KERNFORSCHUNGSZENTRUM

KARLSRUHE

Januar 1970

KFK 1132

Institut für Neutronenphysik und Reaktortechnik

Die numerische Berechnung von interpolierenden Spline-Funktionen mit Blockunterrelaxation

H. Späth

GESELLSCHAFT FÜR KERNFORSCHUNG M. B. H.

KARLSRUHE

KERNFORSCHUNGSZENTRUM KARLSRUHE

Januar 1970 KFK 1132

Institut für Neutronenphysik und Reaktortechnik

Die numerische Berechnung von interpolierenden Spline-Funktionen mit Blockunterrelaxation

von

H. Späth

Gesellschaft für Kernforschung m.b.H., Karlsruhe

Zusammenfassung:

Die numerische Berechnung von interpolierenden Spline-Funktionen für hohe Knotenzahlen und von einem Grad größer als drei ist für praktische Zwecke, z.B. für die Auswertung von nuklearen Meßdaten, sehr wichtig. Für diese Fälle versagen die bekannten Methoden aufgrund numerischer Instabilitäten. Wir leiten ein die Splines bestimmendes Gleichungssystem her, das mit einem iterativen Verfahren (Blockunterrelaxation) gelöst werden kann. In wichtigen Fällen kann man die optimalen Beschleunigungsparameter explizit angeben. Dieses Iterationsverfahren ist in der Rechenzeit mit den bekannten Methoden vergleichbar, bringt aber einen ganz erheblichen Fortschritt bezüglich der numerischen Stabilität.

Die Arbeit ist an der Universität Karlsruhe als Dissertation angenommen. Referenten sind Prof. Dr. J. Weissinger und Privatdozent Dr. H. Brakhage.

Summary:

The numerical calculation of interpolating spline functions for a high number of knots and of a degree greater than three is very important for practical purposes, i.e. for the evaluation of nuclear data points. For these cases the known methods fail because of numerical instabilities. We derive a system of linear equations characterizing the splines that is suitable for iterative solution (blockunderrelaxation). The optimal relaxation parameters can be obtained explicitly in important cases. This iterative method is comparable with the known direct methods concerning the computing time but results in quite an impressive progress with respect to numerical stability.

This work has been accepted as thesis at the University of Karlsruhe. Referents are Prof. Dr. J. Weissinger and Privatdozent Dr. H. Brakhage.

Inhaltsverzeichnis

0.	Einleitung	1
1.	Spline-Interpolation vom Typ K und vom Grad 2m+1. Problemstellung und Ergebnisse	3
2.	Herleitung eines Gleichungssystems für die Spline-Interpolation vom Typ III und m > 2	10
3.	Erweiterung des Gleichungssystems für m = 2 auf die Typen I und II. Mögliche Modifikationen für m > 2 und Typ K	19
4.	Lösung der Gleichungssysteme mit Blockrelaxationsverfahren. Optimale Beschleunigungsparameter	25
5.	Aufwand und Rundungsfehler im Vergleich zu bekannten Verfahren. Einfluß des Typs auf die Gestalt	40
Literatur		53

0. Einleitung

Die Berechnung von interpolierenden Spline-Funktionen für hohe Knotenzahlen und von einem Grad größer als drei ist für praktische Zwecke wichtig. Die bekannten Berechnungsverfahren versagen für diese Fälle aufgrund numerischer Instabilitäten. Die hier vorgeschlagenen iterativen Verfahren sind im Aufwand mit den herkömmlichen direkten Methoden vergleichbar, bringen jedoch einen ganz erheblichen Fortschritt bezüglich der numerischen Stabilität.

Einführend wird in Kapitel 1 der Begriff der Spline-Interpolation vom Typ K und vom Grad 2m+1 erläutert, der eine gewisse Verallgemeinerung der Standardtypen I und II darstellt, und angedeutet, wie sich die wichtigsten Sätze auf Typ K übertragen lassen. Dann können wir Problemstellung und Ergebnisse der Arbeit genau formulieren.

In Kapitel 2 wird für einen speziellen Typ K, den wir Typ III nennen, ein lineares Gleichungssystem hergeleitet und dessen zwei kanonische Blockstrukturen studiert.

Nach m = 1 ist für die Praxis besonders wichtig der Fall m = 2. In Kapitel 3 wird gezeigt, wie und weshalb sich genau für m = 2 das Gleichungssystem für Typ III auf einfache Weise so abändern läßt, daß auch die Standardtypen I und II damit berechnet werden können. Der Einfachheit halber beschränken wir uns für m > 2 im folgenden auf das Gleichungs-

system für Typ III, obwohl dieses prinzipiell für beliebige Typen K modifiziert werden kann.

Aufgrund der beiden natürlichen Blockstrukturen unseres Gleichungssystems bieten sich zur Lösung die iterativen Methoden der Blockrelaxation an. Je nach Blockung ist Blocküber- oder -unterrelaxation anwendbar. In Kapitel 4 werden für m = 2, beliebige Abszissen und die Typen I, II und III optimale Beschleunigungsparameter für das Blockunterrelaxationsverfahren angegeben. Für m > 2 ist die Bestimmung allgemein nur für äquidistante Abszissen durchführbar. Für Typ III wird dies beim Blocküber- und -unterrelaxationsverfahren getan. Die Blockunterrelaxation erweist sich als mindestens zweimal so schnell.

Schließlich werden in Kapitel 5 Aufwand und numerische Stabilität des Blockunterrelaxationsverfahren mit bekannten Methoden zur Spline-Interpolation vom Typ II verglichen. Das Iterationsverfahren ist bei vergleichbarem Rechenaufwand wesentlich günstiger in bezug auf numerische Stabilität. Der Einfluß des Typs auf die Gestalt der Spline-Funktion wird an einem Beispiel demonstriert.

1. Spline-Interpolation vom Typ K und vom Grad 2m+1. Problemstellung und Ergebnisse

Eine Spline-Funktion s von Grad 2m+1 auf einem Abszissengitter $a = x_1 < x_2 < \ldots < x_n = b$ besteht aus n-1 jeweils in $[x_i, x_{i+1}]$ definierten Polynomen, deren Grad höchstens 2m+1 ist, die an den Knotenstellen $x_2, \ldots, x_{n-1}$ 2m-mal stetig differenzierbar aneinandergesetzt sind.

Sind zu den Abszissen x_i Ordinaten y_i vorgegeben, durch die die Spline-Funktion verlaufen soll, so spricht man von Spline-Interpolation. Diese unterscheidet sich u.a. aufgrund ihrer Konvergenzeigenschaften vorteilhaft von anderen Interpolationsverfahren. Für m = 0 haben wir Polygonzüge und für m = 1 zweimal stetig differenzierbar aneinandergesetzte kubische Polynome.

Da jedes Polynom vom Grad 2m+1 durch 2m+2 Konstanten bestimmt ist, müssen, falls die Polynome jeweils durch ihre Koeffizienten charakterisiert werden, bei der Spline-Interpolation (2m+2)(n-1) Konstanten berechnet werden. Berücksichtigen wir, daß neben den Interpolationsbedingungen $s(x_i) = y_i$ (i=1,...,n) gelten soll, daß die Ableitungen der Ordnung $1, 2, \ldots, 2m$ an den Knotenstellen $x_2, \ldots, x_{n-1}$ übereinstimmen sollen, so bleiben noch 2m Freiheitsgrade. Daher gibt man nun im allgemeinen noch in a und b jeweils m Werte für Ableitungen bestimmter Ordnung vor, und zwar aus Symmetriegründen für solche der gleichen Ordnung in a und b. Noch allgemeinere Randbedingungen werden bei GREVILLE (1969) betrachtet. Wir greifen im folgenden eine bestimmte Klasse von (symmetrischen) Randbedingungen heraus:

(1.1) Definition: Spline-Interpolation vom Typ K

Sei $J_1 = \{1,\ldots,m\}$ und $J_2 = \{m+1,\ldots,2m\}$. Es bezeichne $T: J_1 \to J_2$ die durch $T(i) := 2m+1-i$ für $i \in J_1$ definierte bijektive Abbildung der beiden Indexmengen aufeinander. K sei eine Teilmenge von J_1.

Es seien Abszissen $a = x_1 < \ldots < x_n = b$ und zugehörige Ordinaten y_i und Zahlen $y_a^{(k)}$, $y_b^{(k)}$ gegeben.

Eine Spline-Funktion s vom Grad 2m+1 interpoliert vom Typ K, wenn gilt

i) $s(x_i) = y_i \qquad (i = 1,\ldots,n)$

ii) $s^{(k)}(a) = y_a^{(k)}$, $s^{(k)}(b) = y_b^{(k)}$ für $k \in K \subset J_1$

iii) $s^{(k)}(a) = s^{(k)}(b) = 0$ für $k \in J_2 - TK$

Durch Spezialisierung von K ergeben sich die folgenden drei wichtigen Beispiele:

(1.2) Definition: Eine Spline-Funktion vom Grad 2m+1 interpoliert vom

Typ I, wenn $K = J_1$
Typ II, wenn $K = \emptyset$ (leere Menge)
Typ III, wenn $K = \{2i : i = 1,\ldots,\lfloor \frac{m}{2} \rfloor\}$

Bei Typ I sind also in a und b die Ableitungen $s^{(k)}$ $(k=1,\ldots,m)$, bei Typ II die für $k = m+1,\ldots,2m$ mit den Zahlenwerten Null vorgegeben. Diese beiden Typen sind - neben den Splines mit periodischen Randbedingungen, für die wir und hier nicht interessieren - die in dem Standardwerk von AHLBERG, NILSON, WALSH (1967) betrachteten. Bei Typ III werden Werte für die geraden Ableitungen an den Rändern vorgegeben, die für $k > \lfloor \frac{m}{2} \rfloor$ gleich Null sind. Für $m > 1$ ist Typ III anscheinend noch nicht betrachtet worden.

Zunächst zeigen wir, daß für die Spline-Interpolation vom Typ K die in AHLBERG, NILSON, WALSH (1967) nur für Typ I und II formulierten Hauptsätze gelten. Da es sich um eine Übertragung der Grundideen bei Typ I und II auf Typ K handelt, werden wir Beweise nur skizzieren.

Die Einführung insbesondere von Typ III wird sich bei der Herleitung von Gleichungssystemen, die für Typ K, also insbesondere für die Typen I und II nur leicht modifiziert werden müssen, als vorteilhaft erweisen.

Es bezeichne $H^{m+1}[a,b]$ die Menge aller Funktionen f, die auf dem Intervall $[a,b]$ definiert sind, die dort eine absolut stetige m-te Ableitung besitzen und deren (m+1)-te Ableitung quadratisch integrierbar ist.

Zwei Funktionen $f,g \in H^{m+1}[a,b]$ nennen wir äquivalent bezüglich einer Indexmenge $K \subset J_1 = \{1,\dots,m\}$, in Zeichen $f \overset{K}{\sim} g$, wenn für alle $k \in K$ gilt $f^{(k)}(a) = g^{(k)}(a)$ und $f^{(k)}(b) = g^{(k)}(b)$. Die leere Menge $\emptyset$ ist für K zugelassen.

(1.3) Satz: Es sei $a = x_1 < \dots < x_n = b$, $m < n$ und $f \in H^{m+1}[a,b]$.

i) Dann existiert genau eine Spline-Funktion s, die f vom Typ K interpoliert

ii) Dieses s minimalisiert

$$\int_a^b [g^{(m+1)}(x)]^2 dx \tag{1.4}$$

für $g \in G_f := \{g \in H^{m+1}[a,b],\ g \overset{K}{\sim} f,\ f(x_i) = g(x_i), i=1,\dots,n\}$

Der interpolierende Spline s vom Typ II zeichnet sich also dadurch aus, daß (1.4) von s minimalisiert wird, ohne daß zusätzliche Forderungen an G_f bezüglich der Übereinstimmung von Werten für Ableitungen in a und b gestellt werden.

Der Beweis von (1.3) ist bei AHLBERG, NILSON, WALSH (1967) für Typ I und II durchgeführt. Aus der folgenden Beweisskizze wird deutlich, wie der Beweis auf Typ K übertragen werden kann:

Für eine Spline-Funktion s vom Grad 2m+1 und vom Typ K mit $s \overset{K}{\sim} f$ gilt:

$$\begin{aligned}
&\int_a^b \left[f^{(m+1)}(x)-s^{(m+1)}(x)\right]^2 dx \\
&= \int_a^b \left[f^{(m+1)}(x)\right]^2 dx - 2\int_a^b \left[f^{(m+1)}(x) - s^{(m+1)}(x)\right] s^{(m+1)}(x)dx - \\
&\qquad\qquad - \int_a^b \left[s^{(m+1)}(x)\right]^2 dx \\
&= \int_a^b \left[f^{(m+1)}(x)\right]^2 dx - \int_a^b \left[s^{(m+1)}(x)\right]^2 dx \\
&\quad - 2\sum_{i=1}^{n-1} (-1)^m \left[f(x)-s(x)\right] s^{(2m+1)}(x) \Big|_{x_i}^{x_{i+1}} \\
&\quad - 2\sum_{j=1}^{m} (-1)^{j+1} \left[f^{(m+1-j)}(x)-s^{(m+1-j)}(x)\right] s^{(m+j)}(x) \Big|_a^b
\end{aligned}$$

Bei der partiellen Integration wird die Stetigkeit der Terme im letzten Ausdruck benutzt. Der dritte Term verschwindet aufgrund der Interpolationsbedingungen und der vierte gerade aufgrund der Definition (1.1).

