ ist für $n \geq 2$. Für $n \geq 2$ gilt die Rekursionsformel

$$D(n,k) = \sum_{l=2}^{n} \sum_{\substack{n_1+\ldots+n_l=n \\ k_1+\ldots+k_l=k-1}} D(n_1,k_1)D(n_2,k_2)\ldots D(n_l,k_l) \,.$$

Mit Hilfe dieser Formel kann der folgende Satz bewiesen werden (vgl. Rényi /8/).

Satz 4: Für $n \geq 2$ und $k \geq 1$ gilt für die Anzahl D(n,k) der Suchcodes vom Umfang n mit k Verzweigungspunkten

$$D(n,k) = \frac{1}{n} \binom{n-2}{k-1} \binom{n+k-1}{k} \,.$$

Für $n \geq 2$ gilt für die Anzahl $D(n)$ aller Suchcodes vom Umfang n

$$D(n) = \frac{1}{n} \sum_{k=1}^{n-1} \binom{n-2}{k-1} \binom{n+k-1}{k} .$$

Insbesondere besteht für $D(n)$ die asymptotische Formel

$$D(n) \sim \frac{\sqrt{3-2\sqrt{2}}\,(3+2\sqrt{2})^n}{4\sqrt{\pi}\cdot n^{3/2}} .$$

Die ersten Werte von $D(n)$ sind $D(2) = 1$, $D(3) = 3$, $D(4) = 11$, $D(5) = 45$, $D(6) = 197$. Auch für $D_r(n)$, die Anzahl der Suchcodes vom Umfang n mit maximaler Länge r, $r = 1,2,\ldots$, und $B_N(n,r)$, die Anzahl der regulären Suchcodes vom Grade N und des Umfanges n, kann man explizite Formeln herleiten (vgl. Rényi /8/).-

Literatur

/1/ Chorneyko, I. Z., and Mohanty, S. G.: On the enumeration of pseudo-search codes. Studia Sci. Math. Hungar. 7, 47 - 54 (1972)

/2/ Geçkinli, N. C.: Two corollaries to the Huffamn coding procedure. IEEE Trans. Information Theory IT-21, 3, 342 - 344 (1975)

/3/ Huffman, D. A.: A method for the construction of minimum redundancy codes. Proc. of the IRE 40, 1098 - 1101 (1952)

/4/ Knuth, D. E.: Optimum binary search tress. Acta Informat. 1, 14 - 25 (1971)

/5/ Picard, C. F.: Graphes et questionnaires. I,II. Paris 1972

/6/ Pötschke, D.: On the classical approach to the Shannon transmission problem. In: Csiszar, I., and P. Elias (Ed.): Topics in Information Theory. pp. 527 - 538, Amsterdam 1977

/7/ Pötschke, D., und Sobik, F.: Neuere Ergebnisse der mathematischen Informationstheorie. Berlin (in Vorbereitung).

/8/ Rényi, A.: On the enumeration of search codes. Acta Math. Acad. Sci. Hungar. 21, 27 - 33 (1970)

eingegangen: 01. 12. 1978

Anschrift des Verfassers:

Dr. rer. nat. Dieter Pötschke
Akademie der Wissenschaften der DDR
Zentralinstitut für Kybernetik und
Informationsprozesse
DDR-1199 Berlin
Rudower Chaussee 5