Somit haben wir die sogenannte erste Integralbeziehung

$$(1.5) \qquad \int_a^b \left[f^{(m+1)}(x)\right]^2 dx = \int_a^b \left[s^{(m+1)}(x)\right]^2 dx + \int_a^b \left[f^{(m+1)}(x)-s^{(m+1)}(x)\right]^2 dx$$

Daraus folgt dann (1.4) und daraus wiederum Existenz und Eindeutigkeit (AHLBERG, NILSON, WALSH (1967)).

Durch Anwendung des Theorems von ROLLE auf $f^{(i)}(x) - s^{(i)}(x)$ folgt zusammen mit der Minimaleigenschaft (1.4) (AHLBERG, NILSON, WALSH (1967)) der Satz

(1.6) Satz: Zu gegebenem Abszissengitter $a = x_1 < \ldots < x_n = b$ und gegebenem $f \in H^{m+1}[a,b]$ sei s der interpolierende Spline vom Typ K. Dann gilt die universelle Abschätzung

$$|f^{(i)}(x) - s^{(i)}(x)| < c_1(n,i,f,K)\,\beta^{(2m+1-2i)/2} \tag{1.7}$$

mit $c_1(n,i,f,K) < \infty$, $\beta := \max_i \Delta x_i$ und $i = 0,1,\ldots,m$.

Bei der Spline-Interpolation vom Typ K werden also die Funktion und ihre höheren Ableitungen simultan approximiert.

Bemerkung: Eine Verschärfung von (1.6) für $f \in H^{2m+2}[a,b]$ läßt sich durch Erweiterung des Aquivalenzbegriffes erreichen. Man setzt voraus $s \overset{\bar{K}}{\sim} f$ mit $\bar{K} = K \cup (J_2 - TK)$. Steht dann auf der rechten Seite in (1.7) der Exponent $2m+1-i$, so gilt (1.7) mit einer Konstanten $c_2 = c_2(n,i,f,K)$ für $i=0,1,\ldots,2m+1$. Die für den Beweis notwendige sogenannte zweite Integralbeziehung

$$\int_a^b [f^{(m+1)}(x) - s^{(m+1)}(x)]^2 dx = (-1)^{m+1} \int_a^b [f(x) - s(x)] f^{(2m+2)}(x) dx \tag{1.8}$$

wird auf dieselbe Art und Weise wie (1.5) ausgerechnet unter Verwendung von $f \in H^{2m+2}[a,b]$ anstelle von $f \in H^{m+1}[a,b]$.

Die Spline-Interpolation wird u.a. bei folgenden praktischen Problemen angewandt, wobei die Art der Randbedingungen, also der Typ, im allgemeinen nicht so wesentlich ist, da diese die Gestalt der Funktion nur in der Umgebung von a und b beeinflußt.

Wegen der durch (1.4) in einem gewissen Sinne definierten Glattheit (für m = 1 kann man sich vorstellen, daß eine mittlere zweite Ableitung minimalisiert wird) und der Konvergenzeigenschaften (1.6) eignet sie sich hervorragend zur Auswertung (z.B. Differentiation und Integration) von Meßdaten. So kann man sie z.B. in der Multigruppentheorie der Reaktorphysik dazu verwenden, um zu einer vorgegebenen Treppenfunktion eine glatte, in jedem Intervall $[x_i, x_{i+1}]$ flächengleiche Funktion anzugeben (ANSELONE, LAURENT (1968), SPÄTH (1968)). Hierbei treten nicht äquidistante Abszissen mit einer Anzahl $n \leq 200$ auf. In der Kernphysik werden Vielkanalanalysatoren dazu benutzt, Impulshöhenverteilungen und Flugzeitspektren zu messen. Da bei derartigen Messungen die Kanalnummer die Rolle der Abszisse spielt, hat man äquidistante Abszissen mit $n \approx 1000$.

Passt man experimentelle Daten mit Spline-Funktionen im Sinne der kleinsten Quadrate an (SCHOENBERG (1964)), so ist im Verlauf der iterativen Minimalisierung nach REINSCH (1967) eine wiederholte Spline-Interpolation bei gleichen Abszissen notwendig. CURTIS, POWELL (1967) und BOOR, RICE (1968) benutzen die Spline-Interpolation zur Approximation von Funktionen.

Für diese Zwecke ist es wünschenswert, für eine große Anzahl von Abszissen interpolierende Splines, auch für $m > 1$, bestimmen zu können.

Die bekannten numerischen Verfahren (CARASSO (1966), CARASSO, LAURENT (1968), ANSELONE, LAURENT (1968)) berechnen Typ II. Abgesehen von $m = 1$, wo ein lineares Gleichungssystem mit positiv definiter Matrix auftritt, dessen Lösung für beliebige n keine numerischen Schwierigkeiten bildet, sind die beschriebenen Verfahren bei den üblichen Wortlängen von 7-16 Dezimalen nur für $2 \leq m \leq 4$ und für $n \leq N_o(m)$ brauchbar, wobei die Zahl N_o mit wachsendem m fällt. Es gilt ungefähr $60 < N_o < 100$ (CARASSO (1967), CARASSO, LAURENT (1968)). Die Koeffizienten der auftretenden linearen Gleichungssysteme sind dividierte Differenzen der Ordnung 2m+2 und ihre rechten Seiten solche der Ordnung m+1. Bei endlicher Zahlenlänge können diese nur ungenau berechnet werden. Erfahrungsgemäß ist der Einfluß der Rundungsfehler in den rechten Seiten kritischer.

Wir geben nun für Typ III ein Gleichungssystem an, dessen rechte Seiten bei gleicher Zahlenlänge genauer berechnet werden können, da unabhängig von m stets nur Differenzen von zwei Differenzenquotienten auftreten. Für m = 2 läßt sich das Gleichungssystem einfach für die Typen I und II modifizieren. Prinzipiell ist dies auch für m > 2 und beliebige Typen K möglich.

Da die Koeffizientenmatrix in allen Fällen zwei natürliche Blockstrukturen besitzt, wobei einmal in den Blöcken relativ wenige von Null verschiedene Elemente und im anderen Fall nur wenige von Null verschiedene Blöcke auftreten, wenden wir zur Auflösung der Gleichungssysteme iterative Verfahren der Blockrelaxation (VARGA (1962)) an. Im einzelnen erreichen wir:

Für Typ III und äquidistante Abszissen erhalten wir für $2 \leq m \leq 10$ und $n \leq 1000$ mittels einer empirischen, durch theoretische Betrachtungen abgestützten Methode angenähert optimale Beschleunigungsparameter für ein Blockunterrelaxationsverfahren und für $2 \leq m \leq 10$ und beliebige n exakte optimale Beschleunigungsparameter für ein Blocküberrelaxationsverfahren. Bei gleichem Rechenaufwand pro Iteration benötigt die Blockunterrelaxation etwa die halbe Anzahl der Schritte im Vergleich zur Blocküberrelaxation.

Für den nach m = 0 und m = 1 nächstwichtigen Fall m = 2 erhalten wir für die Typen I, II und III und beliebige Abszissen die exakten optimalen Beschleunigungsparameter bei Blockunterrelaxation für beliebige n.

Dies ist unseres Wissens der erste Vorschlag (außer für m = 1 (GREVILLE (1967)), wo es unnötig ist) Spline-Funktionen mit iterativen Methoden zu berechnen und einer der wenigen bekannten Anwendungsfälle, wo sich Blockunterrelaxation als vorteilhaft erweist.

2. Herleitung eines Gleichungssystems für die Spline-Interpolation vom Typ III und $m \geq 2$

Die verschiedenen Methoden zur Berechnung von interpolierenden Spline-Funktionen (CARASSO, LAURENT (1968)) unterscheiden sich in der Wahl und der davon abhängigen Anzahl der Unbekannten. Dabei sind, wie sich in Kapitel 5 zeigen wird, die Verfahren mit geringerer Anzahl von Unbekannten nicht notwendigerweise die zweckmäßigsten.

Für die numerische Auswertung ist es günstig (BOOR, RICE (1968)), den Spline im Intervall $[x_i, x_{i+1}]$ $(i=1,\ldots,n-1)$ in der Form

$$(2.1) \qquad P_{i-1}(t) = \sum_{k=0}^{2m+1} A_k^{(i-1)} (t-x_{i-1})^k \qquad (i=2,\ldots,n)$$

zur Verfügung zu haben. Die $(2m+2)(n-1)$ Koeffizienten $A_k^{(i)}$ sind durch die Interpolationsbedingungen, durch die Übergangsbedingungen

$$(2.2) \qquad P_{i-1}^{(j)}(x_i) = P_i^{(j)}(x_i) \qquad \begin{pmatrix} i=2,\ldots,n-1 \\ j=1,\ldots,2m \end{pmatrix}$$

und schließlich durch die Randbedingungen vom Typ K nach Satz (1.3) eindeutig bestimmt.

Führt man wie HOLLODAY (1957) die $A_k^{(i)}$ jedoch als Unbekannte ein, so erhält man ein i.a. äußerst schlecht konditioniertes Gleichungssystem (CARASSO, LAURENT (1968)).

Die Anzahl der Unbekannten reduziert sich auf die Hälfte, wenn man bei Splines vom Typ K die Werte $y_i^{(k)}$ $(i=2,\ldots,n-1,\ k \in K \cup (J_2 - TK))$ als Unbekannte einführt,

die von den Randbedingungen her für i=1 und i=n bekannt sind. In Kapitel 3 werden wir sehen, wann dies prinzipiell möglich ist. Im folgenden wird gezeigt, daß jedenfalls speziell für die Spline-Interpolation vom Typ III dieses Vorgehen sinnvoll ist, sich die Koeffizienten $A_k^{(i)}$ einfach durch die Unbekannten $y_i^{(2k)}$ ausdrücken lassen, und man für die Unbekannten ein Gleichungssystem mit günstigen Eigenschaften für die numerische Behandlung erhält.

Wir setzen (2.1) in anderer Form an:

$$P_{i-1}(t) = \sum_{k=o}^{m} h_{i-1}^{2k} \left[y_{i-1}^{(2k)} L_k(1-u) + y_i^{(2k)} L_k(u) \right] \tag{2.3}$$

$$\text{mit } h_{i-1} = \Delta x_{i-1} \text{ und } u = \frac{t-x_{i-1}}{h_{i-1}}$$

Dabei sind die Polynome $L_k(u)$ (LIDSTONE (1930)) definiert durch

$$\begin{aligned} L_o(u) &= u \\ L_k''(u) &= L_{k-1}(u) \qquad (k \geq 1) \\ L_k(o) &= L_k(1) = 0 \end{aligned} \tag{2.4}$$

Aufgrund von (2.4) hat (2.3) die gewünschten Eigenschaften

$$P_{i-1}^{(2k)}(x_{i-1}) = y_{i-1}^{(2k)} \text{ und } P_{i-1}^{(2k)}(x_i) = y_i^{(2k)}$$

$$(i = 2,\ldots,n; \quad k = 0,\ldots,m)$$

Nun gilt (WHITTAKER (1934))

$$L_k(u) = \frac{2^{2k+1}}{(2k+1)!} \beta_{2k+1}\left(\frac{1+u}{2}\right) \qquad (k \geq 1) \tag{2.5}$$

wobei mit β die BERNOUILLI'schen Polynome

$$(2.6) \qquad \beta_j(x) = \sum_{k=0}^{j} \binom{j}{k} B_k \, x^{j-k} \qquad (j=0,1,\ldots)$$

bezeichnet sind, worin die BERNOUILLI'schen Zahlen B_k definiert sind durch

$$(2.7) \qquad B_0 = 1, \; B_1 = -\frac{1}{2}, \; \sum_{k=0}^{j-1} \binom{j}{k} B_k = 0 \qquad (j=2,3,\ldots)$$

Es gelten die Beziehungen (NÖRLUND (1954))

$$(2.8) \qquad \beta_j(x+1)-\beta_j(x) = jx^{j-1}$$

$$(2.9) \qquad \beta_j(kx) = k^{j-1} \sum_{i=0}^{k-1} \beta_j\left(x+\frac{i}{k}\right)$$

Nach (2.6) erhalten wir

$$(2.10) \qquad \beta_j'(0) = \beta_j'(1) = jB_{j-1}$$

Setzen wir in (2.9) $k = 2$, differenzieren auf beiden Seiten, setzen $x = 0$ und verwenden (2.10), so erhalten wir

$$(2.11) \qquad \beta_j'\left(\frac{1}{2}\right) = \frac{j(1-2^{j-2})}{2^{j-2}} B_{j-1} \qquad (j \geq 1)$$

Wir berechnen nun die ungeraden Ableitungen von (2.3) und erhalten mit (2.4)

$$(2.12) \qquad P_{i-1}^{(2j+1)}(t) = \sum_{k=j}^{m} h_{i-1}^{2(k-j)-1} \left[y_i^{(2k)} L_{k-j}'(u) - y_{i-1}^{(2k)} L_{k-j}'(1-u) \right]$$

Setzen wir zur Abkürzung

$$(2.13) \qquad \begin{aligned} q_i &= L_i'(1) = \frac{2^{2i}}{(2i+1)!} \beta_{2i+1}'\left(1\right) = \frac{2^{2i}}{(2i)!} B_{2i} \\ r_i &= -L_i'(0) = \frac{2(2^{2i-1}-1)}{(2i)!} B_{2i} \qquad (i=0,1,\ldots) \end{aligned}$$

so erhält man durch Einsetzen von (2.12) und (2.13) in die Bedingungen (2.2) das gesuchte lineare Gleichungssystem

$$\sum_{k=o}^{m-j} \{r_k\, h_{i-1}^{2k-1}\, y_{i-1}^{(2(k+j))} \qquad (j=0,1,\ldots,m-1)\ (i=2,\ldots,n-1)$$

$$+ q_k (h_{i-1}^{2k-1} + h_i^{2k-1})\, y_i^{(2(k+j))} \tag{2.14}$$

$$+ r_k\, h_i^{2k-1}\, y_{i+1}^{(2(k+j))}\} = 0$$

zur Bestimmung der Unbekannten $y_i^{(2k)}$ $(i=2,\ldots,n-1;\ k=1,\ldots,m)$. Numerische Werte für q_k und r_k sind:

Tabelle 1: Zahlenwerte für die Konstanten q_k und r_k

k	q_k	r_k
0	1.0	-1.0
1	.3333333333	.1666666667
2	$-.2222222222_{10}-1$	$-.1944444444_{10}-1$
3	$.2116402116_{10}-2$	$.2050264550_{10}-2$
4	$-.2116402116_{10}-3$	$-.2099867724_{10}-3$
5	$.2137779916_{10}-4$	$.2133604564_{10}-4$
6	$-.2164404281_{10}-5$	$-.2163347443_{10}-5$
7	$.2192594785_{10}-6$	$.2192327134_{10}-6$
8	$-.2221460879_{10}-7$	$-.2221393085_{10}-7$
9	$.2250784651_{10}-8$	$.2250767480_{10}-8$
10	$-.2280515120_{10}-9$	$-.2280510771_{10}-9$

Die Koeffizienten $A_k^{(i-1)}$ in (2.1) erhalten wir aus den $y_{i-1}^{(2k)}$ und $y_i^{(2k)}$ mit (2.12) und (2.13) zu

$$A_{2k}^{(i-1)} = \frac{1}{(2k)!}\, y_{i-1}^{(2k)}, \quad A_{2k+1}^{(i-1)} = -\frac{1}{(2k+1)!} \sum_{j=k}^{m} h_{i-1}^{2(j-k)-1} \left(q_{j-k} y_{i-1}^{(2j)} + r_{j-k} y_i^{(2j)}\right)$$

$$(k=0,\ldots,m;\ \ i=2,\ldots,n)$$

Um einen Überblick über die Struktur des linearen Gleichungssystems (2.14) zu erhalten, führen wir Matrizenschreibweise ein.

Es bezeichne

$$A_k = \operatorname{sign}(q_k) \begin{pmatrix} q_k(h_1^{2k+1} + h_2^{2k-1}) & r_k h_2^{2k-1} & & & \\ r_k h_2^{2k-1} & q_k(h_2^{2k-1} + h_3^{2k-1}) & & & \\ & & \ddots & & r_k h_{n-2}^{2k-1} \\ & & & r_k h_{n-2}^{2k-1} & q_k(h_{n-2}^{2k-1} + h_{n-1}^{2k-1}) \end{pmatrix}$$

$$z_{k,i} = y_i^{(2k)} \qquad (k=0,\ldots,m; \quad i=1,\ldots,n)$$

$$z_k = (z_{k,2},\ldots, z_{k,n-1}) \qquad (k=0,\ldots,m)$$

(2.15)

$$b_o = -A_o z_o = \begin{pmatrix} \sum_{k=o}^{m} h_1^{2k-1} r_k z_{k,1} \\ 0 \\ 0 \\ \vdots \\ 0 \\ \sum_{k=o}^{m} h_{n-1}^{2k-1} r_k z_{k,n} \end{pmatrix}$$

$$b_i = -\begin{pmatrix} \sum_{k=0}^{m-i} h_1^{2k-1} r_k z_{k+i,1} \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \sum_{k=0}^{m-i} h_{n-1}^{2k-1} r_k z_{k+i,n} \end{pmatrix} \qquad (i=1,\ldots,m-1)$$

Dann hat (2.14) die Gestalt

$$(2.16)\quad \begin{pmatrix} A_1 & -A_2 & A_3 & -A_4 & A_5 & \ldots & (-1)^{m+1} A_m \\ A_o & A_1 & -A_2 & A_3 & A_4 & \ldots & \cdot \\ & A_o & A_1 & -A_2 & A_3 & \ldots & \cdot \\ & & A_o & A_1 & -A_2 & \ldots & \cdot \\ & & & A_o & A_1 & \ldots & \cdot \\ & & & & & & \\ & & & & & A_1 & -A_2 \\ & & & & & A_o & A_1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ \cdot \\ \cdot \\ z_{m-1} \\ z_m \end{pmatrix} = \begin{pmatrix} b_o \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ \cdot \\ \cdot \\ b_{m-2} \\ b_{m-1} \end{pmatrix}$$

Die Matrizen A_k $(k=0,\ldots,m)$ sind nach (2.15) alle symmetrisch und tridiagonal und besitzen die Kantenlänge n-2. Für $k=1,\ldots,m$ sind sie wegen $|q_k| > |r_k|$ und $x_1 < \ldots < x_n$ streng diagonal dominant und daher positiv definit. Wegen $|q_o| = |r_o| = 1$ ist A_o zwar nicht streng diagonal dominant, aber wegen $x_1 < \ldots < x_n$ irreduzibel diagonal dominant und folglich ebenfalls positiv definit (VARGA (1962)). Die Koeffizientenmatrix in (2.16) hat obere Block-Hessenberg-Gestalt und die Blockkantenlänge m. In jeder Blockdiagonale sind alle Elemente gleich.

Für den Fall äquidistanter Abszissen besitzt (2.16) noch weitere angenehme Eigenschaften. Setzen wir ohne Beschränkung der Allgemeinheit $h_i = 1$ $(i=1,\ldots,n-1)$ voraus, so erhalten wir mit der Bezeichnung

$$C(a) = \begin{pmatrix} a & 1 & & & \\ 1 & a & 1 & & \\ & 1 & a & \ddots & \\ & & \ddots & a & 1 \\ & & & 1 & a \end{pmatrix} \tag{2.17}$$

aus A_k in (2.15)

$$A_k = \operatorname{sign} q_k \,(2q_k E + r_k C(0)) \tag{2.18}$$

Dabei bezeichnet E die Einheitsmatrix. Daraus folgt:

(2.19) Satz Für äquidistante Abszissen sind die positiv definiten Matrizen A_k $(k=0,\ldots,m)$ paarweise vertauschbar. Es existieren die positiv definiten Inversen und Wurzeln und diese sind untereinander und mit den A_k wiederum vertauschbar.

Der erste Teil der Behauptung folgt aus (2.18) durch Nachrechnen; der zweite Teil aus der Tatsache, daß Inverse und Wurzeln Polynome in A_k sind.

Aus (2.19) folgt sofort, daß die A_k simultan auf Diagonalgestalt transformierbar sind. Da wir später die Eigenwerte und die Transformation benötigen, geben wir diese explizit an:
Die Matrix C(a) aus (2.17) wird diagonalisiert durch

$$U = (u_{ik}) = \left(\sqrt{\frac{2}{n-1}} \sin \left(\frac{ik\pi}{n-1} \right) \right) \tag{2.20}$$

wobei die Kantenlänge von C(a) n-2 ist.

C(a) besitzt die Eigenwerte $a + 2\cos\frac{\pi k}{n-1}$ $(k=1,\ldots,n-2)$.

Also besitzt A_k nach (2.18) die Eigenwerte

$$(2.21) \qquad d_{k,i} = \operatorname{sign} q_k \left(2q_k + 2r_k \cos\left(\frac{\pi i}{n-1}\right)\right) \qquad (i=1,\ldots,n-2)$$

Bemerkung: Transformiert man in (2.16) alle A_k simultan mit U, so stehen anstelle der tridiagonalen A_k Diagonalmatrizen, und das Gleichungssystem der Kantenlänge m(n-2) zerfällt in n-2 Gleichungssysteme der (kleinen) Kantenlänge m. Dieses hiermit angedeutete direkte Verfahren im Fall äquidistanter Abszissen wird man in praxi nicht anwenden, da es, wie schon die Transformation (2.20) zeigt, hohen Aufwand kostet.

Wir betrachten eine weitere, sich zwanglos ergebende Form des Gleichungssystems (2.14). Wenn wir in (2.16) den Zeilenindex der Blöcke und den Index der Blockzeilen vertauschen und die Unbekannten entsprechend umordnen, wird aus der oberen Block-Hessenberg-Matrix, deren Elemente tridiagonale Matrizen sind, eine blocktridiagonale Matrix, deren Elemente gewöhnliche obere Hessenberg-Matrizen sind.

Für äquidistante Abszissen ($h_i = 1$) erhält man so für die umgeordnete Koeffizientenmatrix

(2.22)

$$\begin{pmatrix} A & B & & & \\ B & A & B & & \\ & B & A & \ddots & \\ & & \ddots & \ddots & B \\ & & & B & A \end{pmatrix}$$

mit

$$A = 2\begin{pmatrix} q_1 & q_2 & q_3 & \cdots & q_m \\ q_o & q_1 & q_2 & \cdots & \\ & q_o & q_1 & \cdots & \\ & & & \ddots & q_2 \\ & & & q_o & q_1 \end{pmatrix} \qquad B = \begin{pmatrix} r_1 & r_2 & r_3 & \cdots & r_m \\ r_o & r_1 & r_2 & \cdots & \\ & r_o & r_1 & \cdots & \\ & & & \ddots & r_2 \\ & & & r_o & r_1 \end{pmatrix} \tag{2.23}$$

Die Matrix (2.22) ist symmetrisch und besitzt in den Blockhaupt- und nebendiagonalen jeweils lauter gleiche Elemente. Später wird noch wichtig sein, daß $A^{-1}B$ nur positive Eigenwerte zwischen 0 und 1/2 besitzt. Diese beiden Eigenschaften gelten für nichtäquidistante im allgemeinen nicht.

3. Erweiterung des Gleichungssystems für m = 2 auf die Typen I und II. Mögliche Modifikationen für m > 2 und Typ K

==========

Für m = 2 betrachten wir jetzt das Gleichungssystem (2.16). Es lautet mit $y'' = z_1$ und $y^{iv} = z_2$

$$(3.1) \qquad \begin{pmatrix} A_1 & -A_2 \\ A_0 & A_1 \end{pmatrix} \begin{pmatrix} y'' \\ y^{iv} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

In Komponentenschreibweise wird daraus

$$\begin{aligned} &\tfrac{1}{6} h_{k-1} y''_{k-1} + \tfrac{1}{3}(h_{k-1} + h_k) y''_k + \tfrac{1}{6} h_k y''_{k+1} \\ &-\left(\tfrac{7}{360} h_{k-1}^3 y^{iv}_{k-1} + \tfrac{8}{360}(h_{k-1}^3 + h_k^3) y^{iv}_k + \tfrac{7}{360} h_k^3 y^{iv}_{k+1} \right) \\ (3.2) \qquad &= \frac{\Delta y_k}{h_k} - \frac{\Delta y_{k-1}}{h_{k-1}} \qquad (k=2,\ldots,n-1) \\ &-\frac{1}{h_{k-1}} y''_{k-1} + \left(\frac{1}{h_{k-1}} + \frac{1}{h_k} \right) y''_k - \frac{1}{h_k} y''_{k+1} \\ &+ \tfrac{1}{6} h_{k-1} y^{iv}_{k-1} + \tfrac{1}{3}(h_{k-1} + h_k) y^{iv}_k + \tfrac{1}{6} h_k y^{iv}_{k+1} = 0 \end{aligned}$$

Stellen wir wie in (2.1) die Spline-Funktion für $x \in [x_k, x_{k+1}]$ $(k=1,\ldots,n-1)$ dar durch

$$(3.3) \qquad P_k(t) = A_k u^5 + B_k u^4 + C_k u^3 + D_k u^2 + E_k u + F_k$$

$$\text{mit } u = t - x_k$$

so erhalten wir die Koeffizienten in (3.3) zu

$$
\begin{aligned}
A_k &= \frac{1}{120h_k}\,(y^{iv}_{k+1} - y^{iv}_k)\\
B_k &= \frac{1}{24}\,y^{iv}_k\\
C_k &= \frac{1}{6h_k}\,(y''_{k+1} - y''_k) - \frac{1}{36}\,h_k\,(y^{iv}_{k+1} + 2y^{iv}_k)\\
D_k &= \frac{1}{2}\,y''_k\\
E_k &= \frac{\Delta y_k}{h_k} - \frac{1}{6}\,h_k\,(y''_{k+1} + 2y''_k) + \frac{h_k^3}{360}\,(8y^{iv}_k + 7y^{iv}_{k+1})\\
F_k &= y_k
\end{aligned}
\qquad (3.4)\qquad (k=1,\ldots,n-1)
$$

Wir wollen nun das Gleichungssystem (3.2) so modifizieren, daß es auch zur Berechnung von Splines vom Typ I und II verwendbar ist. Dazu genügt es, für die Intervalle $[x_1,x_2]$ und $[x_{n-1},x_n]$ Polynome der Gestalt (3.3) anzugeben, die die Interpolationsbedingungen erfüllen, für x_2 Werte y''_2 und y^{iv}_2 und für x_{n-1} Werte y''_{n-1} und y^{iv}_{n-1} sowie für Typ I in x_1 Werte y'_1 und y''_1 und in x_n Werte y'_n und y''_n und für Typ II in x_1 Werte y''_1 und y^{iv}_1 und in x_n Werte y''_n und y^{iv}_n annehmen.

Es ist möglich, solche Polynome anzugeben. Wir schreiben nur diejenigen Koeffizienten auf, die sich gegenüber (3.4) ändern. Für Typ I erhalten wir so

$$
\begin{aligned}
A_1 &= \frac{1}{8h_1^2}\left[\frac{3}{h_1^2}\,(\frac{\Delta y_1}{h_1} - y'_1) - \frac{1}{2h_1}\,(2y''_1 + y''_2) + \frac{h_1}{8}\,y^{iv}_2\right]\\
B_1 &= -\frac{1}{8h_1}\left[\frac{15}{2}\,(\frac{\Delta y_1}{h_1} - y'_1) - \frac{5}{2h_1}\,(2y''_1 + y''_2) + \frac{7h_1}{24}\,y^{iv}_2\right]\\
C_1 &= \frac{1}{4}\left[\frac{10}{2}\,(\frac{\Delta y_1}{h_1} - y'_1) - \frac{1}{h_1}\,(4y''_1 + y''_2) + \frac{h_1}{12}\,y^{iv}_2\right]\\
A_{n-1} &= \frac{1}{8h_{n-1}^2}\left[\frac{3}{h_{n-1}^2}\,(\frac{\Delta y_{n-1}}{h_{n-1}} - y'_n) + \frac{1}{2h_{n-1}}\,(2y''_n + y''_{n-1}) - \frac{h_{n-1}}{8}\,y^{iv}_{n-1}\right]\\
C_{n-1} &= -\frac{1}{4h_{n-1}^2}\left[\frac{5}{h_{n-1}^2}\,(\frac{\Delta y_{n-1}}{h_{n-1}} - y'_n) + \frac{1}{h_{n-1}}\,(\frac{3}{2}\,y''_{n-1} + y''_n) + \frac{h_{n-1}}{8}\,y^{iv}_{n-1}\right]\\
E_{n-1} &= \frac{1}{8}\left[15\,\frac{\Delta y_{n-1}}{h_{n-1}} - 7y'_n + h_{n-1}\,(y''_n - \frac{3}{2}\,y''_{n-1}) + \frac{h_{n-1}^3}{24}\,y^{iv}_{n-1}\right]
\end{aligned}
\qquad (3.5)
$$

Aus den Übergangsbedingungen in x_2 und x_{n-1} erhält man dann die notwendigen Modifikationen von (3.2) für Typ I zu

$$
\begin{aligned}
&-\tfrac{1}{8}h_1y_1'' + (\tfrac{3}{16}h_1 + \tfrac{1}{3}h_2)y_2'' + \tfrac{1}{6}h_2y_3'' \\
&-(\tfrac{7}{8}y_1' + (\tfrac{1}{192}h_1^3 + \tfrac{8}{360}h_2^3)y_2^{iv} + \tfrac{7}{360}h_2^3y_3^{iv}) \\
&\qquad = -\tfrac{15}{8}\frac{\Delta y_1}{h_1} + \frac{\Delta y_2}{h_2} \qquad (3.6)\\
&\tfrac{3}{2h_1}y_1'' + (\tfrac{9}{4h_1} + \tfrac{1}{h_2})y_2'' - \tfrac{1}{h_2}y_3'' \\
&-\tfrac{15}{2h_1^2}y_1' + (\tfrac{3}{16}h_1 + \tfrac{1}{3}h_2)y_2^{iv} + \tfrac{1}{6}h_2y_3^{iv} = \tfrac{15}{2h_1^3}\Delta y_1
\end{aligned}
$$

und

$$
\begin{aligned}
&\tfrac{1}{6}h_{n-2}y_{n-2}'' + (\tfrac{1}{3}h_{n-2} + \tfrac{3}{16}h_{n-1})y_{n-1}'' - \tfrac{1}{8}h_{n-1}y_n'' \\
&-(\tfrac{7}{360}h_{n-2}^3y_{n-2}^{iv} + (\tfrac{8}{360}h_{n-2}^3 + \tfrac{1}{192}h_{n-1}^3)y_{n-1}^{iv} - \tfrac{7}{8}y_n') \\
&\qquad = \tfrac{15}{8}\frac{\Delta y_{n-1}}{h_{n-1}} - \frac{\Delta y_{n-2}}{h_{n-2}} \\
&-\frac{1}{h_{n-2}}y_{n-2}'' + (\frac{1}{h_{n-2}} + \frac{9}{4h_{n-1}})y_{n-1}'' + \frac{3}{2h_{n-1}}y_n'' \\
&+\tfrac{1}{6}h_{n-2}y_{n-2}^{iv} + (\tfrac{1}{3}h_{n-2} + \tfrac{3}{16}h_{n-1})y_{n-1}^{iv} - \frac{15}{2h_{n-1}^2}y_n' = -\frac{15}{2h_{n-1}^3}\Delta y_{n-1}
\end{aligned}
$$

Auf dieselbe Weise erhalten wir für Typ II unter Berücksichtigung von $y_1''' = y_1^{iv} = y_n''' = y_n^{iv} = 0$

$$
\begin{aligned}
C_1 &= 0 \\
D_1 &= \tfrac{1}{2}(y_2'' - \tfrac{1}{6}h_1^2y_2^{iv}) \\
E_1 &= \frac{\Delta y_1}{h_1} - \tfrac{1}{2}h_1y_2'' + \tfrac{3}{40}h_1^3y_2^{iv} \\
C_{n-1} &= -\tfrac{1}{12}h_{n-1}y_{n-1}^{iv} \qquad (3.7)\\
E_{n-1} &= \frac{\Delta y_{n-1}}{h_{n-1}} - \tfrac{1}{2}h_{n-1}y_{n-1}'' + \tfrac{1}{20}h_{n-1}^3y_{n-1}^{iv}
\end{aligned}
$$

Die zugehörigen Änderungen für (3.2) lauten

$$
\begin{aligned}
&\left(\tfrac{1}{2}h_1 + \tfrac{1}{3}h_2\right)y_2'' + \tfrac{1}{6}h_2 y_3'' \\
&-\left(\tfrac{1}{20}h_1^3 + \tfrac{8}{360}h_2^3\right)y_2^{iv} - \tfrac{7}{360}h_2^3 y_3^{iv} = \frac{\Delta y_2}{h_2} - \frac{\Delta y_1}{h_1} \\
(3.8)\qquad &\frac{1}{h_2}y_2'' - \frac{1}{h_2}y_3'' + \left(\tfrac{1}{2}h_1 + \tfrac{1}{3}h_2\right)y_2^{iv} + \tfrac{1}{6}h_2 y_3^{iv} = 0 \\
&\left(\tfrac{1}{3}h_{n-2} + \tfrac{1}{2}h_{n-1}\right)y_{n-1}'' + \tfrac{1}{6}h_{n-1}y_{n-2}'' \\
&-\left(\tfrac{8}{360}h_{n-2}^3 + \tfrac{1}{20}h_{n-1}^3\right)y_{n-1}^{iv} - \tfrac{7}{360}h_{n-2}^3 y_{n-2}^{iv} = \frac{\Delta y_{n-1}}{h_{n-1}} - \frac{\Delta y_{n-2}}{h_{n-2}} \\
&\frac{1}{h_{n-2}}y_{n-1}'' - \frac{1}{h_{n-2}}y_{n-2}'' + \left(\tfrac{1}{3}h_{n-2} + \tfrac{1}{2}h_{n-1}\right)y_{n-1}^{iv} + \tfrac{1}{6}h_{n-2}y_{n-2}^{iv} = 0
\end{aligned}
$$

Wir haben die Modifikationen des Gleichungssystems (3.2) für Typ I und II im einzelnen aufgeschrieben, um die folgenden Eigenschaften ablesen zu können. Die Tridiagonalität der einzelnen Blöcke in (3.1) bleibt erhalten. Bei Typ I bleiben die Matrizen A_k (k=0,1,2) positiv definit; bei Typ II wird A_0 positiv semi-definit. Bei beiden Typen wird die Matrix A_1 in (3.1) links oben und rechts unten jeweils auf die gleiche Art geändert. Dies alles wird in Kapitel 4 bei der Bestimmung optimaler Relaxationsparameter wesentlich sein.

Es stellt sich die Frage, ob das Gleichungssystem (2.16) auch für m > 2 so modifiziert werden kann, daß ebenfalls Typ I und II - oder auch Spline-Funktionen von einem beliebigen Typ K - damit berechenbar sind, ohne daß die Blockstruktur zerstört wird. Eine ähnliche Fragestellung ist, ob sich bei Typ K ebenfalls die in den Randbedingungen auftretenden Ableitungen als Unbekannte einführen lassen, so daß die bei Typ III vorliegenden Vorteile von (2.14) erhalten bleiben.

Beide Fragen reduzieren sich auf die Lösbarkeit des folgenden Interpolationsproblems: Für welche Indexmengen

$$
\begin{aligned}
I &= \{i_0 = 0, i_1, \ldots, i_p\} \qquad (i_p \le 2m+1) \\
J &= \{j_0 = 0, j_1, \ldots, j_q\} \qquad (j_q \le 2m+1)
\end{aligned}
$$

und beliebige p+1 Zahlen $y_a^{i_k}$ und q+1 Zahlen $y_b^{j_k}$ mit p+q = 2m existiert jeweils ein eindeutig bestimmtes Polynom f vom Grad 2m+1 mit

$$
\begin{aligned}
&f^{(i_k)}(a) = y_a^{i_k} && i_k \in I && (k=0,\ldots,p)\\
&f^{(i_k)}(b) = y_b^{j_k} && j_k \in J && (k=0,\ldots,q)
\end{aligned}
\tag{3.10}
$$

Darauf gibt Antwort

(3.11) Satz (WHITTAKER (1934))

Es bezeichne P(k) die Anzahl der $i \in I$, die kleiner als k sind und Q(k) die Anzahl der $j \in J$, die kleiner als k sind. Dann ist das Interpolationsproblem (3.10) genau dann jeweils eindeutig lösbar wenn

$$P(k) + Q(k) \geq k \qquad (k=1,\ldots,2m+2) \tag{3.12}$$

Aus (3.11) ziehen wir nun einige Schlüsse:

i) Für m > 2 ist es nicht möglich, durch Modifikation der Matrizen A_k in nur der ersten und letzten Zeile Gleichungssysteme (wie bei m = 2) für die Typen I und II aus (2.16) zu erhalten.
Beispiel: für m = 3 und Typ II ist I = {0,4,5,6} und J = {0,2,4,6}.
Bedingung (3.12) ist für k = 4 verletzt.

ii) Durch Modifikation der Matrizen A_k in mehreren (die Anzahl hängt von m ab) der ersten und letzten Zeilen ist es möglich, aus (2.16) modifizierte Gleichungssysteme für die Typen I und II, j_a für beliebige Typen K, zu erhalten.
Beispiel: für m = 3 und Typ II kann man für x_2 I = (0,2,3,6) wählen. Dann sind die Interpolationsprobleme für $[x_1,x_2]$ und $[x_2, x_3]$ nach (3.12) lösbar. Dabei geht jedoch in allen Fällen die für äquidistante Abszissen geltende Vertauschbarkeit der A_k verloren, wie man an einfachen Beispielen sehen kann. Da diese Eigenschaft bei der Gewinnung optimaler Relaxationsparameter eine gewisse Rolle spielt, betrachten wir die Art, wie diese Modifikationen allgemein durchzuführen sind, nicht weiter.

iii) Aufgrund von (3.11) kann man entscheiden, für welche Typen K Unbekannte analog zu Typ III einführbar sind. Man sieht z.B., daß (3.12) für Typ I und beliebige m erfüllt ist. Andererseits gilt (3.12) nicht für Typ II und $m \geq 2$ wenn $k = 3$.
Bei unserem Konstruktionsprinzip für Typ III war es wesentlich, daß man für beliebige m die das Interpolationsproblem (3.10) lösenden Polynome explizit angeben konnte (siehe (2.3)). Dies gelingt anscheinend bisher nur noch für Typ I (AHLBERG, NILSON, WALSH (1967)). Das auf diese Weise für Typ I entstehende Gleichungssystem hat gegenüber (2.16) einige große Nachteile:
Das interpolierende Polynom (3.10) ist vom vorgegebenen m abhängig. Man hat keine zu (2.4) analoge von m unabhängige Darstellung. Dies bewirkt, daß nicht wie in (2.16) lauter gleiche Blockelemente in jeder Blockdiagonalen stehen. Weiter hat man nicht wie in (2.16) obere Block-Hessenberg-Form, sondern eine voll besetzte Blockmatrix, deren Elemente zwar wieder alle tridiagonal sind, aber nicht wie in (2.16) alle auch positiv definit, sondern abwechselnd symmetrisch (und positiv definit) und schiefsymmetrisch. Ihre Anordnung entspricht der der schwarzen und weißen Felder auf einem Schachbrett. Für äquidistante Abszissen sind nur die symmetrischen bzw. schiefsymmetrischen jeweils untereinander vertauschbar. Schließlich ist es anscheinend nicht möglich, einfache Formeln zur Berechnung der Koeffizienten $A_k^{(i)}$ in (2.1) aus den Unbekannten $y_i^{(k)}$ anzugeben, wie es uns bei Typ III möglich war.

4. Lösung des Gleichungssystems mit Blockrelaxationsverfahren. Optimale Beschleunigungsparameter

Es sei ein lineares Gleichungssystem $Ax = b$ mit $\det A \neq 0$ gegeben und für A eine Blockung derart, daß die Diagonalelemente quadratisch und invertierbar sind. Es bezeichne D die Blockdiagonalmatrix, -U die obere und -L die untere Blockdreiecksmatrix. Dann heißt

$$J = E - D^{-1}A = D^{-1}(L + U) \tag{4.1}$$

die Jacobi-Matrix und

$$L_\omega = (D - \omega L)^{-1} ((1-\omega) D + \omega U) \tag{4.2}$$

$$\text{mit } 0 < \omega < 2$$

die Blockrelaxationsmatrix von A bezüglich der gegebenen Blockung. Mit beiden Matrizen kann man Iterationsverfahren zur Lösung des gegebenen Systems durchführen. Z.B. lautet das zu (4.2) gehörige Verfahren

$$x^{(k+1)} = L_\omega x^{(k)} + (D-\omega L)^{-1} b$$

wobei k den Index der Iteration bedeutet. Diese sind genau dann konvergent wenn die Spektralradien $\rho(J)$ bzw. $\rho(L_\omega)$ kleiner als 1 sind. Nach einem Satz von KAHAN (VARGA (1962)) gilt $\rho(L_\omega) \geq |\omega-1|$, so daß reelle ω nur für $0 < \omega < 2$ sinnvoll sind. Für $\omega = 1$ spricht man vom Block-Gauss-Seidel-, für $0 < \omega < 1$ vom Blockunterrelaxations- und für $1 < \omega < 2$ vom Blocküberrelaxationsverfahren. ω heißt der Relaxations- oder Beschleunigungsparameter: Das Ziel ist es, ω so zu bestimmen daß $\rho(L_\omega)$ minimal (und < 1) wird.

Dies ist im wesentlichen bisher nur möglich, falls A bei der gegebenen Blockung eine p-zyklische, konsistent geordnete Matrix ist und über die Eigenwerte der Jacobi-Matrix gewisse Voraussetzungen erfüllt sind (VARGA (1962)). Die Matrix in (2.16) ist bei der natürlichen Blockung nicht p-zyklisch.

Da wir im folgenden nur einen Spezialfall der 2-zyklischen Matrizen - nämlich blocktridiagonale -, die schon konsistent geordnet sind, benötigen, formulieren wir die allgemeinen Ergebnisse nur für diesen Fall.

(4.3) Satz

A sei blocktridiagonal mit quadratischen, invertierbaren Diagonalelementen.
Dann gilt

i) VARGA (1962): Sei $\lambda \neq 0$ Eigenwert von L_ω und erfüllt μ die Beziehung

(4.4) $$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$$

so ist μ Eigenwert von J. Ist μ Eigenwert von J und erfüllt λ (4.4), so ist λ Eigenwert von L_ω.

ii) VARGA (1962): Die Eigenwerte von J^2 seien reell und nichtnegativ und es gelte $0 \leq \rho(J) < 1$. Dann gilt $\rho(L_\omega) < 1$ für $0 < \omega < 2$. $\rho(L_\omega)$ wird minimal für

(4.5) $$\omega_b = \frac{2}{1 + \sqrt{1 - \rho(J^2)}}$$

und es gilt

$$\rho(L_\omega) = \omega - 1 \qquad \text{für } \omega \geq \omega_b \tag{4.6}$$

iii) NIETHAMMER (1964): Die Eigenwerte von J^2 seien reell und nicht-positiv.
Dann gilt $\rho(L_\omega) < 1 \quad$ für $0 < \omega < \frac{2}{1+\rho(J)}$.
$\rho(L_\omega)$ wird minimalisiert durch

$$\omega_b = \frac{2}{1+\sqrt{1+\rho(J^2)}} \tag{4.7}$$

und es gilt

$$\rho(L_\omega) = 1 - \omega \qquad \text{für } \omega \leq \omega_b \tag{4.8}$$

Bevor wir Satz (4.3) anwenden, betrachten wir die explizite Form des Blockrelaxationsverfahrens bei der Anwendung auf unser Gleichungssystem (2.14) mit den Blockungen (2.16) und (2.22).

Für (2.16) lautet das Verfahren

$$z_i^{(k+1)} = (1-\omega) z_i^{(k)} + \omega A_1^{-1} \left(b_{i-1} - A_o z_{i-1}^{(k+1)} + \sum_{j=2}^{m-i+1} (-1)^j A_j z_{j+i-1}^{(k)} \right)$$

$(i=1,\ldots,m), \quad z_o := 0$

Bemerkung: Man könnte in diesem Fall natürlich die Rollen von L und U vertauschen und ein weiteres Verfahren erhalten. Dieses konvergiert jedoch nicht annähernd so gut.

Bezeichnen die mit einem Querstrich versehenen Größen die Vektoren z und b nach der in (2.22) vorgenommenen Umordnung, so lautet hierfür das Blockrelaxationsverfahren bei äquidistanten Abszissen

$$\overline{z}_i^{(k+1)} = (1-\omega)\overline{z}_i^{(k)} + \omega A^{-1}(\overline{b}_{i-1} - B(\overline{z}_{i-1}^{(k+1)} + \overline{z}_{i+1}^{(k)}))$$

$$(i=1,\ldots,n-2) \qquad \overline{z}_o = \overline{z}_{n-1} = 0$$

In beiden Fällen erweist sich die Eigenschaft, daß in der Blockhauptdiagonalen jeweils lauter identische Matrizen der Kantenlänge n-2 bzw. m stehen, als wesentlich vereinfachend für die Durchführung der Iterationen. A_1 besitzt die (große) Kantenlänge n-2, ist aber nur tridiagonal. Wegen der positiven Definitheit kann eine Dreieckszerlegung ohne Pivotisierung stattfinden, und zwar vor Beginn der Iteration. Die Matrix A besitzt die(kleine) Kantenlänge m und kann ein für allemal invertiert werden.

Nun wollen wir nach Satz (4.3) optimale Beschleunigungsparameter bestimmen. Dazu betrachten wir zunächst (3.1), also den Fall m = 2. Die Koeffizientenmatrix hatte die Gestalt

$$(4.9) \qquad \begin{pmatrix} A_1 & -A_2 \\ A_o & A_1 \end{pmatrix}$$

und es waren dadurch Spline-Funktionen vom Grad 5 und der Typen I, II und III bestimmt. Wir hatten die Eigenschaft, daß für beliebige Abszissen A_1 und A_2 positiv definit waren und A_o positiv semi-definit war. Unter diesen Voraussetzungen gilt:

(4.10) Satz:

Die Eigenwerte der zu (4.9) gehörenden Jacobi-Matrix

$$(4.11) \qquad \begin{pmatrix} 0 & A_1^{-1}A_2 \\ -A_1^{-1}A_0 & 0 \end{pmatrix}$$

liegen auf der imaginären Achse.

Beweis: Die Eigenwerte t von (4.11) genügen der Gleichung

$$\det(t^2E + A_1^{-1}A_2\, A_1^{-1}A_0) = 0$$

Aufgrund der Voraussetzungen ist $B = A_1^{-1}A_2A_1^{-1}$ positiv definit. Dann existiert $B^{1/2}$ und die zu BA_0 ähnliche Matrix $B^{1/2}A_0B^{1/2}$ ist positiv semi-definit. Daraus folgt die Behauptung.

Aus (4.10) folgt zusammen mit (4.3) iii), daß der optimale Relaxationsparameter für (3.1) gegeben ist durch

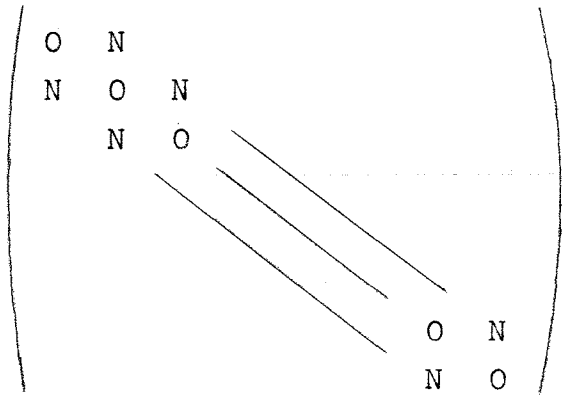
$$(4.12) \qquad \omega_b = \frac{2}{1+\sqrt{1+\rho(A_1^{-1}A_2A_1^{-1}A_0)}}$$

Da wir wissen, daß $A_1^{-1}A_2A_1^{-1}A_0$ nur nichtnegative Eigenwerte besitzt, ist die numerische Bestimmung von $\rho(A_1^{-1}A_2A_1^{-1}A_0)$ mit Hilfe der Potenz-Methode (VARGA (1962)) unproblematisch.

Wir betrachten nun für $m > 2$ die Blockungen (2.16) und (2.22) des Gleichungssystems (2.14), setzen nun aber auch für (2.16) äquidistante Abszissen voraus, so daß die Matrizen A_k nach Satz (2.19) vertauschbar sind.

Zunächst benötigen wir einen Hilfsatz über Eigenwerte von Matrizen der Gestalt

(4.13) $M =$



(4.14) Hilfssatz: Die Matrix N in (4.13) habe die Eigenwerte ν_k $(k=1,\ldots,r)$. M habe die Blockkantenlänge s. Dann hat M die Eigenwerte

$$\mu_{ki} = 2\,\nu_k \cos\frac{i\pi}{s+1} \tag{4.15}$$

$$(i=1,\ldots,s; \quad k=1,\ldots,r)$$

Dies ergibt sich mit Hilfe des Additionstheorems für die Sinus-Funktion aus der durch die Bauweise von M bedingten Tatsache, daß man aus einem Eigenvektor x von N s Eigenvektoren $y_j = (y_{j1},\ldots,y_{js})$ von M erhält mit der Festsetzung $y_{jk} = \sin(\frac{jk\pi}{s+1})\,x \quad (j,k=1,\ldots,s)$.

Betrachten wir nun die Matrix (2.22), so sehen wir, daß die zugehörige Jacobi-Matrix von der Gestalt (4.13) ist mit $N = -A^{-1}B$, wobei A und B in (2.23) definiert sind. Es ist nicht gelungen, die Eigenwerte ν_k von $A^{-1}B$ für beliebige m explizit anzugeben. Daher haben wir sie für $2 \leq m \leq 10$ numerisch berechnet.

Es gilt (vgl. Tabelle 2 mit $\nu_{max} = \frac{1}{2}\rho(J)$):

(4.16) Für die Eigenwerte ν_k $(k=1,\ldots,m)$ von $A^{-1}B$ gilt $0 < \nu_k < \frac{1}{2}$ für $2 \leq m \leq 10$

Die Eigenwerte μ_{ik} der Jacobi-Matrix von (2.22) sind nach (4.14) gegeben durch

$$\mu_{ik} = -2\,\nu_k \cos\left(\frac{i\pi}{n-1}\right) \tag{4.17}$$

$$(k=1,\ldots,m;\quad i=1,\ldots,n-2)$$

Für die Eigenwerte (4.17) gilt also nach (4.16) $0 < \mu_{ik}^2 < 1$. Die Voraussetzungen von (4.3) ii) sind erfüllt und wir haben die

(4.18) Folgerung: Für das zu (2.22) gehörende Blocküberrelaxationsverfahren ist der beste Relaxationsparameter durch (4.5) gegeben mit

$$\rho(J) = 2\,\nu_{max} \cos\frac{\pi}{n-1}$$

wobei $\nu_{max} < \frac{1}{2}$ der größte Eigenwert von $A^{-1}B$ ist.

In Tabelle 2 sind die Werte bei $n = \infty$ für $\rho(J) = 2\,\nu_{max}$ und ω_b angegeben. Für $20 < n < \infty$ ergeben sich wegen obiger Formel nur unwesentlich kleinere Werte für ω_b.

Tabelle 2: Optimale Beschleunigungsparameter für das Blocküberrelaxationsverfahren bei $n = \infty$

m	$\rho(J)$	ω_b
2	.726	1.185
3	.832	1.287
4	.888	1.370
5	.920	1.437
6	.940	1.493
7	.954	1.539
8	.963	1.577
9	.970	1.610
10	.975	1.638

Bei der dritten Anwendung von Satz (4.3) kommen wir wieder auf die Blockung (2.16) zurück. Da (4.3) und die allgemeineren Sätze auf (2.16) nicht anwendbar sind, betrachten wir jetzt die blocktridiagonale Matrix

(4.19)

$$\begin{pmatrix} A_1 & -A_2 & & & \\ A_o & A_1 & -A_2 & & \\ & A_o & A_1 & \ddots & \\ & & \ddots & \ddots & -A_2 \\ & & & A_o & A_1 \end{pmatrix}$$

die aus (2.16) durch Abschneiden einer Dreiecksmatrix erhalten wird, in der nur die Matrizen A_k mit $k \geq 3$ vorkommen. Der Versuch, für diese reduzierte Matrix optimale Beschleunigungsparameter zu erhalten und daraus wiederum Schlüsse für (2.16) zu ziehen, wird legitimiert durch die Tatsache, daß nach (2.18) und Tabelle 1 gilt

$$A_k \sim 2.2_{10}\text{-}k\ C(2) \qquad (k \geq 3) \tag{4.20}$$

wobei C(2) nach (2.17) definiert ist. Für $k \geq 3$ sind die A_k klein gegenüber A_o, A_1 und A_2.

Betrachten wir also jetzt (4.19). Die zugehörige Jacobi-Matrix lautet

$$J = \begin{pmatrix} 0 & A_1^{-1}A_2 & & & \\ -A_1^{-1}A_0 & 0 & \ddots & & \\ & \ddots & \ddots & A_1^{-1}A_2 & \\ & & -A_1^{-1}A_0 & 0 \end{pmatrix} \tag{4.21}$$

Definieren wir

$$W := iA_o^{1/2}A_2^{-1/2} \qquad (i^2 = -1)$$

und die Blockdiagonalmatrix T durch

$$T = \begin{pmatrix} E & & & & \\ & W & & & \\ & & W^2 & & \\ & & & \ddots & \\ & & & & W^{m-1} \end{pmatrix}$$

so erhalten wir aufgrund der Vertauschbarkeitseigenschaften (2.19)

(4.22)
$$T^{-1}JT = i \begin{pmatrix} 0 & V & & & \\ V & 0 & & & \\ & & \ddots & & \\ & & & 0 & V \\ & & & V & 0 \end{pmatrix}$$

mit

(4.23)
$$V = A_1^{-1} A_0^{1/2} A_2^{1/2}$$

V besitzt wegen (2.19) und (2.21) die Eigenwerte

(4.27)
$$\lambda_k = \frac{\sqrt{(q_0 + r_0 \cos\frac{\pi k}{n-1})(|q_2| + |r_2| \cos\frac{\pi k}{n-1})}}{(q_1 + r_1 \cos\frac{\pi k}{n-1})}$$

Also besitzt J nach (4.22) und (4.14) die Eigenwerte

$$\mu_{kj} = 2i\lambda_k \cos\left(\frac{j\pi}{m+1}\right) \qquad (i^2 = -1)$$

$$(k=1,\ldots,n-2;\ j=1,\ldots,m)$$

Die Voraussetzung von (4.3) iii) ist also erfüllt. Um $\rho(J)$ zu finden, genügt es, da $2\cos\frac{j\pi}{m+1}$ für $j = 1$ maximal wird, $\max_k \lambda_k$ zu suchen. Setzen wir $\gamma = \gamma(k) = \cos\frac{\pi k}{n-1}$ und $c_k = q_k/r_k$ $(k=0,1,2)$ so ist also nach (4.24)

$$f(\gamma) = \frac{\sqrt{(c_0+\gamma)(c_2+\gamma)}}{c_1+\gamma}$$

so groß wie möglich zu machen. $f(\gamma)$ besitzt genau ein Maximum, nämlich für

$$\bar{\gamma} = \frac{c_1(c_0+c_2) - 2c_0c_2}{c_0+c_2-2c_1} = -\frac{2}{3}$$

Seien nun k_1 und k_2 diejenigen beiden aufeinanderfolgenden ganzen Zahlen mit $\gamma(k_1) \leq -\frac{2}{3} \leq \gamma(k_2)$. Es sei.

$$(4.25) \qquad k_o = \begin{cases} k_1 & f(\gamma(k_1)) \geq f(\gamma(k_2)) \\ k_2 & \text{sonst} \end{cases}$$

Dann ist der Spektralradius von J gegeben durch

$$(4.26) \qquad \rho(J) = 2\cos\frac{\pi}{m+1}\,\frac{\sqrt{(q_o+r_o\cos\frac{\pi k_o}{n-1})(|q_2|+|r_2|\cos\frac{\pi k_o}{n-1})}}{(q_1+r_1\cos\frac{\pi k_o}{n-1})}$$

und der optimalen Beschleunigungsparameter für die zu der Matrix (4.19) Blockrelaxationsmatrix ist bestimmt durch (4.7). Setzen wir statt $\cos\frac{\pi k_o}{n-1}$ die Größe $\bar{\gamma} = -\frac{2}{3}$ ein, was für größere n realistisch ist, so erhalten wir in Abhängigkeit von m die Zahlenwerte von $\rho(J)$ und ω_b in Tabelle 3. Für $m \to \infty$ geht ω_b gegen $\frac{4}{5}$ und $\rho(L_\omega)$ gegen $\frac{1}{5}$, was ein guter Wert ist. Für m = 2 haben wir den exakten Wert auch für (2.16), da (4.19) und (2.16) in diesem Fall identisch sind.

Wendet man nun das Blockunterrelaxationsverfahren für m > 2 auf (2.16) statt auf (4.19) an, so erwartet man intuitiv eine Konvergenzverschlechterung, da die Matrix des Gleichungssystems voller wird. Eine Verschlechterung würde aber bedeuten, da $\rho(L_\omega)$ größer wird, daß ω_b unter der Voraussetzung der näherungsweisen Richtigkeit von (4.9) kleiner wird.

Wir haben nun ω_b für einen festen Bereich, nämlich für $3 \leq m \leq 10$ und $10 \leq n \leq 1000$, empirisch bestimmt.

Wir hätten wie für den Fall m = 2 und nichtäquidistanter Abszissen, den Spektralradius $\rho(L_\omega)$ in Abhängigkeit von ω, m und n mit Hilfe des Potenzverfahrens bestimmen können. Da wir hier jedoch die Information, daß

alle Eigenwerte von L_ω nichtnegativ sind, nicht besitzen, haben wir von diesem Verfahren abgesehen, da man bei der Potenzmethode über Vielfachheit und Anzahl der betragsgleichen Eigenwerte, die zu $\rho(L_\omega)$ gehören, von vornherein eine Information haben muß, wenn man nicht jeweils all Möglichkeiten (FADDEJEW, FADDEJEWA (1964)) durchprobieren will.

Der einfachere Weg war der folgende. Für $m = 3,4,\ldots,10$ und $n = 10(10)100$, $150(50)350$, $400(100)1000$ wurde das Iterationsverfahren bei verschiedenen Startvektoren und Ordinaten durchgeführt für verschiedene Beschleunigungsparameter und mit dem Abbruchkriterium

$$\max_{i=1,\ldots,m} \left\{ \max_{k=2,\ldots,n-1} \left| \frac{z_{ik}^{(t+1)} - z_{ik}^{(t)}}{z_{ik}^{(t+1)}} \right| \right\} < .0005$$

t bezeichnet den Index der Iteration. Als Werte für ω_b wurden diejenigen genommen, für die die Anzahl der Iterationen gemittelt über die verschiedenen Startwerte minimal war. Diese Minima sind nicht sehr stark ausgeprägt. Da bei Unterrelaxation Unterschätzen des Relaxationsparameter günstiger ist als Überschätzen (NIETHAMMER (1964)), wurde in Zweifelsfällen stets das kleinere ω_b genommen. Eine merkliche Abhängigkeit von $\rho(L_\omega)$ von n ist nicht sichtbar. Die so gewonnenen Werte für $\bar{\omega}_b$ finden sich in Tabelle 3. Wie die Anzahlen der Iterationen zeigten, bleibt die Beziehung $\rho(L_{\bar{\omega}_b}) = 1 - \bar{\omega}_b$ ungefähr richtig.

Tabelle 3: "Beste" Beschleunigungsparameter für den abgeschnittenen und den nichtabgeschnittenen Fall

Spalte 2: Spektralradius der Jacobi-Matrix für den abgeschnittenen Fall

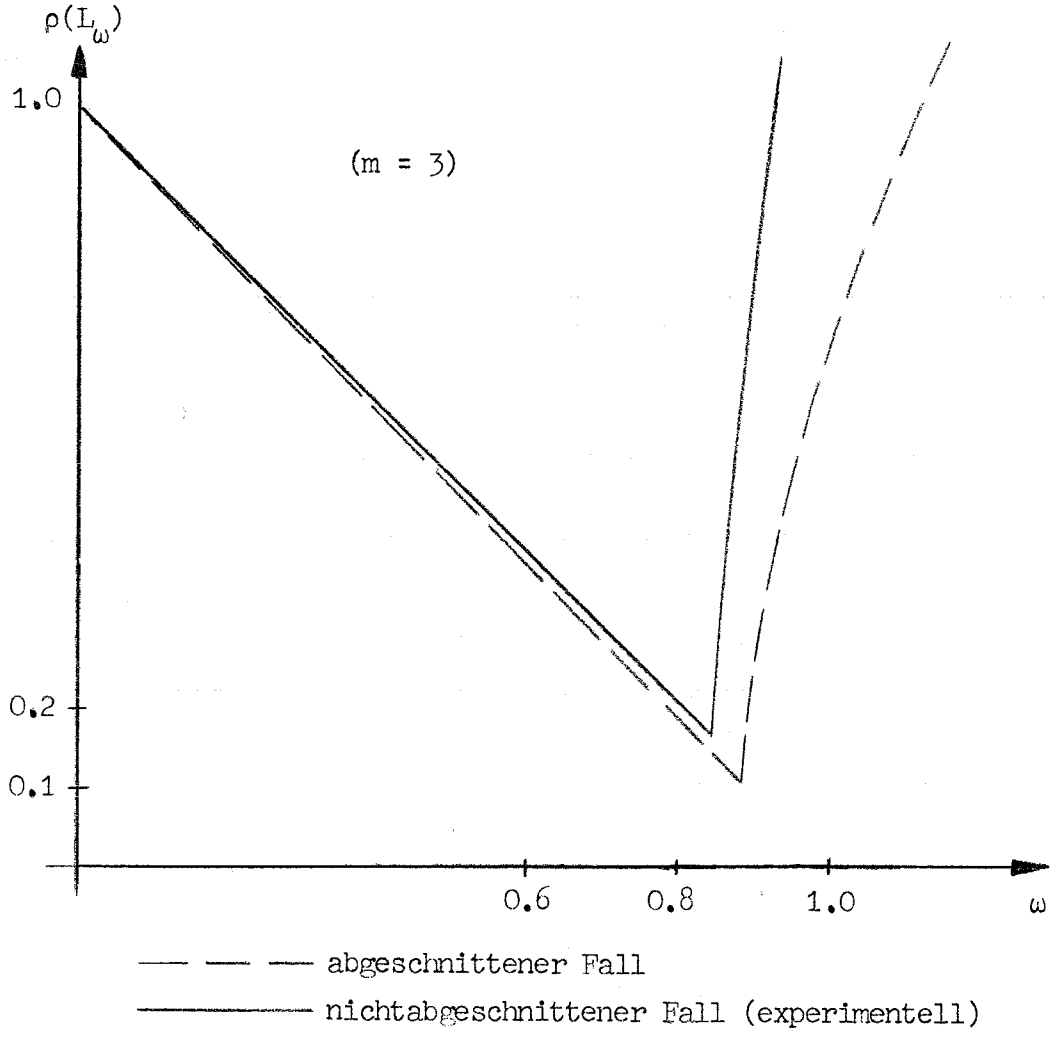
Spalte 3: Exaktes ω_b für den abgeschnittenen Fall

Spalte 4: $\bar{\omega}_b$ für den nichtabgeschnittenen Fall

m	$\rho(J)$	ω_b	$\bar{\omega}_b$
2	.559	.932	.932
3	.791	.879	.847
4	.905	.852	.808
5	.968	.836	.778
6	1.007	.827	.755
7	1.033	.820	.737
8	1.051	.816	.723
9	1.063	.813	.713
10	1.073	.811	.705
∞	$1.118 = \frac{\sqrt{5}}{2}$	$.800 = \frac{4}{5}$	

Vergleicht man die Kurven von $\rho(L_\omega)$ von (4.19) und (2.16) in Abhängigkeit von ω, so ergibt sich unabhängig von m folgendes Bild: Für $\omega \to 0$ gehen die beiden Kurven ineinander über. Die Kurve für (2.16) hat dieselbe Form wie die für (4.19). Jedoch liegt das Minimum $\bar{\omega}_b$ weiter links und für $\omega > \bar{\omega}_b$ besitzt die Kurve eine wesentlich größere Steigung. $\rho(L_\omega) = 1$ wird für kleinere ω erreicht. Das Schaubild illustriert die Verhältnisse für m = 3:

Bild 1



Rechnungen haben gezeigt, daß sich allgemeine empirische Aussagen für nichtäquidistante Abszissen nicht so einfach gewinnen lassen. Da die Matrixelemente und somit der Spektralradius stetig von den Abszissendistanzen abhängen, hat man für nahezu äquidistante Abszissen natürlich auch noch Konvergenz. Die Erfahrung lehrt, daß $\beta = \frac{\Delta x_{max}}{\Delta x_{min}}$ umso näher bei Eins liegen muß, je größer n und vor allem je größer m wird. Man kann Beispiele angeben, für die es kein ω mit $0 < \omega < 2$ gibt, so daß $\rho(L_\omega) < 1$ ist: $m=3$, $n=3$, $\Delta x_1=1$, $\Delta x_2=40$.

Wir fassen die Ergebnisse zusammen:

i) Für die Spline-Interpolation vom Grad 5($m = 2$) und der Typen I, II und III kann man bei beliebigen Abszissen Gleichungssysteme aufstellen, die es erlauben, bei Anwendung des Blockunterrelaxationsverfahrens eine vollständige Theorie zur Bestimmung des optimalen Beschleunigungsparameters aufzustellen.

ii) Für äquidistante Abszissen, $2 \leq m \leq 10$ und beliebige n und Typ III haben wir wiederum die exakten optimalen Relaxationsparameter (Blocküberrelaxation). Dabei ist ein Schönheitsfehler, daß die Eigenschaft der Jacobi-Matrix, nur reelle Eigenwerte kleiner als Eins zu besitzen, nur numerisch nachgewiesen ist. Daher die Einschränkung $2 \leq m \leq 10$.

iii) Für äquidistante Abszissen, $2 \leq m \leq 10$ und $10 \leq n \leq 1000$ und Typ III haben wir, gestützt auf exakte Ergebnisse bei einem verwandten Problem, empirisch optimale Relaxationsparameter ermittelt (Blockunterrelaxation). Bei Vergleich der Werte von $\rho(L_{\omega_b})$ bei Blocküber- und -unterrelaxation, stellt man fest (siehe Tabellen 2 und 3), daß die Unterrelaxation mehr als doppelt so gut wie die Überrelaxation ist.

5. Aufwand und Rundungsfehler im Vergleich zu bekannten Verfahren. Einfluß des Typs auf die Gestalt

Bisher haben wir ein Gleichungssystem für die Spline-Interpolation vom Typ III aufgestellt, dieses für andere Typen K modifiziert, ein iteratives Verfahren zu dessen Auflösung in bestimmten Fällen angegeben und, soweit wie möglich, optimale Beschleunigungsparameter bestimmt. Jetzt werden wir uns mit der Effektivität, d.h. dem Rechenaufwand, und der numerischen Stabilität, d.h. dem Einfluß von Rundungsfehlern, im Vergleich zu bereits bekannten Methoden beschäftigen. Abschließend geben wir konkrete Beispiele für den Einfluß der Randbedingungen an.

Zunächst schildern wir die für Typ II konzipierte, allgemein (CARASSO (1966), CARASSO, LAURENT (1968), SCHUMAKER (1969)) als die numerisch stabilste und am wenigsten aufwendige empfohlene Methode.

Diese geht von dem für eine Spline-Funktion vom Typ II und vom Grad 2m+1 mit den Knoten $x_1,\ldots,x_n$ gültigen Ansatz aus:

$$s(t) = P_m(t) + \sum_{j=1}^{n-m-1} \lambda_j \, h_m(j,t) \tag{5.1}$$

mit

$$P_m(t) = \sum_{i=0}^{m} p_i \, t^i \tag{5.2}$$

und

$$(5.3) \qquad h_m(j,t) = \delta_{x_j,\ldots,x_{j+m+1}} \left[(t-x)_+^{2m+1} \right],$$

wobei bedeutet

$$(t)_+ = \begin{cases} t & \text{für } t \geq 0 \\ 0 & \text{sonst} \end{cases}$$

und weiter

$$\delta_{x_j,\ldots,x_{j+m+1}} \left[f(x) \right] = \sum_{i=1}^{m+1} \frac{f(x_i)}{\prod_{j \neq i}(x_i - x_j)}$$

Diese dividierten Differenzen der Ordnung m+1 können rekursiv berechnet werden. Schreiben wir an den Knotenstellen die Ordinaten $y_1,\ldots,y_n$ vor, so verläuft der Algorithmus zur Bestimmung der Unbekannten $p_0,\ldots,p_m$ und $\lambda_1,\ldots,\lambda_{n-m-1}$ in folgenden zwei Schritten:

A. Bestimmung der $\lambda_1,\ldots,\lambda_{n-m-1}$:

1. Berechne die dividierten Differenzen

$$b_i = \delta_{x_i,\ldots,x_{i+m+1}} y := \sum_{j=1}^{m+1} \frac{y_i}{\prod_{i \neq j} (x_j - x_i)}$$

$(i=1,\ldots,n-m-1)$

und

$$a_{ij} = \delta_{x_i,\ldots,x_{i+m+1}} \left[h_m(j,t) \right]$$

$(i,j=1,\ldots,n-m-1)$

2. Löse das lineare Gleichungssystem

$$(5.4) \qquad \sum_{j=1}^{n-m-1} a_{ij}\,\lambda_j = b_i \qquad (i=1,\ldots,n-m-1)$$

Die Matrix (a_{ij}) ist symmetrisch; es gilt $a_{ij} = 0$ für $|i-j| > m+1$. Die Diagonalelemente sind dem Betrage nach stets größer als jedes einzelne Element in der zugehörigen Zeile. Es gilt $a_{ij} > 0$ wenn $n > 2m+1$.

B. Bestimmung der $p_0,\ldots,p_m$:

Aus den Interpolationsbedingungen $s(x_j) = y_j$ $(j=1,\ldots,m+1)$ erhält man das lineare Gleichungssystem mit VANDERMONDE'scher Matrix

$$(5.5) \qquad \sum_{i=0}^{m} x_j^i\, p_i = y_j - \sum_{i=1}^{n-m-1} \lambda_i\, h(i,x_j)$$

$$(j=1,\ldots,m+1)$$

Bei Schritt A werden bei der Berechnung der b_i dividierte Differenzen der Ordnung m+1 und bei den a_{ij} solche der Ordnung 2m+2 benötigt. Die dadurch bedingten Rundungsfehler können die nach (5.4) zu berechnenden λ_j verfälschen und diese wiederum die p_i nach (5.5). Sehr kritisch ist dann die Auswertung der Spline-Funktion nach (5.1) insbesondere für Werte t in der rechten Hälfte des Intervalls $[x_1, x_n]$, da dort viele λ_i und stark wachsende Zahlen $h_m(j,t)$ beteiligt sind und $P_m(t)$ für große t ausgewertet werden muß.

Um den Rundungsfehler zu verkleinern, werden folgende beiden Modifikationen von Schritt B vorgeschlagen (CARASSO (1966)), die die Auswertung nach (5.1) vermeiden:

C'. Die Zahlen $p_o,\ldots,p_m$ werden für jedes einzelne Intervall $[x_i,x_{i+1}]$ $(i=1,\ldots,n-m-1)$ neu analog zu (5.5) aus den Interpolationsbedingungen $s(x_{i+j}) = y_{i+j}$ $(j=0,\ldots,m)$ berechnet und daraus für jede einzelne Abszisse x_i die Zahlen $s^{(k)}(x_i)$ $(k=1,\ldots,m)$ durch Differentiation von (5.1) gewonnen.

C''. Durch (m+1)-malige Differentiation von (5.1) erhält man

$$(5.6) \qquad s^{(m+1)}(t) = \frac{(2m+1)!}{m!} \sum_{j=1}^{n-m-1} \lambda_j \, h_m^{(m+1)}(j,t)$$

und somit die Zahlen $y_i^{(m+1)} = s^{(m+1)}(x_i)$ allein aus den λ_j. Das weitere Vorgehen schildern wir exemplarisch für $m = 2$. Im Intervall $[x_{i-1},x_{i+2}]$ $(i=2,\ldots,n-2)$ kann $s(t)$ geschrieben werden als

$$(5.7) \qquad s(t) = \sum_{i=0}^{5} a_i \, t^i + \alpha(t - x_i)_+^5 + \beta(t - x_{i+1})_+^5$$

Die acht Zahlen $a_0,\ldots,a_5,\alpha,\beta$ können aus dem linearen Gleichungssystem

$$(5.8) \qquad s(x_k) = y_k, \quad s'''(x_k) = y_k''' \qquad (k= i-1,\ldots,i+2)$$

bestimmt werden. Durch Differentiation von (5.7) kann man wie bei C' die Zahlen $s^{(k)}(x_i)$ $(k=1,\ldots,m)$ gewinnen.

Nach Satz (3.11) hat man nun in beiden Fällen genügend Information, um die Spline-Funktion in der Form

$$s(t) = f_i(t-x_i) \quad \text{für} \quad t \in [x_i,x_{i+1}]$$

also stückweise durch Polynome f_i vom Grad 2m+1, was (2.1) entspricht, darstellen zu können. Letzten Endes hat man also die gleiche Anzahl von Unbe-

kannten, wie wir sie bei Typ III eingeführt hatten.

Bei Schritt C' müssen n-m-1 lineare Gleichungssysteme der Kantenlänge m+1, bei C" solche der Kantenlänge 4m gelöst werden. Die Methode der lokalen Integration C" soll in bezug auf Rundungsfehler günstiger sein (CARASSO (1966)), weil bei der Berechnung der Werte $s^{(k)}(x_i)$ (k=1,...,m) die nach (5.4) berechneten, mit Rundungsfehlern behafteten λ_i nur lokal in die Berechnung der Zahlen $s^{(m+1)}(x_i)$ eingehen ($h_m^{(m+1)}(j,t) = 0$ für $t \leq x_j$ und $t \geq x_{i+m+1}$), während bei C' in die rechten Seiten von (5.5) die λ_i für wachsende x in wachsender Anzahl auftreten und mit immer größer werdenden Zahlen $h_m(i,t)$ multipliziert werden. Andererseits besitzen die bei C" zu lösenden Gleichungssysteme (5.8) im allgemeinen eine schlechtere Kondition als die bei C' zu lösenden aus (5.5), was den oben genannten Vorteil aufheben kann.

Wir können hier nicht allgemein nachprüfen, ob nun Schritt C' oder C" numerisch günstiger ist. Die Erfahrung spricht für C".

An ALGOL-Programmen lag für variables m nur A - C' vor (CARASSO (1967)). Für m = 2 existierte ein ALGOL-Programm für A - C" (CARASSO (1966)), das aber noch Fehler enthielt und in das zur Lösung von (5.8) das GAUSSsche Eliminationsverfahren mit Pivotisierung (statt ohne) eingebaut wurde. Aufgrund von numerischen Erfahrungen wird empfohlen, die Programme nur für $2 \leq m \leq 4$ und $n \leq 100$ zu benutzen (CARASSO, LAURENT (1968)).

Im folgenden führen wir an Beispielen einen numerischen Vergleich des Blockunterrelaxationsverfahren (Abkürzung BU) mit den Methoden A - C' und A - C" durch, soweit dies möglich ist. Für die Rechnungen stand eine Rechenanlage

IBM 360/65 mit etwa 200 000 Multiplikationen bzw. 400 000 Additionen pro Sekunde zur Verfügung. Der ALGOL-Compiler kann per Compilereingabe veranlasst werden, wahlweise mit allen als reell deklarierten Größen einfach (etwa 7 Dezimalen) oder doppelt genau (16 Dezimalen) zu rechnen.

Zunächst betrachten wir den Fall m = 2 und nichtäquidistanter Abszissen. Auf das nach (3.8) modifizierte Gleichungssystem können wir BU mit optimalen Beschleunigungsparameter anwenden und genau wie A - C' bzw. A - C" interpolierende Splines vom Typ II berechnen. Da bei A - C' bzw. A - C" und BU unmittelbar nur die Zahlen y''_k (k=2,...,n-1) als gemeinsame Ergebnisse anfallen, ziehen wir diese zum Vergleich heran.

Im ersten Beispiel vergleichen wir die Werte y''_k beim Übergang von einfacher auf doppelte Genauigkeit:

(5.9) Beispiel 1

$$x_1 := -15$$
$$x_k := .01 \times \text{entier}\,(100 \times (x_{k-1} + .01 + |a \times \sin(b \times k)|)) \quad (k=2,\ldots,n)$$
$$y_k := f(x_k) \quad (k=1,\ldots,n), \qquad f(x) = \frac{2x^2+x-1}{x^2-x+1}$$
$$n = 50$$
$$(a,b) = (1,1) \text{ bzw. } (a,b) = (1,10)$$

Der Parameter a steuert die Größe der maximalen Abszissendistanz, b die Verteilung der Abszissen. Die Abszissen x_k sind so definiert, daß - im Gegensatz zu den y_k bei einfacher Genauigkeit - durch Abschneiden keine Fehler entstehen. Bei Rechnung mit doppelter Genauigkeit stimmen die Ergebnisse für die y''_k bei A - C', A - C" und BU auf mindestens 8 Ziffern

in der Gleitkommadarstellung überein. Beim Übergang von einfacher auf doppelte Genauigkeit stimmen bei BU für (a,b) = (1,1) mindestens vier und bei (a,b) = (1,10) mindestens fünf Ziffern überein; bei A - C" dagegen nur eine bzw. zwei Ziffern. Bei A - C' treten bei einfacher Genauigkeit Fehler der Größenordnung 10^3 bzw. 10^2 auf.

(5.10) Beispiel 2

$$x_1 := 10$$
$$x_k := x_{k-1} + .01 + |a \times \sin(b \times k)| \qquad (k=2,\ldots,n)$$
$$y_k := x_k^2 \qquad (k=1,\ldots,n)$$
$$n = 50, \quad a,b \text{ variabel}$$

Hier besitzen sowohl Abszissen als auch Ordinaten einen durch Abschneiden von Ziffern bedingten Fehler. Da $f(t) = t^2$ eine Spline-Funktion vom Grad 5 und vom Typ II ist, können wir hier die exakten Werte $y_k'' = 2$ $(k=2,\ldots,n-1)$ mit den von den verschiedenen Methoden bei durch das Abschneiden leicht gestörten Abszissen und Ordinaten berechneten Werten vergleichen. Tabelle 4 zeigt diesen Einfluß für verschiedene n und Paare (a,b) bei Rechnung mit einfacher Genauigkeit. In den drei letzten Spalten sind die Werte $\max_k |y_k'' - 2|$ eingetragen:

Tabelle 4

a	b	n	A - C'	A - C"	BU
1	1	40	.5241	.0360	.0170
		70	725.2878	.0803	.0469
		100	3282.959	4.8242	.1534
1	10	40	19.1318	.2703	.0146
		70	219.8710	.2703	.0759
		100	1139.242	1.5501	.3696
10	1	40	.0263	.0500	.0041
		70	27.9618	.0500	.0360
		100	281.0136	.1858	.0876

Bei doppelter Genauigkeit liefern alle Methoden in allen Fällen mindestens 11 richtige Dezimalen.

Für n > 100 ist die Verwendung von A - C' bzw. A - C" nicht empfohlen worden (CARASSO, LAURENT (1968)). Wir geben ein Beispiel an, womit dies auch bei Rechnung mit doppelter Genauigkeit bestätigt wird und das zeigt, daß BU hier A - C" überlegen ist.

(5.11) Beispiel 3

$$x_1 = 10, \qquad y_1 = 0$$

$$x_k = \frac{1}{100} \times \text{entier}\,(100 \times (x_{k-1} + |a \times \sin(b \times k)|) + .01$$

$$y_k = \frac{1}{1000} \times \text{entier}\,(10000 \times \cos(k))$$

$$(k=2,\ldots,n)$$

$$n = 150, \qquad a = 20, \quad b = 1$$

Bei Berechnung mit einfacher und doppelter Genauigkeit stimmten bei BU die Werte y_k'' auf mindestens 3 Dezimalen überein. In Tabelle 5, zweite Spalte, sind die drei betragsgrößten Differenzen der Werte y_k'' bei BU und A - C" bei Rechnung mit doppelter Genauigkeit angegeben. In der dritten Spalte sind die zu den entsprechenden Stützstellen gehörigen Abweichungen von BU bei Rechnung mit einfacher und mit doppelter Genauigkeit tabelliert.

Tabelle 5

k	$\lvert y''_{\text{BU}_{16\text{ Dez.}}} - y''_{\text{A-C''}_{16\text{ Dez.}}} \rvert$	$\lvert y''_{\text{BU}_{7\text{ Dez.}}} - y''_{\text{BU}_{16\text{ Dez.}}} \rvert$
88	$7.69_{10} - 3$	$1.25_{10} - 4$
118	$1.76_{10} - 3$	$6.10_{10} - 5$
132	$3.23_{10} - 2$	$1.05_{10} - 5$

Die Erfahrung hat uns gezeigt, daß BU unabhängig von n in dem Sinne numerisch stabil ist, daß sich die Ergebnisse bei Rechnung mit einfacher und doppelter Genauigkeit wenig unterscheiden. Im allgemeinen stimmen mindestens 3, meistens 4 Dezimalen überein.

Aus den angegebenen Beispielen ist ersichtlich, daß bei genügender Stellenzahl von allen Methoden für $n \leq 100$ etwa gleichgute Resultate erzielt werden. Bei ungenügender Zahlenlänge liefert im Vergleich zu BU die Methode A - C" nicht sehr gute, die Methode A - C' unbrauchbare Ergebnisse.

Da man nun im allgemeinen von vornherein die notwendige Stellenzahl gar nicht kennt und sie auf einer Rechenmaschine eventuell auch nicht zur Verfügung hat, ist hiermit der Hauptvorteil unserer Methode BU demonstriert. Dieser wird dadurch bedingt, daß bei unserer Methode BU unabhängig von m stets nur Differenzen von zwei Differenzquotienten in (2.20) auftreten, während bei A - C' und A - C" dividierte Differenzen der Ordnung m+1 in den b_i und der Ordnung 2m+2 in den a_{ij} eingehen. Es scheint, als ob der Rundungsfehler in den b_i kritischer als in den a_{ij} ist.

Bevor wir dieses Ergebnis für $m > 2$ reproduzieren, vergleichen wir noch den Rechenaufwand für $m = 2$. Dabei benutzen wir als Maßeinheit die Zeit für eine Multiplikation, wobei der Bedarf zweier Additionen dem einer Multiplikation gleichgesetzt wird.

Auf diese Weise erhält man den Rechenaufwand für BU bei $m = 2$ zu

(5.12) $$(n-2)\left(19 + 19\alpha + \frac{43}{2}\beta\right)$$

Dann bedeutet α die Anzahl der notwendigen Iterationen zur Bestimmung von $\rho(A_1^{-1}A_2A_1^{-1}A_o)$ nach der Potenzmethode und β die Anzahl der zum Erreichen einer bestimmten Genauigkeit durchzuführenden Iterationen beim Blockunterrelaxationsverfahren. Nun hängt α vom Verhältnis der beiden größten Eigenwerte von $A_1^{-1}A_oA_1^{-1}A_2$ und β vom Spektralradius der Blockrelaxationsmatrix ab. Beide Zahlen sind äußerst komplizierte Funktionen der vorgegebenen Abszissen und sind a priori nicht bekannt. Für $\frac{\Delta x_{max}}{\Delta x_{min}} \sim 10$ und nicht allzu pathologische Verteilungen der Abszissen sind die Werte $\alpha = 6$ und $\beta = 13$ realistisch. Da für A - C" bei optimaler Programmierung approximativ 400(n-2) Multiplikationszeiten benötigt werden, sind für den genannten Fall die Methoden gleich schnell. Es sind jedoch durchaus praktische Fälle denkbar, wo A - C" schneller ist. So ist für $\alpha = 10$ und $\beta = 50$ A - C" z.B. etwa dreimal schneller als BU. Andererseits zeigt die Erfahrung, daß für größere β der Einfluß der Rundungsfehler bei A - C" stärker wächst als bei BU.

Betrachten wir nun den Fall äquidistanter Abszissen und $m \geq 2$. Hier gäbe es zwar eine elegante, eigens auf äquidistante Abszissen zurecht geschnittene und sehr schnelle Methode (CARASSO (1966)), die jedoch viel schlechtere Ergebnisse liefert als die Methoden A - C' oder A - C". Benutzt man zum Vergleich A - C', so kann in Schritt A die Matrix a_{ij} ein für allemal berechnet werden. Die Bildung dividierter Differenzen der Ordnung m+1 bei den b_i ist jedoch unvermeidbar.

Wir vergleichen A - C' mit der Blockunterrelaxation zunächst an einem Beispiel wo man die Splines vom Typ II und III zusammenfallen lassen kann, und so der Vergleich überhaupt möglich ist:

(5.13) Beispiel 4

$$m = 4$$
$$x_k := k - 1 \qquad (k=1,\ldots,n)$$
$$y_k := T_4\left(\frac{k-1}{n-1}\right), \quad T_4(t) = 8t^4 - 8t^2 + 1$$

Mit dem Blockunterrelaxationsverfahren BU erreichten wir für n = 51,101,201,401,801,1601 bei doppelter Genauigkeit unabhängig von n für y_k'' mindestens 13, für y_k^{iv} mindestens 12, für y_k^{vi} mindestens 10 und für y_k^{viii} mindestens 9 korrekte Dezimalen. Bei A - C' wurden für n = 51 bei y_k'' und y_k^{iv} 12 und bei n = 101 bei y_k'' ebenfalls 12 und bei y_k^{iv} nur 11 korrekte Ziffern erreicht. Größere n konnten hier aus Rechenzeit- und Speicherplatzbegrenzungen nicht verglichen werden. A - C' schneidet für n = 51,101 nur wenig schlechter ab als BU. Für größere n ist jedoch zu erwarten, daß BU überlegen ist.

Anders sieht dies aus bei Rechnung mit einfacher Genauigkeit. Hier werden nach (5.13) bei den y_k neun Dezimalen abgeschnitten. Für die gleichen Werte von n wie oben erhält man bei BU bei y'' drei, bei y^{iv} zwei, bei y^{vi} eine und bei y^{viii} keine exakten Dezimalen. Bei A - C' dagegen erhält man für n = 51 absolute Fehler der Größenordnung 10^3 für y'' und y^{iv}, bei n = 101 gar solche von 10^7.

Im letzten Beispiel wollen wir das Verhalten von BU beim Übergang vom einfacher und doppelter Genauigkeit betrachten:

(5.14) Beispiel 5

$$m = 3,6,9$$
$$n = 200$$
$$x_k := k - 1 \qquad (k=1,\ldots,n)$$
$$y_k := \text{entier}\ (100 \times (\sin(k) + \cos(n-k)))$$

Vergleicht man die bei den beiden Genauigkeiten berechneten Werte für die Ableitungen der Ordnung 2j (j=1,...,m), so unterscheiden sich diese in den

meisten Fällen in den ersten sechs Dezimalen nicht. Manchmal stimmen aber auch nur fünf und ganz selten nur vier Dezimale überein. Diese empirische Aussage gilt unabhängig von m.

Der Aufwand für die Rechenzeit pro Iteration ist bei BU

$$(5.15) \qquad (n-2)\,(7 + \tfrac{27}{4}\,m + \tfrac{7}{4}\,m^2)$$

Multiplikationszeiten. Bei den in der Tabelle 3 angegebenen Relaxationsparametern benötigt man bei einfacher Genauigkeit erfahrungsgemäß 2m+3 Iterationen ($2 \leq m \leq 10$) unabhängig von n, bis sich bei zwei aufeinanderfolgenden Schritten die Werte für die Unbekannten nicht mehr unterscheiden. Somit geht aus (5.15) hervor, daß der Aufwand, wie schon bei m = 2, linear mit n wächst. Ein Vergleich des Rechenaufwands mit A - C' oder A - C" ist nicht auf faire Weise durchführbar, da dort die durch die Äquidistanz möglichen Vereinfachungen nicht überschaubar sind.

Abschließend wollen wir noch einige Bemerkungen über den Einfluß der Randbedingungen, also des Typs K, auf die Gestalt der Kurve und der optimalen Beschleunigungsparameter bei BU machen. Nach Kapitel 3 können wir für m = 2 die Typen I, II und III berechnen. Als Ausgangsdaten wählen wir

(5.12) Beispiel 6

$x_1 := 0,\ x_k := .01 \times \mathrm{entier}(100 \times (x_{k-1} + .01 + |a \times \sin(b \times k)|)) \quad (k=2,\ldots,n)$

$y_k := \mathrm{entier}(100 \times \sin(k)) + \mathrm{entier}(100 \times \cos(n-k)) \quad (k=1,\ldots,n)$

Typ I : $y_1' = y_1'' = y_n' = y_n'' = 0$

Typ II : $y_1''' = y_1^{iv} = y_n''' = y_n^{iv} = 0$

Typ III: $y_1'' = y_1^{iv} = y_n'' = y_n^{iv} = 0$

(a,b) = (1,1), (1,10), (10,1)

n = 100, 500

Hier sind die Zahlenwerte für die Randbedingungen von Typ I und Typ III recht willkürlich gewählt. Trotzdem ergab sich in allen Fällen, also unabhängig vom Typ, von n und von (a,b), daß für k=15,...,n-16 in die Koeffizienten der Polynome, aus denen sich die jeweilige Spline-Funktion zusammensetzt, bei Rechnung mit einfacher Genauigkeit auf 6 Dezimalen übereinstimmen. Bei weniger pathologischen Randwerten für die einzelnen Typen in (5.12) tritt erfahrungsgemäß die Übereinstimmung schon viel näher bei den Rändern ein. Weiter war bei festen Zahlen a und b der berechnete Wert des optimalen Beschleunigungsparameters bei obigem Beispiel unabhängig vom Typ und unterschied sich für n = 100 und n = 500 erst in der fünften Dezimale.

Diese Erfahrungen lassen uns vermuten, daß bei den in Kapitel 3 für m > 2 und Typ K als prinzipiell möglich erkannten Modifikationen von (2.19) die in den Tabellen 2 und 3 angegebenen optimalen Beschleunigungsparameter für Typ III auch für - Typ K entsprechend - abgeänderte Gleichungssysteme bei äquidistanten Abszissen und nicht zu kleiner Stützstellenzahl gültig sind.

Literaturverzeichnis

AHLBERG J.H., NILSON E.N., WALSH J.L. (1967)
The Theory of Splines and their Applications
Academic Press, New York

ANSELONE P.M., LAURENT P.J. (1968)
A General Method for the Construction of Interpolating or Smoothing Spline-Functions
Num. Math. 12, 66-82

BOOR C. de, RICE J.R. (1968)
Least Squares Cubic Spline Approximation
I: Fixed Knots, II: Variable Knots
CSD TR 20 und 21, Purdue University

CARASSO C. (1966)
Méthodes numériques pour l'obtention de Fonctions-Spline
Dissertation, Universität Grenoble

CARASSO C. (1967) in
Procédures ALGOL en Analyse numérique
Centre National de la Recherche Scientifique, Paris

CARASSO C., LAURENT P.J. (1968)
On the numerical Construction and the practical use of interpolating Spline-Functions
IFIP Congress, Edinburgh

CURTIS A.R., POWELL M.J.D. (1967)
Using Cubic Splines to Approximate Functions of One Variable to prescribed Accuracy
AERE-R 5602, Atomic Energy Research Establishment, Harwell, England

FADDEJEW, D.K., FADDEJEWA W.N. (1964)
Numerische Methoden der linearen Algebra
Oldenburg Verlag München-Wien

GREVILLE T.N.E. (1967)
Spline Functions, Interpolation, and Numerical Quadrature
in RALSTON/WILF
Mathematical Methods for Digital Computers, Vol. II
J. Wiley & Sons, p. 156-168

GREVILLE T.N.E. (1969)
Theory and Applications of Spline-Functions
Academic Press, New York

LIDSTONE G.J. (1930/31)
Notes on the Extension of AITKEN's Theorem (for Polynomial Interpolation) to the EVERETT Type
Edingburgh Math. Soc. Proc., Series II, 16-19

NIETHAMMER W. (1964)
Überrelaxation bei linearen Gleichungssystemen mit schiefsymmetrischer Koeffizientenmatrix
Dissertation, Universität Tübingen

NÖRLUND E.N. (1954)
Vorlesungen über Differenzenrechnung
Springer Verlag

REINSCH C.H. (1967)
Smoothing by Spline Functions
Num. Math. 10, 177-183

SCHOENBERG I.J. (1964)
Spline Functions and the Problem of Graduation
Proc. Nat. Acad. Sci. USA 52, 947-950

SCHUMAKER L.L. (1969)
Some Algorithms for the Computation of Interpolating and Approximating Spline Functions
in: GREVILLE (1969)

SPÄTH, H. (1968)
Ein Verfahren zur flächentreuen Approximation von Treppenfunktionen durch glatte Kurven
ZAMM 48, T106-107

VARGA R.S. (1962)
Matrix Iterative Analysis
Prentice Hall

WHITTAKER J.M. (1934)
On LIDSTONE's Series and Two-Point Expansions of Analytic Functions
Proc. London Math. Soc., Vol. 36, 451-469